UNIVERSITY OF CALIFORNIA SAN DIEGO

Large-dimensional Expectile Regression with Heavy-tailed Data

A dissertation submitted in partial satisfaction of the
requirements for the degree Doctor of Philosophy

in

Mathematics with a Specialization in Statistics

by

Zian Wang

Committee in charge:

Professor Wenxin Zhou, Chair
Professor Ery Arias-Castro
Professor Jiawang Nie
Professor Yixiao Sun
Professor Danna Zhang

2023

The Dissertation of Zian Wang is approved, and it is acceptable in quality and form for publication on microfilm and electronically.

University of California San Diego

2023

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

First and foremost, I am extremely grateful to my advisor, Professor Wenxin Zhou, for his constant professional support during my study. He is my role model as mathematicians, his immense knowledge and patient guidance have encouraged me in all the time of my academic research. This work would not have been finished without his guidance.

Besides, I would like to express my sincere gratitude to my committee, including Professor Danna Zhang, Professor Ery Arias-Castro, Professor Jiawang Nie and Professor Yixiao Sun. Their insightful comments and suggestions helped me to further improve the dissertation.

Last but not the least, I wish to thank all of my friends and my parents for their love and support in all my life.

Chapter 1, in part, is a reprint of the material in the paper "Retire: Robust Expectile Regression in High Dimensions", Man, Rebeka; Tan, Kean Ming; Wang, Zian and Zhou, Wen-Xin. The paper has been accepted by *Journal of Econometrics*, 2023. The dissertation author was the primary investigator and author of this paper.

Chapter 2, in part, is a reprint of the material in the paper "Retire: Robust Expectile Regression in High Dimensions", Man, Rebeka; Tan, Kean Ming; Wang, Zian and Zhou, Wen-Xin. The paper has been accepted by *Journal of Econometrics*, 2023. The dissertation author was the primary investigator and author of this paper.

Chapter 3, in part, is a reprint of the material in the paper "A Unified Algorithm for Penalized Robust Expectile Regression", Wang, Zian and Zhou, Wen-Xin. Preprint. The dissertation author was the primary investigator and author of this paper.

VITA

2017	B.S. in Mathematics, University of Science and Technology of China
2017–2023	Graduate Teaching and Research Assistant, University of California San Diego
2023	Ph. D. in Mathematics with a Specialization in Statistics, University of California San Diego

PUBLICATIONS

Man, R., Tan, K. M., Wang, Z., and Zhou, W.-X., Retire: Robust Expectile Regression in High Dimensions (2023). *Journal of Econometrics*, accepted

Wang, Z., and Zhou, W.-X., A Unified Algorithm for Penalized Robust Expectile Regression. Preprint. To be submitted.

PACKAGES

Wang, Z., and Zhou, W.-X., Robust Expectile Regression Package, GitHub.

ABSTRACT OF THE DISSERTATION

Large-dimensional Expectile Regression with Heavy-tailed Data

by

Zian Wang

Doctor of Philosophy in Mathematics with a Specialization in Statistics

University of California San Diego, 2023

Professor Wenxin Zhou, Chair

High-dimensional data can often display heterogeneity due to heteroscedastic variance or inhomogeneous covariate effects. Penalized quantile and expectile regression methods offer useful tools to detect heteroscedasticity in high-dimensional data. However, the former is computationally challenging due to the non-smooth nature of the check loss, and the latter is sensitive to heavy-tailed error distributions. In this thesis, we propose and study (penalized) robust expectile regression (`retire`) with random designs and heavy-tailed noises. Theoretically, we establish the statistical properties of the `retire` estimator under two regimes: (i) low-dimensional regime in which $d \ll n$; (ii) high-dimensional regime in which $s \ll n \ll d$ with s denoting the number of significant predictors. In the low-dimensional setting, we theoretically

establish explicit nonasymptotic high probability error bounds, Bahadur representation and a Berry-Esseen bound, from which we derive asymptotic normality for the `retire` estimator. In the high-dimensional setting, we focus on iteratively reweighted ℓ_1-penalization which reduces the estimation bias from ℓ_1-penalization and leads to oracle properties. We thoroughly analyze the statistical properties of the solution path of iteratively reweighted ℓ_1-penalized `retire` estimation, adapted from the local linear approximation algorithm for folded-concave regularization. Under a mild minimum signal strength condition, we demonstrate that with as few as $\log(\log d)$ iterations, the final iterate of our proposed approach achieves the oracle convergence rate. At each iteration, we solve the weighted ℓ_1-penalized convex program using a local adaptive majorize-minimization algorithm. Moreover, extensions to group-structured penalizations are also studied. Numerical studies demonstrate the promising performance of the proposed procedure in comparison to both non-robust and quantile regression based alternatives.

Chapter 1

Expectile Regression in Low Dimensions

1.1 Motivation and Overview

Simple linear regression has been widely used nowadays. Its focus is primarily on inferring the conditional mean of the response given the a set of predictors/covariates. In many economic applications, however, more aspects than the mean of the conditional distribution of the response given the covariates are of interest, and that the covariate effects may be inhomogeneous and/or the noise variables exhibit heavy-tailed and asymmetric tails. For instance, in the Job Training Partners Act studied in Abadie, Angrist and Imbens (2002), one is more interested in the lower tail than the mean of the conditional distribution of income given a large pool of predictors. To capture heterogeneity in the set of covariates at different locations of the response distribution, methods such as quantile regression (Koenker and Bassett, 1978) and asymmetric least squares regression (Newey and Powell, 1987) have been widely used.

Quantile regression expresses the conditional quantiles of the response as a linear function of the covariates, and finding the regression model involves minimizing a piece-wise linear loss. However, there are two major drawbacks to the quantile regression approach. One is that the loss function (check function) is not continuously differentiable, which leads to computational challenges. The other one, perhaps most importantly, is that quantile regression requires (non-parametric) estimation of density function for errors, and such estimation depends on the degree of smoothing chosen empirically by researchers.

An alternative approach to explore heterogeneity and/or asymmetry in the response distribution is the expectile regression (Newey and Powell, 1987), which is essentially a least squares analogue of regression quantile estimation. Both quantiles and expectiles are useful descriptors of the tail behavior of a distribution in the same way as the median and mean are related to its central behavior. They share similar properties, and as shown by Jones (1994), expectiles are exactly quantiles of a transformed version of the original distribution. Quantiles are naturally more dominant in the literature due to the fact that expectiles lack an intuitive interpretation while quantiles are simply the inverse of the distribution function and directly indicate relative frequency. The key advantage of expectile regression is its computational expediency (for example, via the iteratively-reweighted least squares algorithm), and the asymptotic covariance matrix can be estimated without the need of estimating the conditional density function (nonparametrically). Therefore, it offers a convenient and relatively efficient method of summarizing the conditional response distribution.

We continue to discuss the aforementioned approaches in the following section.

1.2 Quantile/Expectile Regression

In the section, we introduce the (uniform) problem setup for both quantile regression and expectile regression. Let $y \in \mathbb{R}$ be a scalar response variable and $\boldsymbol{x} = (x_1, \ldots, x_d)^{\mathrm{T}} \in \mathbb{R}^d$ be a d-dimensional vector of covariates. The training data $(y_1, \boldsymbol{x}_1), \ldots, (y_n, \boldsymbol{x}_n)$ are independent copies of $(y, \boldsymbol{x})$. Given a location parameter $\tau \in (0,1)$, we consider the linear model

$$y_i = \boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{\beta}^*(\tau) + \varepsilon_i(\tau), \tag{1.1}$$

where $\boldsymbol{\beta}^*(\tau)$ is the unknown d-dimensional vector of regression coefficients, and $\varepsilon_i(\tau)$'s are independent random noise. Model (1.1) allows the regression coefficients $\boldsymbol{\beta}^*(\tau)$ to vary across different values of τ, and thereby offers insights into the entire conditional distribution of y given $\boldsymbol{x}$. We let $x_1 = 1$ so that β_1^* denotes the intercept term, and suppress the dependency of $\boldsymbol{\beta}^*(\tau)$

and $\varepsilon(\tau)$ on τ whenever there is no ambiguity. Here we introduce two popular approaches, the quantile regression and the expectile regression, to estimate $\boldsymbol{\beta}^*$ at various level τ.

Quantile regression is perhaps the most natural way to relate the conditional distribution of y given $\boldsymbol{x}$ and the parameter process $\{\boldsymbol{\beta}^*(\tau), \tau \in (0,1)\}$, under the assumption that $F^{-1}_{y_i|\boldsymbol{x}_i}(\tau) = \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}^*(\tau)$, or equivalently, $\mathbb{P}\{\varepsilon_i(\tau) \leq 0 \,|\boldsymbol{x}_i\} = \tau$. Fitting such conditional quantile model involves minimizing a non-smooth piecewise linear loss function, $\varphi_\tau(u) = u\{\tau - \mathbb{1}(u<0)\}$, typically recast as a linear program, solvable by the simplex algorithm or interior-point methods. For the latter, Portnoy and Koenker (1997) showed that the average-case computational complexity grows as a cubic function of the dimension d, and thus, is computationally demanding for problems with large dimensions.

Expectile regression is the other approach. Adapted from the concept of quantiles, Newey and Powell (1987) and Efron (1991) separately proposed an alternative class of location measures of a distribution, named the expectile according to the former. The resulting regression methods are referred to as the expectile regression or the asymmetric least squares regression, which are easy to compute and reasonably efficient under normality conditions.

We start with the definition of expectiles. Let $Z \in \mathbb{R}$ be a random variable with finite moment, i.e., $\mathbb{E}(|Z|) < \infty$. The τ-th expectile or τ-mean of Z is defined as

$$e_\tau(Z) := \underset{u\in\mathbb{R}}{\operatorname{argmin}} \mathbb{E}\big\{\eta_\tau(Z-u) - \eta_\tau(Z)\big\}, \qquad \tau \in (0,1), \tag{1.2}$$

where

$$\eta_\tau(u) = |\tau - \mathbb{1}(u<0)| \cdot \frac{u^2}{2} = \frac{\tau}{2}\{\max(u,0)\}^2 + \frac{1-\tau}{2}\{\max(-u,0)\}^2 \tag{1.3}$$

is the asymmetric squared/ℓ_2 loss (Newey and Powell, 1987). The quantity $e_\tau(Z)$ is well defined as long as $\mathbb{E}|Z|$ is finite. When $\tau = 1/2$, it can be easily seen that $e_{1/2}(Z) = \mathbb{E}(Z)$. Therefore, expectiles can be viewed as an asymmetric generalization of the mean, and the term "expectile"

stems from a combination of "expectation" and "quantile". Moreover, expectiles are uniquely identified by the first-order condition

$$\tau \cdot \mathbb{E}(Z - e_\tau(Z))_+ = (1-\tau) \cdot \mathbb{E}(Z - e_\tau(Z))_-,$$

where $x_+ = \max(x, 0)$ and $x_- = \max(-x, 0)$. Note also that the τ-expectile of Z defined in (1.2) is equivalent to Efron's ω-mean with $\omega = \tau/(1-\tau)$ (Efron, 1991).

Given independent observations $Z_1, \ldots, Z_n$ from Z, the expectile location estimator is given by $\hat{e}_\tau = \operatorname{argmin}_{u \in \mathbb{R}} \sum_{i=1}^n \eta_\tau(Z_i - u)$, which is uniquely defined due to the strong convexity of the asymmetric ℓ_2-loss. The expectile estimator $\hat{e}_\tau$ can also be interpreted as a maximum likelihood estimator of a normal distributed sample with unequal weights given to disturbances of differing signs, with a larger relative weight given to less variable disturbances (Aigner, Amemiya and Poirier, 1976).

Essentially the asymmetric squared loss $\eta_\tau(\cdot)$ is an ℓ_2-version of the check function $\varphi_\tau(\cdot)$ for quantile regression. Given train data from the linear model (1.1) subject to $e_\tau(\varepsilon_i | \boldsymbol{x}_i) = 0$, the expectile regression estimator (Newey and Powell, 1987) is defined as the minimizer of the following convex optimization problem

$$\underset{\boldsymbol{\beta} \in \mathbb{R}^d}{\text{minimize}} \; \frac{1}{n} \sum_{i=1}^n \eta_\tau(y_i - \boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{\beta}), \tag{1.4}$$

which consistently estimates $\boldsymbol{\beta}^*$ when $d = o(n)$ as $n \to \infty$. In particular, expectile regression with $\tau = 0.5$ reduces to the ordinary least squares regression.

1.3 Interpretation of Quantiles and Expectiles

Expectiles are computational efficient, but not as intuitive as quantiles. Mathematically, the notion of expectiles is a least squares counterpart of quantiles, and can be viewed as an alternative measure of "locations" of a random variable $Z \in \mathbb{R}$ with finite moment, i.e., $\mathbb{E}(|Z|) < \infty$.

For example, $1/2$-expectile and $1/2$-quantile correspond to the mean and median, both of which are related to the central behavior. In general, τ-expectile (e_τ) and τ-quantile (q_τ) with τ close to zero and one describe the lower and higher regions of the distribution of Z, respectively. The latter is the point below which $100\tau\%$ of the mass of Z lies, i.e., $q_\tau = F_Z^{-1}(\tau)$; whereas the former specifies the "position" such that the average distance from Z below e_τ to e_τ itself is $100\tau\%$ of the average distance between Z and e_τ, i.e., $\tau = \{\mathbb{E}(Z - e_\tau)_-\}/\mathbb{E}|Z - e_\tau|$.

Both quantile and expectile regression have found applications in various fields, including risk analysis (Taylor, 2008, Kuan et al., 2009, Xie et al., 2014, Bellini and Bernardino, 2017, Daouia et al., 2018), as well as the study of determinants of inflation (Busetti et al., 2021) and life expectancy and economic production (Schnabel and Eilers, 2009). In finance applications, quantile based Value at Risk (QVaR) is one of the most popular risk measures. Specifically, given $\alpha \in (0,1)$, QVaR is defined as the tail quantiles q_α of some underlying return distribution, and it may be interpreted as the possible maximum loss of a given portfolio over a prescribed holding period with probability $(1-\alpha)$.

Despite its overwhelming popularity in financial risk analysis, QVaR has two major drawbacks. Firstly, QVaR is not a coherent risk measure since it lacks subadditivity, meaning that the total risk of a portfolio can be even larger than the sum of the risks of the portfolio's constituent components, which violates the conventional wisdom that diversification reduces risk. Secondly, perhaps most importantly, QVaR reports only the tail probability of losses while ignoring information on the magnitude of losses. To see this, recall that the α-quantile $q_\alpha \in \mathbb{R}$ minimizes $\mathbb{E}[|\alpha - \mathbb{1}(Z \leq q_\alpha)| \cdot |Z - q_\alpha|]$, and its first order condition implies $\int_{-\infty}^{q_\alpha} \mathrm{d}F(z) = \alpha$. Consequently, q_α depends only on the probability of extreme losses but not their magnitude.

Expected Shortfall (ES) is another popular risk measure in finance (Acerbie and Tasche, 2002). It is defined as the conditional expectation of a loss given that it exceeds the QVaR, i.e., $\mathbb{E}(Z|Z > q_\alpha)$. Although ES is a coherent risk measure that considers the magnitude of losses and possesses subadditivity by the nature of expectations, it is not elicitable (Gneiting, 2011), meaning that the risk measure can not be obtained by minimizing the expectation of a forecasting

objective function, which poses challenges for its estimation.

Expectile based VaR (EVaR) is a risk measure to remedy the aforementioned issues. In fact, the EVaR is the only risk measure that is both coherent and elicitable, making it a valuable tool for financial risk management and decision-making (Bellini and Bernardino, 2017, Ziegel, 2016, Bellini and Bignozzi, 2015). Specifically, the τ-expectile is the quantity $e_\tau \in \mathbb{R}$ that minimizes $\mathbb{E}[|\tau - \mathbb{1}(Z < e_\tau)| \cdot (Z - e_\tau)^2]$. Rearranging the first order condition yields $\tau = \int_{-\infty}^{e_\tau} |y - e_\tau| \mathrm{d}F(y) / \int_{-\infty}^{\infty} |y - e_\tau| \mathrm{d}F(y)$. Hence, e_τ depends on both the magnitude and the probabilities of tail realizations, whereas q_α is determined solely by the (lower) tail probabilities. In other words, expectiles are more "global" than quantiles, since changing the upper tails or the magnitude of Z would only affect the expectiles.

Indeed, let $\tau(\alpha)$ be such that $e_{\tau(\alpha)} = q_\alpha$, Yao, Q. and Tong, H. (1996) pointed out an one-to-one mapping:

$$\tau(\alpha) = \frac{\alpha q_\alpha - \int_{-\infty}^{q_\alpha} z \mathrm{d}F(z)}{\mathbb{E}(Z) - 2\int_{-\infty}^{q_\alpha} z \mathrm{d}F(z) - (1-2\alpha)q_\alpha}.$$

It can be seen that the EVaR with a given τ corresponds to the QVaR with distinct tail probabilities α under different distributions. Thus, EVaR may be interpreted as a flexible QVaR, in the sense that its confidence level $\tau(\alpha)$ is determined by the distribution of Z. This is in contrast with the conventional QVaR with a given α. Note that $\alpha \neq \tau$ generally.

1.4 Retire: Robust Expectile Regression

Expectile regression (1.4), despite its computational advantage over quantile regression, is much more sensitive to heavy-tailed distributions due to the squared loss component in (1.3). This lack of robustness is amplified in the presence of high-dimensional covariates and heavy-tailed random noise, causing the estimated coefficients $\hat{\boldsymbol{\beta}}$ to deviate from the true underlying coefficients $\boldsymbol{\beta}^*$. This necessitates the development of a robust expectile regression approach that utilizes a new class of asymmetric loss functions that preserves the robustness of the check loss

to some degree.

To this end, we construct a class of asymmetric robust loss functions that is more resistant against heavy-tailed error/response distributions. The main idea is to replace the quadratic component in (1.3) with a Lipschitz and locally strongly convex alternative, typified by the Huber loss (Huber, 1964) that is a hybrid ℓ_1/ℓ_2 function. The proposed loss function, $\ell_\gamma(u)$, contains a tuning parameter $\gamma > 0$ that is to be chosen to achieve a balanced trade-off between the robustification bias and the degree of robustness. At a high level, we focus on the class of loss functions that satisfies Condition 1 below.

Condition 1. *Let $\ell_\gamma(u) = \gamma^2 \ell(u/\gamma)$ for $u \in \mathbb{R}$, where the function $\ell : \mathbb{R} \mapsto [0,\infty)$ satisfies: (i) $\ell'(0) = 0$ and $|\ell'(u)| \le \min(a_1, |u|)$ for all $u \in \mathbb{R}$; (ii) $\ell''(0) = 1$ and $\ell''(u) \ge a_2$ for all $|u| \le a_3$; and (iii) $|\ell'(u) - u| \le u^2$ for all $u \in \mathbb{R}$, where a_1, a_2, and a_3 are positive constants.*

Condition 1 encompasses many commonly used robust loss functions such as the Huber loss $\ell(u) = \min\{u^2/2, |u| - 1/2\}$ (Huber, 1964), pseudo-Huber losses $\ell(u) = \sqrt{1+u^2} - 1$ and $\ell(u) = \log(e^u/2 + e^{-u}/2)$, smoothed Huber losses $\ell(u) = \min\{u^2/2 - |u|^3/6, |u|/2 - 1/6\}$ and $\ell(u) = \min\{u^2/2 - u^4/24, (2\sqrt{2}/3)|u| - 1/2\}$, among other smooth approximations of the Huber loss (Lange, 1990).

Consequently, we consider the following asymmetric robust loss

$$L_{\tau,\gamma}(u) := |\tau - \mathbb{1}(u < 0)| \cdot \ell_\gamma(u), \tag{1.5}$$

where $\ell_\gamma(\cdot)$ is subject to Condition 1. Note that τ determines the location (asymmetry level) and γ determines the trade-off between the robustification bias and the degree of robustness.

Given a location parameter $\tau \in (0,1)$, we define the `retire` estimator (when $d < n$) as

$$\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}_\gamma = \underset{\boldsymbol{\beta} \in \mathbb{R}^d}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^{n} L_{\tau,\gamma}(y_i - \boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{\beta}), \tag{1.6}$$

where $\gamma > 0$ is a robustification parameter that will be calibrated adaptively from data as we

detail in Section 1.7. Numerically, the optimization problem (1.6) can be efficiently solved by either gradient descent or quasi-Newton methods (Nocedal and Wright, 1999), such as the Broyden-Fletcher-Goldfarb-Shanno (BFGS) algorithm that can be implemented as on option of the base function `optim()` in R.

Recall that the population parameter $\boldsymbol{\beta}^*$ is uniquely identified as

$$\boldsymbol{\beta}^* = \underset{\boldsymbol{\beta}\in\mathbb{R}^d}{\operatorname{argmin}}\ \mathbb{E}\{L_{\tau,\infty}(y-\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta})\} \ \text{ with } \ L_{\tau,\infty}(u) := |\tau - \mathbb{1}(u<0)|\cdot u^2/2.$$

Meanwhile, $\hat{\boldsymbol{\beta}}$ can be viewed an M-estimator of the following population parameter

$$\boldsymbol{\beta}^*_\gamma := \underset{\boldsymbol{\beta}\in\mathbb{R}^d}{\operatorname{argmin}}\,\mathbb{E}\{L_{\tau,\gamma}(y-\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta})\}.$$

It is worth pointing out that $\boldsymbol{\beta}^*_\gamma$ typically differs from $\boldsymbol{\beta}^*$ for any given $\gamma>0$. To see this, note that the convexity of the robust loss $L_{\tau,\gamma}:\mathbb{R}^d\to\mathbb{R}$ implies the first-order condition, that is, $\mathbb{E}\{|\tau - \mathbb{1}(y<\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta}^*_\gamma)|\cdot\ell'_{\tau,\gamma}(y-\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta}^*_\gamma)\boldsymbol{x}\}=\mathbf{0}$. On the other hand, we have $e_\tau(\varepsilon|\boldsymbol{x}) = e_\tau(y-\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta}^*|\boldsymbol{x})=0$, implying $\mathbb{E}\{|\tau-\mathbb{1}(\varepsilon<0)|\cdot\varepsilon\boldsymbol{x}\}=\mathbf{0}$. Since the random error ε given $\boldsymbol{x}$ is asymmetric around zero, in general we have

$$\mathbf{0}\neq\mathbb{E}\{|\tau-\mathbb{1}(\varepsilon<0)|\cdot\ell'_{\tau,\gamma}(\varepsilon)\boldsymbol{x}\} = \mathbb{E}\{|\tau-\mathbb{1}(y<\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta}^*)|\cdot\ell'_{\tau,\gamma}(y-\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta}^*)\boldsymbol{x}\},$$

which in turn implies that $\boldsymbol{\beta}^*\neq\boldsymbol{\beta}^*_\gamma$. We refer to the difference $\|\boldsymbol{\beta}^*_\gamma-\boldsymbol{\beta}^*\|_2$ as the robustification bias. In Section 1.6, we will show that under mild conditions, the robustification bias is of the order $\mathscr{O}(\gamma^{-1})$, and a properly chosen γ balances bias and robustness.

To perform statistical inference on β^*_j's, we construct normal-based confidence intervals based on the asymptotic theory developed in Section 1.6. To this end, we first introduce some additional notation. Let $\hat{\varepsilon}_i = y_i - \boldsymbol{x}_i^{\mathrm{T}}\hat{\boldsymbol{\beta}}$ be the residuals from the fitted model and let $\mathbf{e}_j\in\mathbb{R}^d$ be the canonical basis vector, i.e., the j-th entry equals one and all other entries equal zero. Let

$\hat{\mathbf{J}} = n^{-1}\sum_{i=1}^{n} |\tau - \mathbb{1}(\hat{\varepsilon}_i < 0)| \cdot \boldsymbol{x}_i \boldsymbol{x}_i^{\mathrm{T}}$. An approximate 95% confidence interval for β_j^* can thus be constructed as

$$\left[\hat{\beta}_j - 1.96\frac{\hat{\sigma}(\mathbf{e}_j)}{\sqrt{n}},\ \hat{\beta}_j - 1.96\frac{\hat{\sigma}(\mathbf{e}_j)}{\sqrt{n}}\right], \tag{1.7}$$

where

$$\hat{\sigma}^2(\mathbf{e}_j) := \mathbf{e}_j^{\mathrm{T}} \hat{\mathbf{J}}^{-1} \left[\frac{1}{n}\sum_{i=1}^{n} \zeta^2(\hat{\varepsilon}_i)\boldsymbol{x}_i \boldsymbol{x}_i^{\mathrm{T}}\right] \hat{\mathbf{J}}^{-1} \mathbf{e}_j,$$

and $\zeta(u) = L'_{\tau,\gamma}(u) = |\tau - \mathbb{1}(u < 0)| \cdot \ell'_\gamma(u)$ is the first-order derivative of $L_{\tau,\gamma}(\cdot)$ given in (1.5).

1.5 Computational Methods

In this section, we provide a computational method (gradient descent with Barzilai-Borwein stepsize) to solve the optimization problems (1.6). Starting from the convex and continuous differentiable loss function $L_{\tau,\gamma}$, the simplest and most intuitive algorithm is perhaps a vanilla gradient descent (GD) algorithm. Let $\mathscr{R}_n(\boldsymbol{\beta}) = \frac{1}{n}\sum_{i=1}^{n} L_{\tau,\gamma}(y_i - \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta})$, given an initializer $\boldsymbol{\beta}^{(0)} \in \mathbb{R}^d$, GD iteratively computes

$$\boldsymbol{\beta}^{(t+1)} = \boldsymbol{\beta}^{(t)} - \eta_t \cdot \nabla \mathscr{R}_n(\boldsymbol{\beta}^{(t)}) \ \text{ for } \ t = 0, 1, \ldots,$$

where $\eta_t \in \mathbb{R}$ is the stepsize for t-th iteration, and we choose a fixed stepsize for vanilla GD at each iteration, i.e., $\eta_t = \eta$. It is worth mentioning that the choice of stepsize is one of the most important issues for GD-type algorithm. Larger stepsizes tend to overshoot, while smaller stepsizes suffer from slow convergence speed. Note that the loss function $L_{\tau,\gamma}$ is twice differentiable, it is natural to employ the Newton-Raphson method that utilizes the inverse of Hessian matrix $\{\nabla^2 \mathscr{R}_n(\boldsymbol{\beta})\}^{-1}$ to serve as adaptive stepsizes. More specifically, Newton-Raphson

method iteratively computes

$$\boldsymbol{\beta}^{(t+1)} = \boldsymbol{\beta}^{(t)} - \left\{\nabla^2 \mathscr{R}_n(\boldsymbol{\beta}^{(t)})\right\}^{-1} \cdot \nabla \mathscr{R}_n(\boldsymbol{\beta}^{(t)}) \ \text{ for } \ t = 0, 1, \ldots.$$

Newton-Raphson has been proven to be very successful in solving nonlinear optimization problems. It enjoys fast convergence rates since it uses second-order information from the Hessian matrix $\nabla^2 \mathscr{R}_n(\boldsymbol{\beta})$. However the computation of the inverse of a $d \times d$ matrix may be quite expensive or numerically unstable, especially when d is large. For this reason, many quasi-Newton methods seek a simple approximation of the inverse of Hessian matrix. Here we introduce the method of gradient descent with Barzilai-Borwein (Barzilai and Browein, 1988) stepsize, which we refer as GD-BB algorithm.

Recall that at the t-th iteration of Newton-Raphson method, we have the secant equation $\mathbf{J}^{(t)}\boldsymbol{\delta}^{(t)} = \boldsymbol{g}^{(t)}$, where

$$\mathbf{J}^{(t)} = \nabla^2 \mathscr{R}_n(\boldsymbol{\beta}^{(t)}), \ \boldsymbol{\delta}^{(t)} = \boldsymbol{\beta}^{(t)} - \boldsymbol{\beta}^{(t-1)} \ \text{ and } \ \boldsymbol{g}^{(t)} = \nabla \mathscr{R}_n(\boldsymbol{\beta}^{(t)}) - \nabla \mathscr{R}_n(\boldsymbol{\beta}^{(t-1)}).$$

To find the approximation of $\mathbf{J}^{(t)}$, Barzilai and Browein (1988) considered choosing the stepsize η that satisfies $\eta \mathrm{I_d} \nabla \mathscr{R}_\mathrm{n}(\boldsymbol{\beta}^{(\mathrm{t})}) \approx \left\{\mathbf{J}^{(\mathrm{t})}\right\}^{-1} \nabla \mathscr{R}_\mathrm{n}(\boldsymbol{\beta}^{(\mathrm{t})})$. By the secant equation $\mathbf{J}^{(t)}\boldsymbol{\delta}^{(t)} = \boldsymbol{g}^{(t)}$, it suffices to choose η that satisfies $\eta^{-1}\boldsymbol{\delta}^{(t)} \approx \boldsymbol{g}^{(t)}$ or $\boldsymbol{\delta}^{(t)} \approx \eta \boldsymbol{g}^{(t)}$. Via least squares approximations, one may use $\eta_{1,t}^{-1} = \mathrm{argmin}_\alpha \, ||\alpha \boldsymbol{\delta}^{(t)} - \boldsymbol{g}^{(t)}||_2^2$ or $\eta_{2,t} = \mathrm{argmin}_\eta \, ||\boldsymbol{\delta}^{(t)} - \eta \boldsymbol{g}^{(t)}||_2^2$. Therefore the BB stepsizes have the following explicit forms

$$\eta_{1,t} = \frac{\langle \boldsymbol{\delta}^{(t)}, \boldsymbol{\delta}^{(t)} \rangle}{\langle \boldsymbol{\delta}^{(t)}, \boldsymbol{g}^{(t)} \rangle} \ \text{ and } \ \eta_{2,t} = \frac{\langle \boldsymbol{\delta}^{(t)}, \boldsymbol{g}^{(t)} \rangle}{\langle \boldsymbol{g}^{(t)}, \boldsymbol{g}^{(t)} \rangle}.$$

Consequently, the BB iteration takes the form

$$\boldsymbol{\beta}^{(t+1)} = \boldsymbol{\beta}^{(t)} - \min\left\{\eta_{1,t}, \eta_{2,t}, \eta_{max}\right\} \cdot \nabla \mathscr{R}_n(\boldsymbol{\beta}^{(t)}) \ \text{ for } \ t = 1, 2, \ldots,$$

where η_{max} is the max learning rate to avoid overshoots. It is worth mentioning that the BB iteration starts at the 1-th iteration, while $\boldsymbol{\beta}^{(1)}$ is simply the one-step update using classic vanilla gradient descent with stepsize 1 from some initializer $\boldsymbol{\beta}^{(0)}$. We summarize the aforementioned procedure into the following Algorithm 1

Algorithm 1. Gradient descent with Barzilai-Borwein stepsize (GD-BB) for solving (1.6).

Input: Expectile level τ, Huber loss tuning parameter γ, and convergence criterion ε.

Initialization: $\hat{\boldsymbol{\beta}}^{(0)} = 0, \eta_{max} = 50$.

Compute: $\boldsymbol{\beta}^{(1)} \leftarrow \boldsymbol{\beta}^{(0)} - \nabla \mathscr{R}_n(\boldsymbol{\beta}^{(0)})$.

Iterate: the following until the stopping criterion $\|\hat{\boldsymbol{\beta}}^{(k)} - \hat{\boldsymbol{\beta}}^{(k-1)}\|_2 \leq \varepsilon$ is met, where $\hat{\boldsymbol{\beta}}^{(k)}$ is the value of $\boldsymbol{\beta}$ obtained at the k-th iteration.

1. $\boldsymbol{\delta}^{(t)} \leftarrow \boldsymbol{\beta}^{(t)} - \boldsymbol{\beta}^{(t-1)}$ and $\boldsymbol{g}^{(t)} = \nabla \mathscr{R}_n(\boldsymbol{\beta}^{(t)}) - \nabla \mathscr{R}_n(\boldsymbol{\beta}^{(t-1)})$.
2. $\eta_{1,t} \leftarrow \langle \boldsymbol{\delta}^{(t)}, \boldsymbol{\delta}^{(t)} \rangle / \langle \boldsymbol{\delta}^{(t)}, \boldsymbol{g}^{(t)} \rangle$ and $\eta_{2,t} \leftarrow \langle \boldsymbol{\delta}^{(t)}, \boldsymbol{g}^{(t)} \rangle / \langle \boldsymbol{g}^{(t)}, \boldsymbol{g}^{(t)} \rangle$.
3. $\eta_t = \min\{\eta_{1,t}, \eta_{2,t}, \eta_{max}\}$.
4. $\boldsymbol{\beta}^{(t+1)} = \boldsymbol{\beta}^{(t)} - \eta_t \cdot \nabla \mathscr{R}_n(\boldsymbol{\beta}^{(t)})$

Output: the final iterate $\hat{\boldsymbol{\beta}}^{(k)}$.

1.6 Statistical Analysis

In this section, we consider the robust expectile regression (`retire`) estimator $\hat{\boldsymbol{\beta}}$ that is defined in (1.6) under the classical setting that $d < n$. Its statistical properties, both asymptotic and nonasymptotic, will be given under the so-called "many regressor" model (Belloni et al., 2015) in which the dimension $d = d_n$ is allowed to grow with n subject to the constraint $d_n = o(n^a)$ for some $0 < a \leq 1$. Note that our proposed estimator relies on the choice of robust loss function in Condition 1. For simplicity, we focus on the Huber loss $\ell(u) = u^2/2 \cdot \mathbb{1}(|u| \leq 1) + (|u| - 1/2) \cdot \mathbb{1}(|u| > 1)$ throughout our analysis, i.e., $a_1 = a_2 = a_3 = 1$ in Condition 1, but note that similar results hold for any robust loss that satisfies Condition 1. Throughout the theoretical analysis, we assume that the location measure $\tau \in (0,1)$ is fixed.

We first defined the empirical loss function and its gradient as

$$\mathscr{R}_n(\boldsymbol{\beta}) = \frac{1}{n}\sum_{i=1}^{n} L_{\tau,\gamma}(y_i - \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}) \quad \text{and} \quad \nabla\mathscr{R}_n(\boldsymbol{\beta}) = -\frac{1}{n}\sum_{i=1}^{n} L'_{\tau,\gamma}(y_i - \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta})\boldsymbol{x}_i,$$

respectively. Moreover, we impose some common conditions on the random covariates $\boldsymbol{x}$ and the random noise ε. In particular, we assume that the random covariates $\boldsymbol{x} \in \mathbb{R}^d$ are sub-exponential and that the random noise ε is heavy-tailed with finite second moment.

Condition 2. *Let $\boldsymbol{\Sigma} = \mathbb{E}(\boldsymbol{x}\boldsymbol{x}^{\mathrm{T}})$ be a positive definite matrix with $\lambda_u \geq \lambda_{\max}(\boldsymbol{\Sigma}) \geq \lambda_{\min}(\boldsymbol{\Sigma}) \geq \lambda_l > 0$ and assume that $\lambda_l = 1$ for simplicity. There exists $\nu_0 \geq 1$ such that $\mathbb{P}(|\boldsymbol{u}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1/2}\boldsymbol{x}| \geq \nu_0\|\boldsymbol{u}\|_2 \cdot t) \leq e^{-t}$ for all $t \in \mathbb{R}$ and $\boldsymbol{u} \in \mathbb{R}^d$. For notational convenience, let $\sigma_{\boldsymbol{x}}^2 = \max_{1\leq j\leq d}\sigma_{jj}$, where σ_{jj} is the j-th diagonal entry of $\boldsymbol{\Sigma}$.*

Condition 3. *The random noise ε has a finite second moment, i.e., $\mathbb{E}(\varepsilon^2|\boldsymbol{x}) \leq \sigma_\varepsilon^2 < \infty$. Moreover, the conditional τ-expectile of ε satisfies $\mathbb{E}[w_\tau(\varepsilon)\varepsilon|\boldsymbol{x}] = 0$, where $w_\tau(u) := |\tau - \mathbb{1}(u < 0)|$.*

Next, we provide nonasymptotic error bounds for the `retire` estimator, $\hat{\boldsymbol{\beta}}$, under the regime in which $n > d$ but d is allowed to diverge. Moreover, we establish a nonasymptotic Bahadur representation for $\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*$, based on which we construct a Berry-Esseen bound for a normal approximation. As mentioned in Section 1.4, the robustification bias $||\boldsymbol{\beta}_\gamma^* - \boldsymbol{\beta}^*||_2$ is inevitable due to the asymmetry nature of error term ε. Let $\underline{\tau} = \min(\tau, 1-\tau)$, $\bar{\tau} = \max(\tau, 1-\tau)$, and $A_1 \geq 1$ be a constant satisfying $\mathbb{E}(\boldsymbol{u}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1/2}\boldsymbol{x})^4 \leq A_1^4\|\boldsymbol{u}\|_2^4$ for all $\boldsymbol{u} \in \mathbb{R}^d$. The following proposition reveals the fact that the robustification bias scales at the rate γ^{-1}, which decays as γ grows.

Proposition 1.6.1. *Assume Conditions 1, 2, and 3 hold. Provided that $\gamma \geq 2\sigma_\varepsilon A_1^2\bar{\tau}/\underline{\tau}$, we have*

$$\|\boldsymbol{\Sigma}^{1/2}(\boldsymbol{\beta}_\gamma^* - \boldsymbol{\beta}^*)\|_2 \leq 2\gamma^{-1}(\bar{\tau}/\underline{\tau})\sigma_\varepsilon^2.$$

The key to our subsequent analysis for the `retire` estimator $\hat{\boldsymbol{\beta}}$ is the strong convexity property of the empirical loss function $\mathscr{R}_n(\cdot)$ uniformly over a local ellipsoid centered at $\boldsymbol{\beta}^*$ with

high probability. Let $\kappa_1 = \min_{|u|\le 1} \ell''(u)$, $\mathbb{B}_{\mathbf{\Sigma}}(r) = \{\boldsymbol{\delta} \in \mathbb{R}^d : \|\mathbf{\Sigma}^{1/2}\boldsymbol{\delta}\|_2 \le r\}$ be an ellipsoid. We characterize the strong convexity of $\mathscr{R}_n(\cdot)$ in Lemma 1.6.1. With the aid of Lemma 1.6.1, we establish a non-asymptotic error bound for the `retire` estimator $\hat{\boldsymbol{\beta}}$ in Theorem 1.6.1.

Lemma 1.6.1. *Let (γ, n) satisfy $\gamma \ge 4\sqrt{2}\max\{\sigma_\varepsilon, 2A_1^2 r\}$ and $n \gtrsim (\gamma/r)^2(d+t)$. Under Conditions 1, 2, and 3, with probability at least $1-e^{-t}$, we have*

$$\langle \nabla\mathscr{R}_n(\boldsymbol{\beta}) - \nabla\mathscr{R}_n(\boldsymbol{\beta}^*), \boldsymbol{\beta} - \boldsymbol{\beta}^* \rangle \ge \frac{1}{2}\kappa_1 \underline{\tau} \|\mathbf{\Sigma}^{1/2}(\boldsymbol{\beta} - \boldsymbol{\beta}^*)\|_2^2 \ \text{ uniformly over } \ \boldsymbol{\beta} \in \boldsymbol{\beta}^* + \mathbb{B}_{\mathbf{\Sigma}}(r).$$

Theorem 1.6.1. *Assume Conditions 1, 2, and 3 hold. For any $t > 0$, the `retire` estimator $\hat{\boldsymbol{\beta}}$ in* (1.6) *with $\gamma = \sigma_\varepsilon\sqrt{n/(d+t)}$ satisfies the bound*

$$\|\mathbf{\Sigma}^{1/2}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)\|_2 \le C(\bar{\tau}/\underline{\tau})\kappa_1^{-1}\sigma_\varepsilon v_0 \sqrt{\frac{d+t}{n}},$$

with probability at least $1-2e^{-t}$ as long as $n \gtrsim d+t$, where $C > 0$ is an absolute constant.

Theorem 1.6.1 shows that under the sub-exponential design with heavy-tailed random errors with bounded second moment, the `retire` estimator $\hat{\boldsymbol{\beta}}$ exhibits a sub-Gaussian type deviation bound, provided that the robustification parameter is properly chosen, i.e., $\gamma = \sigma_\varepsilon\sqrt{n/(d+t)}$. In other words, the proposed `retire` estimator gains robustness to heavy-tailed random noise without compromising statistical accuracy.

Remark 1.6.1. *The choice of $\gamma = \sigma_\varepsilon\sqrt{n/(d+t)}$ in Theorem 1.6.1 is a reflection of the bias and robustness trade-off for the `retire` estimator $\hat{\boldsymbol{\beta}}$. Intuitively, a large γ creates less robustification bias but sacrifices robustness. More specifically, we shall see from the proof of Theorem 1.6.1 that conditioning on the event $\{\hat{\boldsymbol{\beta}} \in \boldsymbol{\beta}^* + \mathbb{B}_{\mathbf{\Sigma}}(r_{\mathrm{loc}})\}$,*

$$\|\mathbf{\Sigma}^{1/2}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)\|_2 \lesssim \underbrace{\frac{\sigma_\varepsilon^2}{\gamma}}_{\text{robustification bias}} + \underbrace{\sigma_\varepsilon\sqrt{\frac{d}{n}} + \gamma\frac{d}{n}}_{\text{statistical error}}$$

with high probability. Therefore, we choose $\gamma = \sigma_\varepsilon \sqrt{n/(d+t)}$ *to minimize the right-hand side as a function of* γ.

To proceed, we establish nonasymptotic Bahadur representation for the difference $\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*$. To this end, we need slightly stronger conditions on both the random covariates $\boldsymbol{x}$ and the random noise ε. In particular, we require that the random covariates $\boldsymbol{x}$ to be sub-Gaussian and that the conditional density of the random noise ε is upper bounded. We formalize the above into the following conditions.

Condition 4. *There exists* $\nu_1 \geq 1$ *such that* $\mathbb{P}(|\boldsymbol{u}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1/2}\boldsymbol{x}| \geq \nu_1 ||\boldsymbol{u}||_2 t) \leq 2e^{-t^2/2}$ *for all* $t \in \mathbb{R}$ *and* $\boldsymbol{u} \in \mathbb{R}^d$.

Condition 5. *Let* $f_{\varepsilon|\boldsymbol{x}}(\cdot)$ *be the conditional density function of the random noise* ε. *There exists* $\bar{f}_{\varepsilon|\boldsymbol{x}} > 0$ *such that* $\sup_{u\in\mathbb{R}} f_{\varepsilon|\boldsymbol{x}}(u) \leq \bar{f}_{\varepsilon|\boldsymbol{x}}$ *almost surely (for all* $\boldsymbol{x}$*).*

Recall that $w_\tau(u) = |\tau - \mathbb{1}(u<0)|$ and that $\zeta(u) = L'_{\tau,\gamma}(u) = w_\tau(u)\ell'_\gamma(u)$. Moreover, let $\mathbf{J} = \mathbb{E}\{w_\tau(\varepsilon)\boldsymbol{x}\boldsymbol{x}^{\mathrm{T}}\}$ be the Hessian matrix. Theorem 1.6.2 establishes the Bahadur representation of the `retire` estimator $\hat{\boldsymbol{\beta}}$. Specifically, we show that the remainder of the Bahadur representation for $\hat{\boldsymbol{\beta}}$ exhibits sub-exponential tails, which we will use to establish the Berry-Esseen bound for linear projections of $\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*$ in Theorem 1.6.3.

Theorem 1.6.2. *Assume Conditions 1, 3, 4, and 5 hold. For any* $t > 0$, *the* `retire` *estimator* $\hat{\boldsymbol{\beta}}$ *given in* (1.6) *with* $\gamma = \sigma_\varepsilon \sqrt{n/(d+t)}$ *satisfies the following nonasymptotic Bahadur representation*

$$\left\| \boldsymbol{\Sigma}^{-1/2}\mathbf{J}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*) - \frac{1}{n}\sum_{i=1}^{n} \zeta(\varepsilon_i)\boldsymbol{\Sigma}^{-1/2}\boldsymbol{x}_i \right\|_2 \leq C\sigma_\varepsilon \cdot \frac{d+t}{n} \tag{1.8}$$

with probability at least $1 - 3e^{-t}$ *as long as* $n \gtrsim d + t$, *where* $C > 0$ *is a constant independent of* (n, d) *and* t.

Theorem 1.6.3. *Under the same set of conditions as in Theorem 1.6.2, assume further that* $\mathbb{E}(|\varepsilon|^3|\boldsymbol{x}) \le v_3 < \infty$ *(almost surely). Then, the* `retire` *estimator* $\hat{\boldsymbol{\beta}}$ *in* (1.6) *with robustness parameter* $\gamma = \sigma_\varepsilon \sqrt{n/(d+\log n)}$ *satisfies*

$$\sup_{\boldsymbol{u}\in\mathbb{R}^d, z\in\mathbb{R}} \left|\mathbb{P}(n^{1/2}\langle \boldsymbol{u}, \hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\rangle \le \sigma z) - \Phi(z)\right| \lesssim \frac{d+\log n}{\sqrt{n}},$$

where $\sigma^2 = \sigma^2(\boldsymbol{u}) := \boldsymbol{u}^{\mathrm{T}}\mathbf{J}^{-1}\mathbb{E}\{\zeta^2(\varepsilon)\boldsymbol{x}\boldsymbol{x}^{\mathrm{T}}\}\mathbf{J}^{-1}\boldsymbol{u}$ *and* $\Phi(\cdot)$ *is the standard normal cumulative distribution function.*

Theorem 1.6.3 shows that with a diverging parameter $\gamma = \sigma_\varepsilon\sqrt{n/(d+\log n)}$, for any $\boldsymbol{u} \in \mathbb{R}^d$, the linear projection of $\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*$ is asymptotically normal after some standardization as long as (n,d) satisfies the scaling condition $d = o(\sqrt{n})$.

1.7 Numerical Experiments

We evaluate the performance of the proposed `retire` estimator (1.6) *via* numerical studies. For all of the numerical studies, we generate the covariates $\boldsymbol{x}_i$ from a multivariate normal distribution $N(\mathbf{0}, \boldsymbol{\Sigma} = (\sigma_{jk})_{1\le j,k\le d})$ with $\sigma_{jk} = 0.5^{|j-k|}$. We then generate the response variable y_i from one of the following three models:

1. Homoscedastic model:

$$y_i = \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}^* + \varepsilon_i, \tag{1.9}$$

2. Quantile heteroscedastic model:

$$y_i = \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}^* + (0.5|x_{id}| + 0.5)\{\varepsilon_i - F_{\varepsilon_i}^{-1}(\tau)\}, \tag{1.10}$$

3. Expectile heteroscedastic model:

$$y_i = \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}^* + (0.5|x_{id}| + 0.5)\{\varepsilon_i - e_\tau(\varepsilon_i)\}, \tag{1.11}$$

where ε_i is the random noise, $F_{\varepsilon_i}^{-1}(\cdot)$ denotes the inverse cumulative distribution function of ε_i, and $e_\tau(\varepsilon_i)$ denotes the inverse of the expectile function of ε_i. Note that under Gaussian and t-distributed noises, the two models (1.11) and (1.10) are the same for $\tau = 0.5$. We set the regression coefficient vector $\boldsymbol{\beta}^* = (\beta_1^*, \beta_2^*, \ldots, \beta_d^*)^{\mathrm{T}}$ as $\beta_1^* = 2$ (intercept), $\beta_j^* = \{1.8, 1.6, 1.4, 1.2, 1, -1, -1.2, -1.4, -1.6, -1.8\}$ for $j = 2, 3, \ldots, 11$. The random noise is generated from either a Gaussian distribution, $N(0,2)$, or a t distribution with 2.1 degrees of freedom. For the heteroscedastic models, we consider two quantile/expectile levels $\tau = \{0.5, 0.8\}$.

We propose to select γ using a heuristic tuning method that involves updating γ at the beginning of each iteration in Algorithm 1. More specifically, let $r_i^k = y_i - \boldsymbol{x}_i^{\mathrm{T}} \hat{\boldsymbol{\beta}}^{k-1}, i = 1, \ldots, n$ be the residuals, where $\hat{\boldsymbol{\beta}}^{k-1}$ is obtained from the $(k-1)$-th iteration of Algorithm 1. Let $\tilde{r}_i^k = (1-\tau) r_i^k \mathbb{1}_{r_i^k \le 0} + \tau r_i^k \mathbb{1}_{r_i^k > 0}$ be the asymmetric residuals, and let $\tilde{\mathbf{r}}^k = (\tilde{r}_1^k, \ldots, \tilde{r}_n^k)^{\mathrm{T}}$. We define $\mathrm{mad}(\tilde{\mathbf{r}}^k) = \{\Phi^{-1}(0.75)\}^{-1} \mathrm{median}(|\tilde{\mathbf{r}}^k - \mathrm{median}(\tilde{\mathbf{r}}^k)|)$ as the median absolute deviation of the asymmetric residuals, adjusted by a factor $\Phi^{-1}(0.75)$, where $\Phi(\cdot)$ is the cumulative distribution function of standard normal distribution. We start with setting $\gamma = \sqrt{n/(d + \log n)}$. At the k-th iteration of Algorithm 1, we update the robustification parameter by

$$\gamma^k = \mathrm{mad}(\tilde{\mathbf{r}}^k) \cdot \sqrt{\frac{n}{d + \log n}}.$$

Throughout our numerical studies, we have found that γ chosen using the above heuristic approach works well across different scenarios. Our computational results are reproducible using codes available from https://github.com/ZianWang0128/Retire.

1.7.1 Estimation

In this subsection, we compare `retire` to three other competitive methods: (i) Huber regression (`huber`); (ii) asymmetric least squares regression (`als`), and (iii) quantile regression (`qr`) implemented via the `R` package `quantreg`. To assess the performance across different methods, we report the estimation error under the ℓ_2-norm, i.e., $\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2$ and its standard

errors.

Note that `huber` and `als` are special cases of `retire` by taking $\tau = 0.5$ and $\gamma \to \infty$, respectively. Thus, both `huber` and `als` can be implemented via Algorithm 1. Also note that both `huber` and `retire` require tuning an additional robustness parameter γ, which is tuned by the aforementioned heuristic approach. The results, averaged over 1000 repetitions, are reported in Table 1.1 for three low-dimensional settings ($n = 200/400/800,\ d = 10$).

Table 1.1. Estimation error under ℓ_2-norm (and its standard errors) are reported, averaged over 1000 repetitions.

		Homo-model (1.9)		Quantile hetero-model (1.10)				Expectile hetero-model (1.11)	
	Method	$\tau = 0.5$		$\tau = 0.5$		$\tau = 0.8$		$\tau = 0.8$	
		$\varepsilon \sim N(0,2)$	$\varepsilon \sim t_{2.1}$	$\varepsilon \sim N(0,2)$	$\varepsilon \sim t_{2.1}$	$\varepsilon \sim N(0,2)$	$\varepsilon \sim t_{2.1}$	$\varepsilon \sim N(0,2)$	$\varepsilon \sim t_{2.1}$
$n = 200$ $d = 10$	retire	0.421 (0.108)	0.427 (0.117)	0.387 (0.098)	0.386 (0.105)	0.498 (0.105)	0.515 (0.126)	0.457 (0.114)	0.508 (0.129)
	huber	0.421 (0.108)	0.427 (0.117)	0.387 (0.098)	0.386 (0.105)	1.107 (0.101)	0.995 (0.099)	0.779 (0.099)	0.964 (0.098)
	als	0.416 (0.108)	0.767 (0.521)	0.402 (0.101)	0.725 (0.487)	0.574 (0.102)	1.051 (1.171)	0.433 (0.107)	1.049 (1.175)
	qr	0.509 (0.125)	0.430 (0.113)	0.438 (0.109)	0.368 (0.100)	0.500 (0.128)	0.539 (0.159)	0.603 (0.140)	0.542 (0.162)
$n = 400$ $d = 10$	retire	0.287 (0.073)	0.319 (0.080)	0.272 (0.068)	0.287 (0.073)	0.400 (0.072)	0.385 (0.083)	0.305 (0.076)	0.373 (0.084)
	huber	0.287 (0.073)	0.319 (0.080)	0.272 (0.068)	0.287 (0.073)	1.087 (0.070)	0.968 (0.076)	0.738 (0.069)	0.936 (0.075)
	als	0.287 (0.073)	0.569 (0.441)	0.279 (0.070)	0.543 (0.408)	0.477 (0.073)	0.799 (0.919)	0.301 (0.076)	0.797 (0.922)
	qr	0.360 (0.088)	0.292 (0.075)	0.305 (0.074)	0.248 (0.064)	0.353 (0.090)	0.372 (0.099)	0.493 (0.098)	0.375 (0.103)
$n = 800$ $d = 10$	retire	0.199 (0.050)	0.240 (0.060)	0.193 (0.048)	0.218 (0.055)	0.375 (0.051)	0.301 (0.058)	0.213 (0.053)	0.285 (0.059)
	huber	0.199 (0.050)	0.240 (0.060)	0.193 (0.048)	0.218 (0.055)	1.083 (0.049)	0.957 (0.054)	0.722 (0.048)	0.924 (0.054)
	als	0.199 (0.050)	0.404 (0.219)	0.195 (0.049)	0.389 (0.200)	0.428 (0.051)	0.572 (0.354)	0.213 (0.054)	0.570 (0.358)
	qr	0.253 (0.063)	0.202 (0.050)	0.214 (0.052)	0.170 (0.041)	0.243 (0.060)	0.258 (0.065)	0.420 (0.061)	0.261 (0.068)

Table 1.1 shows the results for all methods under various models. Generally, the robustified expectile method `retire` has the smallest estimation errors across all settings. For the Huber regression method `huber`, its performance deteriorates when $\tau = 0.8$. This is not surprising since `huber` implicitly assumes $\tau = 0.5$, thus non-negligible bias is introduced when $\tau = 0.8$. For the asymmetric least-square method `als`, (comparing with `retire`) it has similar performance under Gaussian noises, but worse performance under $t_{2.1}$ noises, indicating that the robustness of `retire` loss comparing to asymmetric least-square loss is beneficial especially under heavy-tailed noises. The quantile regression method `qr` performs the best under $t_{2.1}$ noises, but it loses its advantages under Gaussian noises. This may due to the fact that quantile loss (check loss) is the most robust to outliers among the other losses, while losing the sensitivity for small variations.

In summary the numerical studies suggest that `retire` is a robust alternative to its least

squares counterpart `als` and a flexible extension from `huber` to accommodate asymmetry, while maintaining sensitivity for light-tailed noises comparing to the quantile regression approach `qr`.

1.7.2 Inference for Confidence Intervals

In this subsuction, we apply Multiple Bootstrap (MB) technique to obtain confidence intervals for signals β_j^*. Here we briefly outline the Multiple Bootstrap (MB) procedure to obtain confidence intervals for signals β_j^* under low-dimensional settings.

Consider data $\{(y_i, \boldsymbol{x}_i)\}_{i=1}^n$ that follows the data generation process detailed in Section 1.7, and recall that $L_{\tau,\gamma}$ is the loss function defined in (1.5) that satisfies Condition 1. Let $w_1^b, \ldots, w_n^b$ be i.i.d. random bootstrap weights that satisfy $\mathbb{E}(w_i^b) = \mathrm{var}(w_i^b) = 1$. For convenience, we focus on the Huber loss for which $\ell(u) = u^2/2 \cdot \mathbb{1}(|u| \le 1) + (|u| - 1/2) \cdot \mathbb{1}(|u| > 1)$, and choose exponential i.i.d. bootstrap weights, i.e., $w_i^b \sim \exp(1)$.

First we compute an initial `retire` estimator, denoted as $\hat{\boldsymbol{\beta}}^{ini}$, by minimizing the objective function $n^{-1}\sum_{i=1}^n L_{\tau,\gamma}(y_i - \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta})$. Next, we obtain low-dimensional bootstrap samples $\{\hat{\boldsymbol{\beta}}_1^{boot}, \ldots, \hat{\boldsymbol{\beta}}_B^{boot}\}$ by repeatedly minimizing randomly weighted objective functions $n^{-1}\sum_{i=1}^n w_i^b \cdot L_{\tau,\gamma}(y_i - \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta})$, based on which we construct confidence intervals. More specifically, we can construct MB confidence intervals for all the slope coefficients using one of the three classical methods, the percentile method, the pivotal method, and the normal-based method. Let $\alpha \in (0,1)$ be a prespecified confidence level.

1. Efron's percentile method: For each $q \in (0,1)$ and $2 \le j \le d$, define the conditional q-quantile of $\hat{\beta}_j^{boot}$ given the observed data as

$$c_j^b(q) = \inf\{t \in \mathbb{R} : \mathbb{P}^*(\hat{\beta}_j^{boot} \le t) \ge q\}.$$

Then then Efron's percentile interval for β_j^* takes the form

$$\Big[c_j^b(\alpha/2), c_j^b(1-\alpha/2)\Big]. \tag{1.12}$$

2. Pivotal method: The pivotal interval approximates the conditional distribution of $\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*$ by the bootstrap quantity $\hat{\boldsymbol{\beta}}^{boot} - \hat{\boldsymbol{\beta}}$. More specifically, the pivotal confidence interval for β_j^* takes the form

$$\Big[2\hat{\beta}_j^{ini} - c_j^b(1-\alpha/2), 2\hat{\beta}_j^{ini} - c_j^b(\alpha/2)\Big]. \tag{1.13}$$

Pivotal confidence intervals are connected to percentile confidence intervals in sense that the latter are the pivotal confidence intervals reflected about the point $\hat{\beta}_j^{boot}$.

3. Normal-based method: Let $\Phi^{-1}(\cdot)$ be the inverse of the cumulative distribution function of a standard normal random variable. Denote $\mathrm{std}(\cdot)$ as the sample standard deviation. Then the normal-based confidence interval for β_j^* takes the form

$$\Big[\hat{\beta}_j^{ini} - \Phi^{-1}(1-\alpha/2)\cdot \mathrm{std}\big\{\hat{\beta}_{\cdot,j}^{boot}\big\}, \hat{\beta}_j^{ini} + \Phi^{-1}(1-\alpha/2)\cdot \mathrm{std}\big\{\hat{\beta}_{\cdot,j}^{boot}\big\}\Big]. \tag{1.14}$$

We summarize the whole process as follow:

Procedure 2. Multiple Bootstrap Inference for β^* under low-dimensional settings.

Input: generated data $\{(y_i, \boldsymbol{x}_i)\}_{i=1}^n$, bootstrap weights w_i, Huber loss tuning parameter γ.
Initialization: $B = 200$.

1. Compute an initializer $\hat{\boldsymbol{\beta}}^{ini}$ based on the dataset $\{(y_i, \boldsymbol{x}_i)\}_{i=1}^n$ by solving

$$\hat{\boldsymbol{\beta}}^{ini} \in \underset{\boldsymbol{\beta}\in\mathbb{R}^d}{\text{minimize}} \frac{1}{n}\sum_{i=1}^n L_{\tau,\gamma}(y_i - \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta})$$

2. For $b = 1, \ldots, B$, obtain low-dimensional bootstrapped estimators by solving

$$\hat{\boldsymbol{\beta}}_b^{boot} \in \underset{\boldsymbol{\beta}\in\mathbb{R}^d}{\text{minimize}} \frac{1}{n}\sum_{i=1}^n w_i^b \cdot L_{\tau,\gamma}(y_i - \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}).$$

3. Calculate confidence intervals from $\{\hat{\boldsymbol{\beta}}_1^{boot}, \ldots, \hat{\boldsymbol{\beta}}_B^{boot}\}$ by (1.12), (1.13) and (1.14).

Output: Multiple bootstrap confidence intervals.

For low-dimensional confidence interval inference, we follow the same data generation

process, model settings, and hyper-parameter selection methods as Section 1.7. And we use the approximate 95% confidence interval (1.7) derived from asymptotic normality as the benchmark. All inference results, averaged over 1000 repetitions, are reported in the following table.

Table 1.2. Inference results for low-dimensional settings. Coverage rate (and the width of confidence intervals) are reported, averaged over 1000 repetitions.

		Homo-model (2.7)		Quantile hetero-model (2.8)				Expectile hetero-model (2.9)	
	CI type	$\tau = 0.5$		$\tau = 0.5$		$\tau = 0.8$		$\tau = 0.8$	
		$\varepsilon \sim N(0,2)$	$\varepsilon \sim t_{2.1}$	$\varepsilon \sim N(0,2)$	$\varepsilon \sim t_{2.1}$	$\varepsilon \sim N(0,2)$	$\varepsilon \sim t_{2.1}$	$\varepsilon \sim N(0,2)$	$\varepsilon \sim t_{2.1}$
$n = 200$ $d = 10$	Percentile	0.924 (0.473)	0.922 (0.574)	0.921 (0.453)	0.923 (0.541)	0.916 (0.461)	0.912 (0.678)	0.911 (0.466)	0.912 (0.679)
	Pivotal	0.916 (0.473)	0.942 (0.574)	0.915 (0.453)	0.940 (0.541)	0.906 (0.461)	0.925 (0.678)	0.899 (0.466)	0.925 (0.679)
	MB-normal	0.926 (0.484)	0.942 (0.591)	0.922 (0.465)	0.941 (0.549)	0.920 (0.469)	0.930 (0.687)	0.916 (0.475)	0.929 (0.687)
	Normal	0.930 (0.491)	0.919 (0.537)	0.925 (0.466)	0.920 (0.508)	0.924 (0.476)	0.902 (0.629)	0.923 (0.484)	0.902 (0.630)
$n = 400$ $d = 10$	Percentile	0.933 (0.337)	0.934 (0.429)	0.931 (0.326)	0.932 (0.407)	0.923 (0.336)	0.914 (0.526)	0.924 (0.343)	0.914 (0.527)
	Pivotal	0.935 (0.337)	0.944 (0.429)	0.935 (0.326)	0.945 (0.407)	0.921 (0.336)	0.930 (0.526)	0.923 (0.343)	0.930 (0.527)
	MB-normal	0.941 (0.343)	0.947 (0.436)	0.942 (0.332)	0.946 (0.413)	0.930 (0.342)	0.934 (0.532)	0.932 (0.349)	0.934 (0.533)
	Normal	0.946 (0.349)	0.939 (0.419)	0.945 (0.337)	0.942 (0.398)	0.939 (0.351)	0.927 (0.517)	0.938 (0.359)	0.926 (0.518)
$n = 800$ $d = 10$	Percentile	0.938 (0.239)	0.932 (0.324)	0.935 (0.233)	0.934 (0.309)	0.929 (0.243)	0.924 (0.419)	0.927 (0.248)	0.924 (0.420)
	Pivotal	0.937 (0.239)	0.941 (0.324)	0.934 (0.233)	0.942 (0.309)	0.926 (0.243)	0.937 (0.419)	0.927 (0.248)	0.937 (0.420)
	MB-normal	0.945 (0.244)	0.945 (0.329)	0.943 (0.238)	0.944 (0.314)	0.937 (0.248)	0.938 (0.424)	0.935 (0.253)	0.939 (0.425)
	Normal	0.948 (0.247)	0.941 (0.324)	0.946 (0.240)	0.943 (0.309)	0.943 (0.253)	0.939 (0.421)	0.942 (0.259)	0.939 (0.422)

From Table 1.2 we see that all confidence intervals perform fairly well, higher coverage and narrower width are observed as the dimension n increases. All three types of bootstrap confidence intervals perform similarly, while the benchmark confidence interval derived from asymptotic normality tends to perform slightly better under normal noises.

1.7.3 Data Application: Job Training Partners Act Data

In this subsection, we analyze the Job Training Partners Act (JTPA) data, previously studied in Abadie, Angrist and Imbens (2002), using the `retire` estimator proposed in Section 1.4. The JTPA began funding federal training programs in 1983, and its largest component Title II supports training for the economically disadvantaged. Specifically, applicants who faced "barriers to employment", the most common of which were high-school dropout status and long periods of unemployment, were typically considered eligible for JTPA training. The services offered as a part of training included classroom training, basic education, on-the-job training, job search assistance, and probationary employment.

In this data set, applicants who applied for training evaluation between November 1987

and September 1989 were randomly selected to enroll for the JTPA training program. Of the 6,102 adult women in the study, 4,088 were offered training and 2,722 enrolled in the JTPA services, and of the 5,102 adult men in the study, 3,399 were offered training and 2,136 enrolled in the services. The goal is to assess the effect of subsidized training program on earnings. Motivated by Abadie, Angrist and Imbens (2002), we use the 30-month earnings data collected from the Title II JTPA training evaluation study as the response variable. Moreover, we consider the following covariates: (1) whether or not the individual enrolled in the JTPA services (yes=1, no=0), (2) individual's sex (male=1, female=0), (3) whether or not the individual graduated high school or obtained a GED (yes=1, no=0), (4) whether or not the individual is black (yes=1, no=0), (5) whether or not the individual is Hispanic (yes=1, no=0), (6) marriage status (married=1, not married=0) and (7) whether or not the individual worked less than 13 weeks in the 12 months preceding random assignment (yes=1, no=0). We study the conditional distribution of 30-month earnings at different expectile levels $\tau = \{0.1, 0.5, 0.9\}$. Our proposed method involves robustification parameter γ, which we select using the tuning method described in Section 1.7.

The regression coefficients and their associated 95% confidence intervals are shown in Table 1.3. We find that covariates with positive regression coefficients for all quantile levels are enrollment for JTPA services, individual's sex, high school graduation or GED status, and marriage status. Black, hispanic, and worked less than 13 weeks in the past year had negative regression coefficients. The regression coefficients varied across the three different expectile levels we considered. The positive regression coefficients increase as the τ level increases and the negative regression coefficients decrease as the τ level increases. That is, for the lower expectile level of 30-month earnings, the covariates have a smaller in magnitude effect on the individual's earnings compared to the higher expectile level. The regression coefficient for enrollment in JTPA services was 1685.34, 2637.57, and 2714.57 at $\tau = \{0.1, 0.5, 0.9\}$, respectively. The τ-expectile of 30-month earnings for $\tau = \{0.1, 0.5, 0.9\}$ is 5068.02, 15815.29, and 32754.89 dollars, respectively. Compared to the expectile at the given τ, the (relative) effect of subsidized

training was larger for lower expectile levels. Notably, if an individual is a male, conditional on other covariates, their 30-month earnings increase by 5,005 dollars for $\tau = 0.5$ and increase by 10,311 dollars for $\tau = 0.9$. From the confidence intervals, we see that all variables are statistically significant except Hispanic.

Table 1.3. Regression coefficients (and their associated 95% confidence intervals) for the `retire` estimator.

Variable	$\tau = 0.1$	$\tau = 0.5$	$\tau = 0.9$
enrolled in services	1685.34 (1401.03, 1969.65)	2637.57 (2079.74, 3195.40)	2714.57 (1766.01, 3663.13)
male	1706.87 (1435.04, 1978.69)	5005.12 (4449.07, 5561.17)	10310.62 (9338.91, 11282.34)
high school or GED	1477.19 (1218.33, 1736.06)	3656.13 (3140.12, 4172.14)	5718.62 (4803.60, 6633.63)
black	-580.04 (-917.86, -242.21)	-1567.03 (-2265.51, -868.56)	-2459.81 (-3686.14, -1233.48)
hispanic	-130.72 (-588.11, 326.66)	-669.76 (-1626.83, 287.32)	-1495.33 (-3306.12, 315.46)
married	1268.30 (933.66, 1602.94)	3343.63 (2668.95, 4018.30)	4518.43 (3376.92, 5659.93)
worked less than 13 wks	-3677.98 (-3957.24, -3398.72)	-6879.14 (-7438.20, -6320.08)	-8206.16 (-9151.81, -7260.50)

Chapter 2

Expectile Regression in High Dimensions

2.1 Motivation and Overview

In this chapter, we focus on high-dimensional regression models in which the number of covariates, d, is considerably larger than the number of observations, n. Recall that the goal is to infer the conditional distribution of the response variable y given the covariates $\boldsymbol{x}$ based on the training data $(y_1, \boldsymbol{x}_1), \ldots, (y_n, \boldsymbol{x}_n) \in \mathbb{R} \times \mathbb{R}^d$. For high-dimensional settings, ordinary (unpenalized) methods are generally inconsistent due to high dimensions, and the mainstream approach is to use penalization techniques to perform variable selection and estimation simultaneously.

The most intuitive method to explore conditional distribution in high dimensions is the sparse quantile regression (QR). Belloni and Chernozhukov (2011) studied quantile regression with ℓ_1-penalization in order to remove covariates whose population coefficients are zero. They established the uniform (over a range of quantile levels) convergence rate of $\sqrt{s\log(d \vee n)/n}$, where s is the sparsity level—cardinality of the true active set $\mathscr{S} = \mathrm{supp}(\boldsymbol{\beta}^*)$. To alleviate the bias induced by the ℓ_1 penalty, Wang, Wu and Li (2012) proposed concave penalized quantile regression, and showed that the oracle estimator is a local solution to the resulting optimization problem. Via the one-step local linear approximation (LLA) algorithm, Fan, Xue and Zou (2014) proved that the oracle estimator can be obtained (with high probability) as long as the magnitude of true nonzero regression coefficients is at least of order $\sqrt{s\log(d)/n}$. We refer to Wang and He (2022) for a unified analysis of global and local optima of penalized quantile regressions. While

quantile regression offers the flexibility to model the conditional response distribution and is robust to outliers, together the non-differentiability of the check function and the non-convexity of the penalty pose substantial technical and computational challenges. To our knowledge, the theoretical guarantee of the convergence of a computationally efficient algorithm to the oracle QR estimator under the weak minimum signal strength condition—$\min_{j\in\mathscr{S}}|\beta_j^*| \gtrsim \sqrt{\log(d)/n}$ with $\mathscr{S} = \text{supp}(\boldsymbol{\beta}^*)$—remains unclear.

An alternative method for analyzing the conditional distribution in high dimensions is the penalized expectile regression. Gu and Zou (2016) considered the penalized expectile regression using both convex and concave penalty functions. Since the expectile loss is convex and twice-differentiable, scalable algorithms, such as the cyclic coordinate decent and proximal gradient descent, can be employed to solve the resulting optimization problem. Theoretically, the consistency of penalized expectile regression in the high-dimensional regime "$\log(d) \ll n \ll d$" requires *sub-Gaussian* error distributions (Gu and Zou, 2016). This is in strong contrast to penalized QR, the consistency of which requires no moment condition (Belloni and Chernozhukov, 2011, Wang and He, 2022) although certain regularity conditions on the conditional density function are still needed. Lack of robustness to heavy-tailedness for expectile regression is also observed in numerical studies. Since expectile regression is primarily introduced to explore the tail behavior of the conditional response distribution, its sensitivity to the tails of the error distributions, particularly in the presence of high-dimensional covariates, raises a major concern from a robustness viewpoint.

Therefore, we aim to shrink the gap between quantile and expectile regressions, specifically in high dimensions, by proposing a penalized robust expectile regression (penalized-`retire`) method that inherits the computational expediency and statistical efficiency of expectile regression and is nearly as robust as quantile regression against heavy-tailed response distributions. The main idea, which is adapted from Sun, Zhou and Fan (2020), is to replace the asymmetric squared loss associated with expectile regression with a Lipschitz and locally quadratic robust alternative, parameterized by a data-dependent parameter to achieve a desirable

trade-off between bias and robustness.

The class of robustified expectile (`retire`) loss is already detailed in Section 1.4, and we will discuss various penalty techniques in the following section.

2.2 Penalization Techniques

For high-dimensional regression models in which the number of covariates, d, is considerably larger than the number of observations, n, our goal is to infer the conditional distribution of the response variable y given the covariates $\boldsymbol{x}$ based on the training data $(y_1, \boldsymbol{x}_1), \ldots, (y_n, \boldsymbol{x}_n) \in \mathbb{R} \times \mathbb{R}^d$. Meanwhile, by the nature of high dimensionality, some sort of low-dimensional structures are inevitable for the regression problem to be solvable. We impose strict sparsity assumption on high-dimensional data—only a small number of significant predictors are associated with the response.

When fitting high-dimensional regression models, we aim to both select the significant variables and estimate their coefficients correctly. One famous approach is the stepwise selection, which first selects a subset of important variables by various criteria, and then performs a ordinary least squares regression on the selected variables. Albeit being logically intuitive and practically useful, stepwise selection is generally computationally expensive, and it ignores stochastic errors inherited in the stages of variable selections. Hence, its theoretical properties are hard to understand.

Another famous approach is the penalization or regularization technique in hope to perform variable selection and coefficient estimation simultaneously. To this end, we may use various convex and non-convex penalty functions so as to achieve a desirable trade-off between model complexity and statistical accuracy (Bühlmann and van de Geer, 2011, Wainwright, 2019, Fan *et al.*, 2020). For example, the ridge regression proposed by Hoerl, A.E. and Kennard, R.W. (1970) and the least absolute shrinkage and selection operator (Lasso) proposed by Tibshirani (1996). However, ridge regression rarely introduces sparsity despite being computationally

efficient, and Lasso introduces sparsity at the cost of inducing non-negligible bias, which may prevent consistent variable selection. To see this, consider the case of orthogonal design where all the vectors $\boldsymbol{z}_i$ are orthogonal to each other. It can be shown that adding ℓ_1 penalty results in coordinate-wise soft thresholdings $\hat{\beta}_j = \text{sign}(z_j)(z_j - \lambda)_+$. Therefore, small coefficients are shrunk to zero to provide sparsity, and large coefficients are shrunk towards zero to induce non-negligible bias. Intuitively, the ℓ_1 penalty $||\boldsymbol{\beta}||_1$ penalizes the true parameter $\boldsymbol{\beta}^*$ differently. Large coefficients are penalized much more than small coefficients, thus inducing non-negligible bias. Consequently, the selected model with a relatively small prediction error tends to include many false positives, unless stringent assumptions are imposed on the design matrix (Zhang and Zhang, 2012, Zou and Li, 2008, Su, Bogdan and Candés, 2017, Lahiri, 2021).

To construct a penalty that performs variable selection and coefficient estimation simultaneously while inducing as least bias as possible, Fan and Li (2001) considered the folded concave penalties. They showed that such penalty functions have to, (i) be singular at the origin to produce sparse solutions whose estimated coefficients are mostly zero; (ii) be bounded by a constant to produce nearly unbiased estimates for large coefficients; (iii) be symmetric, continuously differentiable and non-convex over $(0, \infty)$ to produce continuous models so that the variable selection process is stable. In summary, they proposed the smoothly clipped absolute deviation (SCAD) penalty $p_\lambda(\theta)$ where $p'_\lambda(\theta) = \lambda\{\mathbb{1}(\theta \leq \lambda) + \frac{(a\lambda - \theta)_+}{(a-1)\lambda}\mathbb{1}(\theta > \lambda)\}$ for some $a > 2$, usually $a = 3.7$ from a Bayes risk perspective. Other widely used nonconvex penalties like the capped-ℓ_1 penalty, the minimax concave penalty (MCP) (Zhang, 2010a) follows the similar principle. It can be shown that with proper choice of regularization parameters λ, the proposed estimators perform as well as the oracle procedure in variable selection, i.e., they work as well as if the correct submodel were known.

Next, we discuss from the computational perspective. Even though the folded concave penalized estimator is proven to achieve the oracle property provided the signals are sufficiently strong, i.e., the estimator has the same rate of convergence as that of the oracle estimator obtained by fitting the regression model with true active predictors that are unknown in practice, the

singularity and nonconvexity of the penalty function challenge us computationally. Directly minimizing the concave penalized loss raises numerical instabilities, and standard gradient-based algorithms are often guaranteed to find a stationary point, while oracle results are primarily derived for the hypothetical global minimum. To overcome the aforementioned challenges, Zou and Li (2008) proposed a unified algorithm for folded-concave penalized estimation based on local linear approximation (LLA). It relaxes the non-convex optimization problem into a sequence of iteratively reweighted ℓ_1-penalized subproblems. It is also shown that LLA is the best convex majorization–minimization (MM) algorithm, thus proving the convergence of the LLA algorithm by the ascent property of MM algorithms.

Furthermore, Fan, Xue and Zou (2014) pointed out that, although the sequence of LLA estimators are guaranteed to converge to a local stationary point instead of the global minimizer, the computed local solution satisfies oracle property with high probability. More specifically, the probability that this specific local solution exactly equals the oracle estimator is lower-bounded by $1-\delta_0-\delta_1-\delta_2$, where δ_0 corresponds to the exception probability of the localizability of the underlying model, δ_1 and δ_2 represent the exception probabilities of the regularity of the oracle estimator, which are usually very small under weak regularity conditions, and irrelevant to the actual estimation method.

2.3 Penalized Retire: Penalized Robust Expectile Regression

In this section, we propose the penalized `retire` estimator for high-dimensional data with $d > n$, obtained by minimizing the robust loss $\frac{1}{n}\sum_{i=1}^{n} L_{\tau,\gamma}(y_i - \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta})$, plus a penalty function $p_\lambda(\cdot)$ that induces sparsity on the regression coefficients. As mentioned in Section 2.2, the non-negligible estimation bias introduced by convex penalties (e.g., the Lasso penalty) can be reduced by folded-concave regularization when the signals are sufficiently strong, that is, the minimum of magnitudes of all nonzero coefficients are away from zero to some extent. The latter, however, is computationally more challenging and unstable due to non-convexity.

Adapted from the local linear approximation algorithm proposed by Zou and Li (2008), we apply an iteratively reweighted ℓ_1-penalized algorithm for fitting sparse robust expectile regression models with the robust loss $L_{\tau,\gamma}(\cdot)$. At each iteration, the penalty weights depend on the previous iterate and the choice of a (folded) concave regularizer satisfying Condition 6 (Zhang and Zhang, 2012) below. Some popular examples include the smoothly clipped absolute deviation (SCAD) penalty (Fan and Li, 2001), the minimax concave penalty (Zhang, 2010a), and the capped-ℓ_1 penalty. We refer the reader to Zhang and Zhang (2012) and Section 4.4 of Fan *et al.* (2020) for more details.

Condition 6. *The penalty function p_λ ($\lambda > 0$) is of the form $p_\lambda(t) = \lambda^2 p_0(t/\lambda)$ for $t \geq 0$, where the function $p_0 : \mathbb{R}_+ \to \mathbb{R}_+$ satisfies: (i) $p_0(\cdot)$ is non-decreasing on $[0,\infty)$ with $p_0(0) = 0$; (ii) $p_0(\cdot)$ is differentiable almost everywhere on $(0,\infty)$ and $\lim_{t\downarrow 0} p_0'(t) = 1$; (iii) $p_0'(t_1) \leq p_0'(t_2)$ for all $t_1 \geq t_2 > 0$.*

Let $p_\lambda(\cdot)$ be a prespecified concave regularizer that satisfies Condition 6, and let $p_\lambda'(\cdot)$ be its first-order derivative. Starting at iteration 0 an initial estimate $\hat{\boldsymbol{\beta}}^{(0)}$, we sequentially solve the following weighted ℓ_1-penalized convex optimization problems:

$$\hat{\boldsymbol{\beta}}^{(t)} \in \underset{\boldsymbol{\beta}\in\mathbb{R}^d}{\text{minimize}} \left\{ \frac{1}{n}\sum_{i=1}^{n} L_{\tau,\gamma}(y_i - \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}) + \sum_{j=2}^{d} p_\lambda'(|\hat{\beta}_j^{(t-1)}|)|\beta_j| \right\}, \tag{2.1}$$

where $\hat{\boldsymbol{\beta}}^{(t)} = (\hat{\beta}_1^{(t)}, \ldots, \hat{\beta}_d^{(t)})^{\mathrm{T}}$. At each iteration, $\hat{\boldsymbol{\beta}}^{(t)}$ is a weighted ℓ_1-penalized robust expectile regression estimate, where the weight $p_\lambda'(|\hat{\beta}_j^{(t-1)}|)|\beta_j|$ can be viewed as a local linear approximation of the concave regularizer $p_\lambda(|\beta_j|)$ around $|\hat{\beta}_j^{(t-1)}|$. With the trivial initialization $\hat{\boldsymbol{\beta}}^{(0)} = \mathbf{0}$, the first optimization problem (2.1) (when $t = 1$) reduces to the ℓ_1-penalized robust expectile regression because $p_\lambda'(0) = \lambda$. This iterative procedure outputs a sequence of estimates $\hat{\boldsymbol{\beta}}^{(1)}, \ldots, \hat{\boldsymbol{\beta}}^{(T)}$, where the number of iterations T can either be set before running the algorithm or depend on a stopping criterion. Throughout this thesis, we refer to the sequence of estimates $\{\hat{\boldsymbol{\beta}}^{(t)}\}_{t=1,\ldots,T}$ given in (2.1) as the *iteratively reweighted ℓ_1-penalized retire* estimators. We will

describe the computation algorithm in Section 2.4, and characterize their statistical properties in Section 2.5, including the theoretical choice of T in order to obtain a statistically optimal estimator.

Remark 2.3.1. *Our work differs from Gu and Zou (2016) in three main aspects. Firstly, we extend the ordinary expectile loss to a class of robustified expectile loss which can handle heavy-tailed error/response distributions. Our proposed robustified expectile loss retains local strong convexity near the origin, behaves (at most) linearly when the input is large, and is differentiable everywhere. Therefore, gradient descent based algorithms can be employed to solve the resulting optimization problem while being robust to heavy-tailed distributions at the mean time. Secondly, we consider a much broader error distribution that could better suit the heavy-tailed distribution. We only require finite second moments and zero conditional $\tau-$expectile for the random error ε, i.e., $\mathbb{E}(\varepsilon^2|\boldsymbol{x}) \leq \sigma_\varepsilon^2 < \infty$ and $\mathbb{E}[w_\tau(\varepsilon)\varepsilon|\boldsymbol{x}] = 0$, where $w_\tau(u) := |\tau - \mathbb{1}(u < 0)|$. As a comparison, Gu and Zou (2016) requires a much more stringent i.i.d. sub-Gaussian condition and zero conditional $\tau-$expectile for the random error ε. Lastly, we work on random designs and imposes less conditions on the design matrix—the design vector $\boldsymbol{x}$ is sub-Exponential. In contrast, Gu and Zou (2016) considers the fixed design that requires the restricted eigenvalue condition and the generalized invertability factor (GIF) condition, both of which are difficult to verify in practice.*

It will be shown that with an appropriate choice of the robust parameters γ and penalty parameter λ, ℓ_1-penalized `retire` with bounded-second-moment noises satisfies sub-Gausssian deviation bounds with near-optimal convergence rate as if sub-Gaussian random noise were assumed in Gu and Zou (2016). Therefore, our work with less stringent conditions for broader scenarios can be viewed as an extension over the previous work.

2.4 Computational Methods

In this section we introduce the Local Adaptive Majorize-minimization (LAMM) algorithm (Fan et al., 2018) for solving the iteratively reweighted ℓ_1-penalized convex optimization

problem in (2.1). On a high level, the LAMM algorithm can be viewed as a high dimensional generalization of the majorize-minimization (MM) algorithm (Hunter, D. R. and Lange, K. , 2004) and the iterative shrinkage-thresholding algorithm (ISTA) (Beck and Teboulle, 2009). The main idea of the LAMM algorithm is to construct an isotropic quadratic objective function that locally majorizes the `retire` loss function, while permitting closed-form updates at each iteration. The quadratic coefficient used for local majorization is adaptively chosen in order to guarantee the decrease of the objective function.

Note that the penalized `retire` regression can be formulated as a linear programming or a second-order cone programming (SOCP) problem, depending on the type of sparsity-inducing penalties. Therefore, general-purpose optimization toolboxes (e.g. interior point methods) can be applied. However, such toolboxes are only adapted to small-scale problems and usually lead to solutions with high precision. For large-scale problems, they tend to be too slow or often run out of memory. As a contrast, the LAMM algorithm is a simpler gradient-based algorithm which is particularly suited for large-scale problems, and the dominant computational effort is a relatively cheap matrix-vector multiplication. The (local) strong convexity of the `retire` loss function facilitates the convergence of such a first order method. Moreover, the LAMM algorithm can be extended to a broad class of convex penalties, we will detail this in Chapter 3.

To elaborate the main idea of the LAMM algorithm, consider the minimization of a general smooth non-linear function $f(\boldsymbol{\beta})$. The LAMM algorithm locally majorizes $f(\boldsymbol{\beta})$ at $\hat{\boldsymbol{\beta}}^{(k-1)}$ by a properly constructed function $g(\boldsymbol{\beta}|\hat{\boldsymbol{\beta}}^{(k-1)})$ that satisfies the following local property

$$f(\hat{\boldsymbol{\beta}}^{(k)}) \leq g(\hat{\boldsymbol{\beta}}^{(k)}|\hat{\boldsymbol{\beta}}^{(k-1)}) \text{ and } g(\hat{\boldsymbol{\beta}}^{(k-1)}|\hat{\boldsymbol{\beta}}^{(k-1)}) = f(\hat{\boldsymbol{\beta}}^{(k-1)}), \tag{2.2}$$

where $\hat{\boldsymbol{\beta}}^{(k)} = \text{argmin}_{\boldsymbol{\beta}}\, g(\boldsymbol{\beta}|\hat{\boldsymbol{\beta}}^{(k-1)})$. It can be checked that $f(\hat{\boldsymbol{\beta}}^{(k)}) \leq f(\hat{\boldsymbol{\beta}}^{(k-1)})$, i.e., the objective function $f(\boldsymbol{\beta})$ decreases at each iteration. Therefore, the minimization of $f(\boldsymbol{\beta})$ decomposes into two parts, constructing a series of local majorization functions $g(\cdot)$, and solving the minimization problem $\hat{\boldsymbol{\beta}}^{(k)} = \text{argmin}_{\boldsymbol{\beta}}\, g(\boldsymbol{\beta}|\hat{\boldsymbol{\beta}}^{(k-1)})$ for each constructed $g(\cdot)$. Note that (2.2) is a

relaxation of the global majorization requirement used in general MM algorithms that require $f(\boldsymbol{\beta}) \leq g(\boldsymbol{\beta}|\hat{\boldsymbol{\beta}}^{(k-1)})$ for all $\boldsymbol{\beta} \in \mathbb{R}^d$. This relaxation was observed by Fan et al. (2018).

Motivated by the local property in (2.2), we now derive an iterative algorithm for solving the series of iteratively reweighted ℓ_1-penalized convex optimization problems in (2.1). To proceed, we consider the special case of weighted ℓ_1-penalized problem

$$\hat{\boldsymbol{\beta}} \in \underset{\boldsymbol{\beta}\in\mathbb{R}^d}{\text{minimize}} \left\{ \frac{1}{n}\sum_{i=1}^{n} L_{\tau,\gamma}(y_i - \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}) + \sum_{j=2}^{d} \lambda_j|\beta_j| \right\}, \tag{2.3}$$

which is indeed a sub-problem of (2.1). Let $\mathscr{R}_n(\boldsymbol{\beta}) = n^{-1}\sum_{i=1}^{n} L_{\tau,\gamma}(y_i - \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta})$ and $\nabla\mathscr{R}_n(\boldsymbol{\beta})$ be its gradient, and let $P(\boldsymbol{\beta}) = \sum_{j=2}^{d}\lambda_j|\beta_j|$. We locally majorize $\mathscr{R}_n(\boldsymbol{\beta})$ at $\hat{\boldsymbol{\beta}}^{(k-1)}$ by constructing an isotropic quadratic function $G_n(\cdot)$ of the form

$$G_n(\boldsymbol{\beta}|\phi_k, \hat{\boldsymbol{\beta}}^{(k-1)}) = \mathscr{R}_n(\hat{\boldsymbol{\beta}}^{(k-1)}) + \langle \nabla\mathscr{R}_n(\hat{\boldsymbol{\beta}}^{(k-1)}), \boldsymbol{\beta} - \hat{\boldsymbol{\beta}}^{(k-1)}\rangle + \frac{\phi_k}{2}||\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}^{(k-1)}||_2^2,$$

where $\phi_k > 0$ is a quadratic parameter to be determined at the k-th iteration. Then define the k-th iterate $\hat{\boldsymbol{\beta}}^{(k)}$ as the solution to

$$\underset{\boldsymbol{\beta}\in\mathbb{R}^d}{\text{minimize}}\ G_n(\boldsymbol{\beta}|\phi_k, \hat{\boldsymbol{\beta}}^{(k-1)}) + P(\boldsymbol{\beta}). \tag{2.4}$$

Clearly we have $G_n(\hat{\boldsymbol{\beta}}^{(k-1)}|\phi_k, \hat{\boldsymbol{\beta}}^{(k-1)}) = \mathscr{R}_n(\hat{\boldsymbol{\beta}}^{(k-1)})$ by construction. To ensure such constructed isotropic quadratic function $G_n(\cdot)$ is indeed a majorization of $\mathscr{R}_n(\cdot)$, we may pick an adaptive and sufficiently large quadratic parameter $\phi_k > 0$ such that $\mathscr{R}_n(\hat{\boldsymbol{\beta}}^{(k)}) \leq G_n(\hat{\boldsymbol{\beta}}^{(k)}|\phi_k, \hat{\boldsymbol{\beta}}^{(k-1)})$. Below we show that such construction also ensures the descent of the

objective function $\mathscr{R}_n(\boldsymbol{\beta}) + P(\boldsymbol{\beta})$ at each iteration.

$$\begin{aligned}\mathscr{R}_n(\hat{\boldsymbol{\beta}}^{(k)}) + P(\hat{\boldsymbol{\beta}}^{(k)}) &\leq G_n(\hat{\boldsymbol{\beta}}^{(k)}|\phi_k, \hat{\boldsymbol{\beta}}^{(k-1)}) + P(\hat{\boldsymbol{\beta}}^{(k)}) \\ &\leq G_n(\hat{\boldsymbol{\beta}}^{(k-1)}|\phi_k, \hat{\boldsymbol{\beta}}^{(k-1)}) + P(\hat{\boldsymbol{\beta}}^{(k-1)}) \\ &= \mathscr{R}_n(\hat{\boldsymbol{\beta}}^{(k-1)}) + P(\hat{\boldsymbol{\beta}}^{(k-1)}),\end{aligned}$$

where the second inequality is due to the fact that $\hat{\boldsymbol{\beta}}^{(k)}$ is a minimizer of (2.4). In practice, we choose ϕ_k by starting from a small value $\phi_0 = 0.01$ and successively inflate it by a factor $\Gamma = 1.25$ until the local majorization requirement $\mathscr{R}_n(\hat{\boldsymbol{\beta}}^{(k)}) \leq G_n(\hat{\boldsymbol{\beta}}^{(k)}|\phi_k, \hat{\boldsymbol{\beta}}^{(k-1)})$ is met at each iteration of the LAMM algorithm.

The main crux of our approach is the isotropic form of $G_n(\boldsymbol{\beta}|\phi_k, \hat{\boldsymbol{\beta}}^{(k-1)})$. As a function of $\boldsymbol{\beta}$, $G_n(\cdot)$ permits a simple analytic solution solution $\hat{\boldsymbol{\beta}}^{(k)}$ for weighted Lasso penalty $P(\boldsymbol{\beta}) = \sum_{j=2}^{d} \lambda_j|\beta_j|$. It is easy to check that $\hat{\boldsymbol{\beta}}^{(k)}$ takes a simple explicit form

$$\begin{cases} \hat{\beta}_1^{(k)} & = \hat{\beta}_1^{(k-1)} - \phi_k^{-1}\nabla_{\beta_1}\mathscr{R}_n(\hat{\boldsymbol{\beta}}^{(k-1)}), \\ \hat{\beta}_j^{(k)} & = S\big(\hat{\beta}_j^{(k-1)} - \phi_k^{-1}\nabla_{\beta_j}\mathscr{R}(\hat{\boldsymbol{\beta}}^{(k-1)}), \phi_k^{-1}\lambda_j\big) \ \text{ for } j \geq 2, \end{cases}$$

where $S(a,b) = \text{sign}(a)\cdot(|a|-b)_+$ denotes the shrinkage operator, $\text{sign}(\cdot)$ is the sign function and $(c)_+ = \max(c,0)$. As a result, the minimization problem at each iteration of LAMM algorithm has an explicit formula, thus $\hat{\boldsymbol{\beta}}^{(k)}$ can be updated efficiently by vector-matrix multiplications. We summarize the whole procedure in the following Algorithm 3 and Algorithm 4.

2.5 Statistical Analysis

In this section, we analyze the sequence of estimators $\{\hat{\boldsymbol{\beta}}^{(t)}\}_{t=1}^{T}$ obtained in (2.1) under the high-dimensional regime in which $d > n$. Throughout the theoretical analysis, we assume that the regression parameter $\boldsymbol{\beta}^* \in \mathbb{R}^d$ in model $y_i = \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}^*(\tau) + \varepsilon_i(\tau)$ is exactly sparse, i.e., $\boldsymbol{\beta}^*$ has s non-zero coordinates. Let $\mathscr{S} = \{1 \leq j \leq d : \beta_j^* \neq 0\}$ be the active set of $\boldsymbol{\beta}^*$ with

Algorithm 3. Local Adaptive Majorize-minimization (LAMM) Algorithm for Solving (2.3) with `retire` loss.

Input: regularization parameters λ_j, expectile level τ, Huber loss tuning parameter γ, inflation factor $\Gamma = 1.25$ and convergence criterion ε.

Initialization: $\hat{\boldsymbol{\beta}}^{(0)} = 0$, $\phi_0 = 0.01$.

Iterate: the following until the stopping criterion $\|\hat{\boldsymbol{\beta}}^{(k)} - \hat{\boldsymbol{\beta}}^{(k-1)}\|_2 \leq \varepsilon$ is met, where $\hat{\boldsymbol{\beta}}^{(k)}$ is the value of $\boldsymbol{\beta}$ obtained at the k-th iteration.

1. Set $\phi_k \leftarrow \max(\phi_0, \phi_{k-1}/\Gamma)$.
2. **repeat**
3. $\hat{\beta}_1^{(k)} \leftarrow \hat{\beta}_1^{(k-1)} - \phi_k^{-1}\nabla_{\beta_1}\mathscr{R}_n(\hat{\boldsymbol{\beta}}^{(k-1)})$, $\hat{\beta}_j^{(k)} \leftarrow S(\hat{\beta}_j^{(k-1)} - \phi_k^{-1}\nabla_{\beta_j}\mathscr{R}_n(\hat{\boldsymbol{\beta}}^{(k-1)}), \phi_k^{-1}\lambda_j)$ for $j \geq 2$.
4. **if** $\mathscr{R}_n(\hat{\boldsymbol{\beta}}^{(k)}) > G_n(\hat{\boldsymbol{\beta}}^{(k)}|\phi_k, \hat{\boldsymbol{\beta}}^{(k-1)})$, set $\phi_k \leftarrow \Gamma\phi_k$.
5. **until** $\mathscr{R}_n(\hat{\boldsymbol{\beta}}^{(k)}) \leq G_n(\hat{\boldsymbol{\beta}}^{(k)}|\phi_k, \hat{\boldsymbol{\beta}}^{(k-1)})$.

Output: the final iterate $\hat{\boldsymbol{\beta}}^{(k)}$.

Algorithm 4. Local Adaptive Majorize-minimization (LAMM) Algorithm for Solving (2.1) with `retire` loss.

Input: regularization parameters λ, penalty function p_λ, expectile level τ, Huber loss tuning parameter γ, inflation factor $\Gamma = 1.25$ and convergence criterion ε.

Initialization: $\hat{\boldsymbol{\beta}}^{(0)} = 0$, $\phi_0 = 0.01$.

Iterate: the following until the stopping criterion $\|\hat{\boldsymbol{\beta}}^{(k)} - \hat{\boldsymbol{\beta}}^{(k-1)}\|_2 \leq \varepsilon$ is met, where $\hat{\boldsymbol{\beta}}^{(k)}$ is the value of $\boldsymbol{\beta}$ obtained at the k-th iteration.

1. Set $\lambda_1 \leftarrow 0$, $\lambda_j \leftarrow p'_\lambda(|\hat{\beta}_j^{(k-1)}|)$ for $j \geq 2$.
2. **apply** Algorithm 3 with parameters $(\lambda_j, \gamma, \Gamma, \varepsilon)$.
3. **update** $\hat{\boldsymbol{\beta}}^{(k)}$.

Output: the final iterate $\hat{\boldsymbol{\beta}}^{(k)}$.

cardinality $|\mathscr{S}| = s$. Recall that $\underline{\tau} = \min(\tau, 1-\tau)$, $\kappa_1 = \min_{|u|\le 1} \ell''(u)$ and $A_1 > 0$ is a constant that satisfies $\mathbb{E}(\boldsymbol{u}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1/2}\boldsymbol{x})^4 \le A_1^4\|\boldsymbol{u}\|_2^4$ for all $\boldsymbol{u} \in \mathbb{R}^d$, where $\boldsymbol{x}$ satisfies Condition 2. Similar to the low-dimensional setting, the key to our high-dimensional analysis is an event $\mathscr{E}_{\mathrm{rsc}}$ that characterizes the local restricted strong convexity property of the empirical loss function $\mathscr{R}_n(\cdot)$ over the intersection of an ℓ_1-cone and a local ℓ_2-ball centered at $\boldsymbol{\beta}^*$ (Loh and Wainwright, 2015). Lemma 2.5.1 below shows that the event $\mathscr{E}_{\mathrm{rsc}}$ occurs with high probability for suitably chosen parameters.

Definition 2.5.1. *Given radii parameters $r, L > 0$ and a curvature parameter $\kappa > 0$, define the event*

$$\mathscr{E}_{\mathrm{rsc}}(r, L, \kappa) = \left\{ \inf_{\boldsymbol{\beta} \in \boldsymbol{\beta}^* + \mathbb{B}(r) \cap \mathbb{C}(L)} \frac{\langle \nabla \mathscr{R}_n(\boldsymbol{\beta}) - \nabla \mathscr{R}_n(\boldsymbol{\beta}^*), \boldsymbol{\beta} - \boldsymbol{\beta}^* \rangle}{\|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_2^2} \ge \kappa \right\},$$

where $\mathbb{B}(r) = \{\boldsymbol{\delta} \in \mathbb{R}^d : \|\boldsymbol{\delta}\|_2 \le r\}$ is an ℓ_2-ball with radius r, and $\mathbb{C}(L) = \{\boldsymbol{\delta} : \|\boldsymbol{\delta}\|_1 \le L\|\boldsymbol{\delta}\|_2\}$ is an ℓ_1-cone.

Lemma 2.5.1. *Let the radii parameters (r, L) and the robustification parameter γ satisfy*

$$\gamma \ge 4\sqrt{2}\lambda_u \max\{\sigma_\varepsilon, 2A_1^2 r\} \qquad \text{and} \qquad n \gtrsim (\sigma_{\boldsymbol{x}} \nu_0 \gamma / r)^2 (L^2 \log d + t).$$

Then, under Conditions 1, 2, and 3, event $\mathscr{E}_{\mathrm{rsc}}(r, L, \kappa)$ with $\kappa = \kappa_1 \underline{\tau}/2$ occurs with probability at least $1 - e^{-t}$.

Under the local restricted strong convexity, in Theorem 2.5.1, we provide an upper bound on the estimation error of $\hat{\boldsymbol{\beta}}^{(1)}$, i.e., the ℓ_1-penalized `retire` estimator.

Theorem 2.5.1. *Assume Conditions 1, 2, and 3 hold. Then, the ℓ_1-penalized `retire` estimator $\hat{\boldsymbol{\beta}}^{(1)}$ with $\gamma = \sigma_\varepsilon \sqrt{n/(\log d + t)}$ and $\lambda \asymp \sqrt{(\log d + t)/n}$ satisfies the bounds*

$$\|\hat{\boldsymbol{\beta}}^{(1)} - \boldsymbol{\beta}^*\|_2 \le 3(\kappa_1 \underline{\tau})^{-1} s^{1/2} \lambda \qquad \text{and} \qquad \|\hat{\boldsymbol{\beta}}^{(1)} - \boldsymbol{\beta}^*\|_1 \le 12(\kappa_1 \underline{\tau})^{-1} s \lambda,$$

with probability as least $1-3e^{-t}$.

Theorem 2.5.1 shows that with an appropriate choice of the tuning parameters γ and λ, the ℓ_1-penalized robust expectile regression satisfies exponential deviation bounds with near-optimal convergence rate as if sub-Gaussian random noise were assumed (Gu and Zou, 2016).

Remark 2.5.1. *Condition 3 can be further relaxed to accommodate heavy-tailed random error with finite* $(1+\phi)$ *moment with* $0<\phi<1$. *Specifically, it can be shown that under the* ℓ_2 *norm, the estimation error of the* ℓ_1*-penalized Huber regression estimator takes the form* $s^{1/2}\{\log(d)/n\}^{\min\{\phi/(1+\phi),1/2\}}$ *(Sun, Zhou and Fan, 2020, Tan, Sun and Witten, 2022). Similar results can be obtained for the proposed* ℓ_1*-penalized* `retire` *estimator and we leave it for future work.*

Remark 2.5.2. *Throughout this section, we assume that the underlying regression parameter* $\boldsymbol{\beta}^*\in\mathbb{R}^d$ *is exactly sparse. In this case, iteratively reweighted* ℓ_1*-penalization helps reduce the estimation bias from* ℓ_1*-penalization as signal strengthens. For weakly sparse vectors* $\boldsymbol{\beta}^*$ *satisfying* $\sum_{j=1}^d|\beta_j^*|^q\le R_q$ *for some* $0<q\le 1$ *and* $R_q>0$, *Fan, Li and Wang (2017) showed that the convergence rate (under* ℓ_2*-norm) of the* ℓ_1*-penalized adaptive Huber estimator with a suitably chosen robustification parameter is of order* $\mathscr{O}(\sigma\sqrt{R_q}\{\log(d)/n\}^{1/2-q/4})$. *Using the same argument, the results in Theorem 2.5.1 can be directly extended to the weakly sparse case where* $\boldsymbol{\beta}^*$ *belongs to an* L_q*-ball for some* $0<q\le 1$. *For recovering weakly sparse signals, folded-concave penalization no longer improves upon* ℓ_1*-penalization, and therefore we will not provide details on such an extension.*

Next, we establish the statistical properties for the entire sequence of estimators $\{\hat{\boldsymbol{\beta}}^{(1)}, \hat{\boldsymbol{\beta}}^{(2)},\ldots,\hat{\boldsymbol{\beta}}^{(T)}\}$ obtained from solving the convex optimization problem (2.1) iteratively. Let $\|\boldsymbol{\beta}^*_{\mathscr{S}}\|_{\min}=\min_{j\in\mathscr{S}}|\beta_j^*|$ be the smallest (in absolute value) non-zero regression coefficient. Under a beta-min condition, we show that the estimation error of $\hat{\boldsymbol{\beta}}^{(1)}$ stated in Theorem 2.5.1 can be refined. More specifically, given the previous iterate $\hat{\boldsymbol{\beta}}^{(T-1)}$, the estimation error of the subsequent estimator, $\hat{\boldsymbol{\beta}}^{(T)}$, can be improved by a δ-fraction for some constant $\delta\in(0,1)$.

Theorem 2.5.2. *Let $p_0(\cdot)$ be a penalty function satisfying Condition 6. Under Conditions 1, 2 and 3, assume there exist some constants $a_1 > a_0 > 0$ such that*

$$a_0 > \sqrt{5}/(\kappa_1 \underline{\tau}), \quad p_0'(a_0) > 0, \quad p_0'(a_1) = 0.$$

Assume further the minimum signal strength condition $\|\boldsymbol{\beta}^*_{\mathscr{S}}\|_{\min} \geq (a_0 + a_1)\lambda$ and the sample size requirement $n \gtrsim s\log d + t$. Picking $\gamma \asymp \sigma_\varepsilon \sqrt{n/(s + \log d + t)}$ and $\lambda \asymp \sigma_\varepsilon \sqrt{(\log d + t)/n}$, we have

$$\|\hat{\boldsymbol{\beta}}^{(T)} - \boldsymbol{\beta}^*\|_2 \lesssim \delta^{T-1}\sigma_\varepsilon \sqrt{\frac{s\,(\log d + t)}{n}} + \frac{\sigma_\varepsilon}{1-\delta}\sqrt{\frac{s + \log d + t}{n}},$$

with probability at least $1 - 4e^{-t}$. Furthermore, setting $T \gtrsim \frac{\log\{\log(d)+t\}}{\log(1/\delta)}$, we have

$$\|\hat{\boldsymbol{\beta}}^{(T)} - \boldsymbol{\beta}^*\|_2 \lesssim \sigma_\varepsilon \sqrt{\frac{s + \log d + t}{n}} \tag{2.5}$$

$$\text{and} \quad \|\hat{\boldsymbol{\beta}}^{(T)} - \boldsymbol{\beta}^*\|_1 \lesssim \sigma_\varepsilon s^{1/2} \sqrt{\frac{s + \log d + t}{n}} \tag{2.6}$$

with probability at least $1 - 4e^{-t}$, where $\delta = \sqrt{5}/(a_0 \kappa_1 \underline{\tau}) < 1$.

Theorem 2.5.2 shows that under the beta-min condition $\|\boldsymbol{\beta}^*_{\mathscr{S}}\|_{\min} \gtrsim \sqrt{\log(d)/n}$, the iteratively reweighted ℓ_1-penalized `retire` estimator $\hat{\boldsymbol{\beta}}^{(T)}$ with $T \asymp \log\{\log(d)\}$ achieves the near-oracle convergence rate, i.e., the convergence rate of the oracle estimator that has access to the true support of $\boldsymbol{\beta}^*$. This is also known as the weak oracle property. Picking $t = \log d$, we see that iteratively reweighted ℓ_1-penalization refines the statistical rate from $\sqrt{s\log(d)/n}$ for $\hat{\boldsymbol{\beta}}^{(1)}$ to $\sqrt{(s + \log d)/n}$ for $\hat{\boldsymbol{\beta}}^{(T)}$.

Remark 2.5.3. *Theorem 2.5.2 reveals the so-called weak oracle property in the sense that the regularized estimator $\hat{\boldsymbol{\beta}}^{(T)}$ enjoys the same convergence rate as the oracle estimator defined by regressing only on the significant predictors. To obtain such a result, the required minimum signal strength $\|\boldsymbol{\beta}^*_{\mathscr{S}}\|_{\min} \gtrsim \sqrt{\log(d)/n}$ is almost necessary and sufficient. To see this, consider*

the linear model $y_i = \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}^* + \varepsilon_i$ *with* $\varepsilon_i \sim N(0,\sigma^2)$ *independent of* $\boldsymbol{x}_i$*, and define the parameter space* $\Omega_{s,a} = \{\boldsymbol{\beta} \in \mathbb{R}^d : \|\boldsymbol{\beta}\|_0 \le s, \min_{j:\beta_j\neq 0}|\beta_j| \ge a\}$ *for* $a > 0$*. Under the assumption that the design matrix* $\mathbb{X} = (\boldsymbol{x}_1,\ldots,\boldsymbol{x}_n)^{\mathrm{T}} \in \mathbb{R}^{n\times d}$ *satisfies a restricted isometry property and has normalized columns, Ndaoud (2019) derived the following sharp lower bounds for the minimax risk* $\psi(s,a) := \inf_{\hat{\boldsymbol{\beta}}}\sup_{\boldsymbol{\beta}^*\in\Omega_{s,a}}\mathbb{E}\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2^2$*: for any* $\varepsilon \in (0,1)$*,*

$$\psi(s,a) \ge \{1+o(1)\}\frac{2\sigma^2 s\log(ed/s)}{n} \ \text{ for any } a \le (1-\varepsilon)\sigma\sqrt{\frac{2\log(ed/s)}{n}}$$
$$\text{and } \ \psi(s,a) \ge \{1+o(1)\}\frac{\sigma^2 s}{n} \ \text{ for any } a \ge (1+\varepsilon)\sigma\sqrt{\frac{2\log(ed/s)}{n}},$$

where the limit corresponds to $s/d \to 0$ *and* $s\log(ed/s)/n \to 0$*.*

The minimax rate $2\sigma^2 s\log(ed/s)/n$ *is attainable by both Lasso and Slope (Bellec, Lecué and Tsybakov, 2018), while the oracle rate* $\sigma^2 s/n$ *can only be achieved when the magnitude of the minimum signal is of order* $\sigma\sqrt{\log(d/s)/n}$*. The beta-min condition imposed in Theorem 2.5.2 is thus (nearly) necessary and sufficient, and is the weakest possible within constant factors.*

Under a stronger beta-min condition $\|\boldsymbol{\beta}^*_{\mathscr{S}}\|_{\min} \gtrsim \sqrt{s\log(d)/n}$*, Gu and Zou (2016) showed that with high probability, the IRW-*ℓ_1 *expectile regression estimator (initialized by zero) coincides with the oracle estimator after three iterations. This is known as the strong oracle property. Based on the more refined analysis by Pan, Sun and Zhou (2021), we conjecture that the IRW-*ℓ_1 `retire` *estimator* $\hat{\boldsymbol{\beta}}^{(T)}$ *with* $T \asymp \log(s\vee\log d)$ *achieves the strong oracle property provided* $\|\boldsymbol{\beta}^*_{\mathscr{S}}\|_{\min} \gtrsim \sqrt{\log(d)/n}$ *without the* $\sqrt{s}$*-factor.*

2.6 Numerical Experiments

In this section, we access the performance of the proposed penalized `retire` estimator via extensive numerical studies. For all of the numerical studies, we generate the covariates $\boldsymbol{x}_i$ from a multivariate normal distribution $N(\mathbf{0}, \boldsymbol{\Sigma} = (\sigma_{jk})_{1\le j,k\le d})$ with $\sigma_{jk} = 0.5^{|j-k|}$. We then generate the response variable y_i from one of the following three models:

1. Homoscedastic model:

$$y_i = \boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{\beta}^* + \varepsilon_i, \tag{2.7}$$

2. Quantile heteroscedastic model:

$$y_i = \boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{\beta}^* + (0.5|x_{id}| + 0.5)\{\varepsilon_i - F_{\varepsilon_i}^{-1}(\tau)\}, \tag{2.8}$$

3. Expectile heteroscedastic model:

$$y_i = \boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{\beta}^* + (0.5|x_{id}| + 0.5)\{\varepsilon_i - e_\tau(\varepsilon_i)\}, \tag{2.9}$$

where ε_i is the random noise, $F_{\varepsilon_i}^{-1}(\cdot)$ denotes the inverse cumulative distribution function of ε_i, and $e_\tau(\varepsilon_i)$ denotes the inverse of the expectile function of ε_i. Note that under Gaussian and t-distributed noises, the two models (2.9) and (2.8) are the same for $\tau = 0.5$. We set the regression coefficient vector $\boldsymbol{\beta}^* = (\beta_1^*, \beta_2^*, \ldots, \beta_d^*)^{\mathrm{T}}$ as $\beta_1^* = 2$ (intercept), $\beta_j^* = \{1.8, 1.6, 1.4, 1.2, 1, -1, -1.2, -1.4, -1.6, -1.8\}$ for $j = 2, 4, \ldots, 20$, and 0 otherwise. The random noise is generated from either a Gaussian distribution, $N(0,2)$, or a t distribution with 2.1 degrees of freedom. For the heteroscedastic models, we consider two quantile/expectile levels $\tau = \{0.5, 0.8\}$.

Similar to Section 1.7, we update the robustification parameter γ using a heuristic tuning method

$$\gamma^k = \mathrm{mad}(\tilde{\mathbf{r}}^k) \cdot \sqrt{\frac{n}{\log(nd)}}. \tag{2.10}$$

Throughout our numerical studies, we have found that γ chosen using the above heuristic approach works well across different scenarios. Our computational results are reproducible using codes available from https://github.com/ZianWang0128/Retire.

2.6.1 Estimation

In this subsection, we implement the ℓ_1-penalized `retire` and the IRW-ℓ_1-penalized `retire` using SCAD-based weights with $T = 3$, which we compare to three other competitive methods: (i) ℓ_1-penalized Huber regression (`huber`); (ii) ℓ_1-penalized asymmetric least squares regression (`sales`) proposed by Gu and Zou (2016), and (iii) ℓ_1-penalized quantile regression (`qr`) implemented via the `R` package `rqPen` (Sherwood and Maidman, 2020). To assess the performance across different methods, we report the estimation error under the ℓ_2-norm, i.e., $\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2$, the true positive rate (TPR), and the false positive rate (FPR). Here, TPR is defined as the proportion of the number of correctly identified non-zeros and the false positive rate is calculated as the proportion of the number of incorrectly identified nonzeros.

Note that `huber` and `sales` are special cases of `retire` by taking $\tau = 0.5$ and $\gamma \to \infty$, respectively. Thus, both `huber` and `sales` can be implemented via Algorithm 4. For all methods, the sparsity inducing tuning parameter λ is selected via ten-fold cross-validation. Specifically, for methods ℓ_1-penalized `retire`, `huber`, and `sales`, we select the largest tuning parameter that yields a value of the asymmetric least squares loss that is less than the minimum of the asymmetric least squares loss plus one standard error. For `qr`, we use the default cross validation function in `R` package `rqPen` to select the largest tuning parameter that yields a value of its corresponding loss function that is less than the minimum of the quantile loss plus one standard error. For IRW `retire`, we select the largest tuning parameter that yields a value of the asymmetric least squares loss that is the minimum of the asymmetric least squares. Also note that both `huber` and `retire` require tuning an additional robustness parameter γ. We select γ using the heuristic tuning method (2.10) to update γ at the beginning of each iteration in Algorithm 4.

The results, averaged over 100 repetitions, are reported in Tables 2.1–2.4 for the moderate- ($n = 400,\ d = 200$) and high-dimensional ($n = 400,\ d = 500$) settings.

Table 2.1 contains results ($\tau = 0.5$) under the homoscedastic model (2.7) with normally and t-distributed noise. For Gaussian noise, the four ℓ_1-penalized estimators have similar

Table 2.1. Homoscedastic model (2.7) with Gaussian noise ($\varepsilon \sim N(0,2)$) and $t_{2.1}$ noise ($\varepsilon \sim t_{2.1}$). Estimation error under ℓ_2-norm (and its standard deviation), true positive rate (TPR) and false positive rate (FPR), averaged over 100 repetitions, are reported.

		$n = 400, d = 200$			$n = 400, d = 500$		
Noise	Method	ℓ_2 error	TPR	FPR	ℓ_2 error	TPR	FPR
Gaussian	ℓ_1 retire	0.569 (0.076)	1.000 (0.000)	0.026 (0.016)	0.604 (0.094)	1.000 (0.000)	0.012 (0.009)
	IRW retire (SCAD)	**0.256** (0.059)	1.000 (0.000)	0.014 (0.027)	**0.261** (0.068)	1.000 (0.000)	0.008 (0.015)
	ℓ_1 huber	0.569 (0.076)	1.000 (0.000)	0.026 (0.016)	0.604 (0.094)	1.000 (0.000)	0.012 (0.009)
	ℓ_1 sales	0.569 (0.076)	1.000 (0.000)	0.026 (0.016)	0.605 (0.094)	1.000 (0.000)	0.012 (0.009)
	ℓ_1 qr	0.566 (0.084)	1.000 (0.000)	0.153 (0.036)	0.661 (0.073)	1.000 (0.000)	0.154 (0.030)
$t_{2.1}$	ℓ_1 retire	1.206 (0.349)	0.996 (0.020)	0.006 (0.005)	1.266 (0.398)	0.992 (0.031)	0.003 (0.003)
	IRW retire (SCAD)	**0.299** (0.083)	1.000 (0.000)	0.017 (0.035)	**0.299** (0.079)	1.000 (0.000)	0.011 (0.020)
	ℓ_1 huber	1.206 (0.349)	0.996 (0.020)	0.006 (0.005)	1.266 (0.398)	0.992 (0.031)	0.003 (0.003)
	ℓ_1 sales	1.317 (0.384)	0.994 (0.024)	0.010 (0.007)	1.380 (0.425)	0.985 (0.041)	0.004 (0.004)
	ℓ_1 qr	0.497 (0.090)	1.000 (0.000)	0.126 (0.036)	0.560 (0.074)	1.000 (0.000)	0.141 (0.030)

Table 2.2. Heteroscedastic model (2.8) with Gaussian noise ($\varepsilon \sim N(0,2)$) and quantile levels $\tau = \{0.5, 0.8\}$. Estimation error under ℓ_2-norm (and its standard deviation), true positive rate (TPR) and false positive rate (FPR), averaged over 100 repetitions, are reported.

		$n = 400, d = 200$			$n = 400, d = 500$		
τ	Method	ℓ_2 error	TPR	FPR	ℓ_2 error	TPR	FPR
0.5	ℓ_1 retire	0.555 (0.077)	1.000 (0.000)	0.022 (0.015)	0.584 (0.094)	1.000 (0.000)	0.011 (0.011)
	IRW retire (SCAD)	**0.237** (0.052)	1.000 (0.000)	0.012 (0.024)	**0.243** (0.064)	1.000 (0.000)	0.008 (0.016)
	ℓ_1 huber	0.555 (0.077)	1.000 (0.000)	0.022 (0.015)	0.584 (0.094)	1.000 (0.000)	0.011 (0.011)
	ℓ_1 sales	0.560 (0.079)	1.000 (0.000)	0.023 (0.015)	0.594 (0.096)	1.000 (0.000)	0.010 (0.010)
	ℓ_1 qr	0.478 (0.078)	1.000 (0.000)	0.150 (0.037)	0.551 (0.066)	1.000 (0.000)	0.151 (0.030)
0.8	ℓ_1 retire	0.628 (0.083)	1.000 (0.000)	0.030 (0.033)	0.651 (0.097)	1.000 (0.000)	0.013 (0.010)
	IRW retire (SCAD)	**0.385** (0.076)	1.000 (0.000)	0.009 (0.019)	**0.383** (0.066)	1.000 (0.000)	0.005 (0.011)
	ℓ_1 huber	1.200 (0.080)	1.000 (0.000)	0.023 (0.017)	1.212 (0.086)	1.000 (0.000)	0.011 (0.011)
	ℓ_1 sales	0.669 (0.083)	1.000 (0.000)	0.026 (0.024)	0.696 (0.093)	1.000 (0.000)	0.010 (0.008)
	ℓ_1 qr	0.563 (0.093)	1.000 (0.000)	0.176 (0.038)	0.661 (0.089)	1.000 (0.000)	0.176 (0.025)

Table 2.3. Heteroscedastic model (2.8) with $t_{2.1}$ noise ($\varepsilon \sim t_{2.1}$) and quantile levels $\tau = \{0.5, 0.8\}$. Estimation error under ℓ_2-norm (and its standard deviation), true positive rate (TPR) and false positive rate (FPR), averaged over 100 repetitions, are reported.

		$n = 400, d = 200$			$n = 400, d = 500$		
τ	Method	ℓ_2 error	TPR	FPR	ℓ_2 error	TPR	FPR
0.5	ℓ_1 retire	1.149 (0.364)	0.998 (0.014)	0.006 (0.005)	1.187 (0.396)	0.994 (0.028)	0.002 (0.003)
	IRW retire (SCAD)	**0.264** (0.079)	1.000 (0.000)	0.014 (0.030)	**0.269** (0.077)	1.000 (0.000)	0.012 (0.022)
	ℓ_1 huber	1.149 (0.364)	0.998 (0.014)	0.006 (0.005)	1.187 (0.396)	0.994 (0.028)	0.002 (0.003)
	ℓ_1 sales	1.263 (0.397)	0.995 (0.022)	0.009 (0.007)	1.297 (0.421)	0.992 (0.031)	0.004 (0.003)
	ℓ_1 qr	0.414 (0.079)	1.000 (0.000)	0.120 (0.034)	0.466 (0.068)	1.000 (0.000)	0.136 (0.032)
0.8	ℓ_1 retire	1.479 (0.710)	0.972 (0.064)	0.007 (0.006)	1.455 (0.640)	0.980 (0.051)	0.003 (0.003)
	IRW retire (SCAD)	**0.378** (0.168)	1.000 (0.000)	0.015 (0.033)	**0.362** (0.114)	1.000 (0.000)	0.008 (0.017)
	ℓ_1 huber	1.503 (0.286)	0.998 (0.014)	0.006 (0.005)	1.534 (0.314)	0.995 (0.022)	0.003 (0.003)
	ℓ_1 sales	1.619 (0.684)	0.967 (0.064)	0.013 (0.013)	1.592 (0.616)	0.972 (0.062)	0.005 (0.005)
	ℓ_1 qr	0.621 (0.133)	1.000 (0.000)	0.155 (0.045)	0.704 (0.104)	1.000 (0.000)	0.167 (0.039)

Table 2.4. Heteroscedastic model (2.9) with Gaussian noise ($\varepsilon \sim N(0,2)$) and $t_{2.1}$ noise ($\varepsilon \sim t_{2.1}$), under the τ-expectile $= 0.8$. Estimation error under ℓ_2-norm (and its standard deviation), true positive rate (TPR) and false positive rate (FPR), averaged over 100 repetitions, are reported.

		$n = 400, d = 200$			$n = 400, d = 500$		
Noise	Method	ℓ_2 error	TPR	FPR	ℓ_2 error	TPR	FPR
Gaussian	ℓ_1 retire	0.607 (0.091)	1.000 (0.000)	0.030 (0.033)	0.645 (0.114)	1.000 (0.000)	0.013 (0.010)
	IRW retire (SCAD)	**0.274** (0.102)	1.000 (0.000)	0.011 (0.029)	**0.271** (0.068)	1.000 (0.000)	0.006 (0.011)
	ℓ_1 huber	0.887 (0.080)	1.000 (0.000)	0.023 (0.017)	0.904 (0.088)	1.000 (0.000)	0.011 (0.012)
	ℓ_1 sales	0.606 (0.093)	1.000 (0.000)	0.026 (0.028)	0.639 (0.113)	1.000 (0.000)	0.011 (0.009)
	ℓ_1 qr	0.653 (0.101)	1.000 (0.000)	0.178 (0.036)	0.701 (0.087)	1.000 (0.000)	0.177 (0.026)
$t_{2.1}$	ℓ_1 retire	1.486 (0.711)	0.973 (0.063)	0.007 (0.007)	1.462 (0.645)	0.980 (0.051)	0.003 (0.003)
	IRW retire (SCAD)	**0.363** (0.170)	1.000 (0.000)	0.014 (0.032)	**0.349** (0.118)	1.000 (0.000)	0.008 (0.017)
	ℓ_1 huber	1.481 (0.290)	0.998 (0.014)	0.006 (0.005)	1.511 (0.318)	0.995 (0.022)	0.003 (0.003)
	ℓ_1 sales	1.625 (0.688)	0.967 (0.064)	0.013 (0.013)	1.598 (0.621)	0.972 (0.062)	0.005 (0.005)
	ℓ_1 qr	0.624 (0.136)	1.000 (0.000)	0.154 (0.045)	0.706 (0.105)	1.000 (0.000)	0.168 (0.039)

performance, except that ℓ_1 `qr` has a much worse FPR. IRW `retire` (with SCAD) significantly reduces estimation error while remaining comparable FPR. For the $t_{2.1}$ noise, we see that IRW `retire` gains considerable advantage over all other methods in estimation error while maintaining model selection accuracy, suggesting that the proposed estimator gains robustness without compromising statistical accuracy. $t_{2.1}$ noise is fairly heavy-tailed, thus large biases are introduced due to the ℓ_1 penalty. IRW `retire` reduces such bias by iteratively solving a series of optimization problems with smaller λ levels, at the cost of more computation involved.

Tables 2.2 and 2.3 show results under the quantile heteroscedastic model (2.8) with the Gaussian and $t_{2.1}$ noise, respectively. Two quantile levels $\tau = \{0.5, 0.8\}$ are considered. We see that `huber` and ℓ_1-penalized `retire` have the same performance when $\tau = 0.5$ since they are indeed equivalent for the case when $\tau = 0.5$. When $\tau = 0.8$, the performance of `huber` deteriorates since `huber` implicitly assumes $\tau = 0.5$ and there is a non-negligible bias when $\tau = 0.8$. Moreover, IRW `retire` has the lowest estimation error among all methods. Similar results can also be found in Table 2.4 for the expectile heteroscedastic model (2.9).

We want to point out that in general the quantile regression method `qr` is quite 'stable'. It always produces acceptable estimation error (especially under the $t_{2.1}$ noise) and much higher FPR than other methods. In fact, the quantile loss is more robust to outliers than the asymmetric square losses, which contributes to low estimation errors under the $t_{2.1}$ noise. As a side effect of being robust for the quantile loss, it hardly shows curvature for the cross-validation process

when tuning the penalty levels λ. Therefore, the performance of `qr` method highly depends on the candidate λ sequence generated by the `rqPen` package, which might cause the consistent high FPR for `qr` than other methods. In summary, the numerical studies confirm IRW `retire` as a robust alternative with better statistical accuracy to its least squares counterpart `sales`.

2.6.2 Inference for Confidence Intervals

In this subsection, we apply both Multiple Bootstrap (MB) technique and Post Selection Inference (PSI) to obtain confidence intervals for signals β_j^*. Details of Multiple Bootstrap technique and Post Selection Inference will be found in the following subsections.

We use the following 95% confidence interval derived from asymptotic normality as a benchmark. It is a simple consequence of Theorem 1.6.3 and its derivation can be found in (1.7).

$$\left[\hat{\beta}_j - 1.96\frac{\hat{\sigma}(\mathbf{e}_j)}{\sqrt{n}},\ \hat{\beta}_j - 1.96\frac{\hat{\sigma}(\mathbf{e}_j)}{\sqrt{n}}\right],$$

where

$$\hat{\sigma}^2(\mathbf{e}_j) := \mathbf{e}_j^{\mathrm{T}}\hat{\mathbf{J}}^{-1}\left[\frac{1}{n}\sum_{i=1}^{n}\zeta^2(\hat{\varepsilon}_i)\boldsymbol{x}_i\boldsymbol{x}_i^{\mathrm{T}}\right]\hat{\mathbf{J}}^{-1}\mathbf{e}_j,$$

and $\zeta(u) = L'_{\tau,\gamma}(u) = |\tau - \mathbb{1}(u<0)|\cdot\ell'_\gamma(u)$ is the first-order derivative of $L_{\tau,\gamma}(\cdot)$ given in (1.5).

Multiple Bootstrap

Here we briefly outline three main steps of Multiple Bootstrap (MB) procedure to obtain confidence intervals for signals β_j^*.

Firstly, given data $\{(y_i,\boldsymbol{x}_i)\}_{i=1}^n$ that follows the data generation process detailed in Section 2.6.1, we obtain bootstrap samples $\{\hat{\boldsymbol{\beta}}_1^{boot},\ldots,\hat{\boldsymbol{\beta}}_B^{boot}\}$ by repeatedly minimizing randomly weighted objective functions $\{n^{-1}\sum_{i=1}^n w_i^b\cdot L_{\tau,\gamma}(y_i-\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta})+\sum_{j=2}^d p'_\lambda(|\hat{\beta}_j^{b,(t-1)}|)|\beta_j|\}$, where $L_{\tau,\gamma}$ is defined in (1.5) that satisfies Condition 1, p'_λ satisfies Condition 6, and $w_1^b,\ldots,w_n^b$ are i.i.d. random bootstrap weights that satisfy $\mathbb{E}(w_i^b)=\mathrm{var}(w_i^b)=1$. For convenience, we focus on the Huber loss for which $\ell(u) = u^2/2\cdot\mathbb{1}(|u|\le 1)+(|u|-1/2)\cdot\mathbb{1}(|u|>1)$. Moreover we choose

SCAD penalty and exponential i.i.d. bootstrap weights, i.e., $w_i^b \sim \exp(1)$.

Secondly, we perform a majority vote to obtain bootstrap estimated active set $\mathscr{S} := \{j = 2,\ldots,d : \hat{p}_j \geq 0.5\}$, where $\hat{p}_j = \frac{1}{B}\sum_{b=1}^{B} \mathbb{1}(|\hat{\beta}_{b,j}^{boot}| > 0)$ is the selection rate for each covariate (intercept excluded). Construct the selected low-dimensional dataset $\{(y_i, \boldsymbol{x}_{i,\{1\}\cup\mathscr{S}})\}_{i=1}^n$, based on which compute a (unregularized) `retire` estimator, denoted as $\hat{\boldsymbol{\beta}}^{ini(low)}$. The benchmark confidence interval (1.7) can be constructed based on $\hat{\boldsymbol{\beta}}^{ini(low)}$.

Lastly, we obtain low-dimensional bootstrap samples $\{\hat{\boldsymbol{\beta}}_1^{boot(low)}, \ldots, \hat{\boldsymbol{\beta}}_B^{boot(low)}\}$ by repeatedly minimizing randomly weighted objective functions $n^{-1}\sum_{i=1}^n w_i^b \cdot L_{\tau,\gamma}(y_i - \boldsymbol{x}_{i,\{1\}\cup\mathscr{S}}^{\mathrm{T}}\boldsymbol{\beta})$, and construct confidence intervals. More specifically, we can construct MB confidence intervals for all the slope coefficients using one of the three classical methods, the percentile method, the pivotal method, and the normal-based method. Let $\alpha \in (0,1)$ be a prespecified confidence level.

1. Efron's percentile method: For each $q \in (0,1)$ and $2 \leq j \leq d$, define the conditional q-quantile of $\hat{\beta}_j^{boot(low)}$ given the observed data as

$$c_j^b(q) = \inf\{t \in \mathbb{R} : \mathbb{P}^*(\hat{\beta}_j^{boot(low)} \leq t) \geq q\}.$$

   Then then Efron's percentile interval for β_j^* takes the form

$$\Big[c_j^b(\alpha/2), c_j^b(1-\alpha/2)\Big]. \tag{2.11}$$

2. Pivotal method: The pivotal interval approximates the conditional distribution of $\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*$ by the bootstrap quantity $\hat{\boldsymbol{\beta}}^{boot} - \hat{\boldsymbol{\beta}}$. More specifically, the pivotal confidence interval for β_j^* takes the form

$$\Big[2\hat{\beta}_j^{ini(low)} - c_j^b(1-\alpha/2), 2\hat{\beta}_j^{ini(low)} - c_j^b(\alpha/2)\Big]. \tag{2.12}$$

   Pivotal confidence intervals are connected to percentile confidence intervals in sense that

the latter are the pivotal confidence intervals reflected about the point $\hat{\beta}_j^{\,boot}$.

3. Normal-based method: Let $\Phi^{-1}(\cdot)$ be the inverse of the cumulative distribution function of a standard normal random variable. Denote $\mathrm{std}(\cdot)$ as the sample standard deviation. Then the normal-based confidence interval for β_j^* takes the form

$$\Big[\hat{\beta}_j^{ini(low)} - \Phi^{-1}(1-\alpha/2)\cdot \mathrm{std}\big\{\hat{\beta}_{\cdot,j}^{boot(low)}\big\}, \hat{\beta}_j^{ini(low)} + \Phi^{-1}(1-\alpha/2)\cdot \mathrm{std}\big\{\hat{\beta}_{\cdot,j}^{boot(low)}\big\}\Big]. \tag{2.13}$$

We summarize the whole process as follow:

Procedure 5. Multiple Bootstrap Inference for β^*.

Input: generated data $\{(y_i, \boldsymbol{x}_i)\}_{i=1}^n$, bootstrap weights w_i, cross-validation selected regularization parameter λ, Huber loss tuning parameter γ, penalty function p_λ, and corresponding solution $\hat{\boldsymbol{\beta}}_{ini}$.
Initialization: $B = 200, T = 3$.

1. For $b = 1, \ldots, B$, obtain bootstrapped regularized estimators by iteratively solving

$$\hat{\boldsymbol{\beta}}^{b,(t)} \in \underset{\boldsymbol{\beta}\in\mathbb{R}^d}{\text{minimize}} \left\{ \frac{1}{n}\sum_{i=1}^n w_i^b \cdot L_{\tau,\gamma}(y_i - \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}) + \sum_{j=2}^d p'_\lambda(|\hat{\beta}_j^{b,(t-1)}|)|\beta_j| \right\},\ t = 1, 2, \ldots, T,$$

   where $\hat{\boldsymbol{\beta}}^{b,(0)} = \hat{\boldsymbol{\beta}}_{ini}$, and denote the final iterate $\hat{\boldsymbol{\beta}}^{b,(T)}$ by $\hat{\boldsymbol{\beta}}_b^{\,boot}$.

2. Perform majority vote to obtain the estimated active set $\mathscr{S} = \{j = 2, \ldots, d : \hat{p}_j \geq 0.5\}$, where $\hat{p}_j = \frac{1}{B}\sum_{b=1}^B \mathbb{1}(|\hat{\beta}_{b,j}^{boot}| > 0)$.

3. Compute an (unregularized) estimator $\hat{\boldsymbol{\beta}}^{ini(low)}$ based on the selected dataset $\{(y_i, \boldsymbol{x}_{i,\{1\}\cup\mathscr{S}})\}_{i=1}^n$.

4. For $b = 1, \ldots, B$, obtain low-dimensional bootstrapped estimators by solving

$$\hat{\boldsymbol{\beta}}_b^{boot(low)} \in \underset{\boldsymbol{\beta}\in\mathbb{R}^{|\mathscr{S}|+1}}{\text{minimize}} \frac{1}{n}\sum_{i=1}^n w_i^b \cdot L_{\tau,\gamma}(y_i - \boldsymbol{x}_{i,\{1\}\cup\mathscr{S}}^{\mathrm{T}}\boldsymbol{\beta}).$$

5. Calculate confidence intervals from $\{\hat{\boldsymbol{\beta}}_1^{boot(low)}, \ldots, \hat{\boldsymbol{\beta}}_B^{boot(low)}\}$ by (2.11), (2.12) and (2.13).

Output: Multiple bootstrap confidence intervals.

Post Selection Inference

Recall that for a linear regression model

$$y_i = \boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{\beta}^* + \varepsilon_i,$$

where $y_i \in \mathbb{R}$ is a response variable and $\boldsymbol{x}_i \in \mathbb{R}^d$ is a d-dimensional covariate, the noise term $\varepsilon_i \in \mathbb{R}$ may be heavy-tailed and asymmetrically distributed. Under moderate to high-dimensional settings where d can be of the same order as n or greater than n, our goal is to construct confidence intervals for the true signals β_j^* based on some sparse estimators/initializers. However, sparse estimators such as the Lasso do not have a tractable limiting distribution, therefore statistical inference with high-dimensional data is challenging.

More specifically, starting with an initial estimator $\hat{\boldsymbol{\beta}}^{ini}$, we aim to debias the estimator for the true signal β_j^* by solving

$$\hat{\beta}_j \in \underset{\beta_j \in \mathbb{R}}{\text{minimize}} \, \frac{1}{n} \sum_{i=1}^{n} L_{\tau,\gamma}(y_i - \boldsymbol{x}_{i,-j}^{\mathrm{T}} \hat{\boldsymbol{\beta}}_{-j}^{ini} - x_{i,j}\beta_j),$$

where $L_{\tau,\gamma}$ is defined in (1.5) that satisfies Condition 1, $\boldsymbol{x}_{i,-j}$ and $\hat{\boldsymbol{\beta}}_{-j}^{ini}$ are, respectively, the sub-vectors of $\boldsymbol{x}_i$ and $\hat{\boldsymbol{\beta}}^{ini}$, deleting the j-th element. However, it is well known that the asymptotic normality of $\hat{\beta}_j$ can not be established if the initial estimator $\hat{\boldsymbol{\beta}}^{ini}$ is not $n^{1/2}$-consistent. Inspired by orthogonalization (Neyman, J. , 1959, Zhang and Zhang, 2014, Belloni, Chernozhukov and Kato, 2015) and decorrelated score (Ning and Liu, 2017), we conjecture the following orthogonal property

$$\frac{\partial}{\partial \eta} \mathbb{E}\Big\{ (-x_{i,j} + \boldsymbol{x}_{i,-j}^{\mathrm{T}} \boldsymbol{v}_j) L'_{\tau,\gamma}(y_i - \boldsymbol{x}_{i,-j}^{\mathrm{T}} \boldsymbol{\beta}_{-j} - x_{i,j}\beta_j^*) \Big\} \Big|_{\eta=\eta^*} = 0, \tag{2.14}$$

where $\boldsymbol{v}_j^* = \mathrm{argmin}_{\boldsymbol{v}_j} \mathbb{E}(x_{i,j} - \boldsymbol{x}_{i,-j}^{\mathrm{T}} \boldsymbol{v}_j)^2$, $\eta = (\boldsymbol{v}_j^{\mathrm{T}}, \boldsymbol{\beta}_{-j}^{\mathrm{T}})^{\mathrm{T}}$ and $\eta^* = (\boldsymbol{v}_j^{*\mathrm{T}}, \boldsymbol{\beta}_{-j}^{*\mathrm{T}})^{\mathrm{T}}$. Corresponding to

orthogonal property (2.14), we consider its empirical version as an estimation equation for β_j^*

$$\frac{1}{n}\sum_{i=1}^{n}(-x_{i,j}+\boldsymbol{x}_{i,-j}^{\mathrm{T}}\hat{\boldsymbol{v}}_j)L'_{\tau,\gamma}(y_i-\boldsymbol{x}_{i,-j}^{\mathrm{T}}\hat{\boldsymbol{\beta}}_{-j}^{ini}-x_{i,j}\beta_j), \tag{2.15}$$

where $\hat{\boldsymbol{v}}_j$ is a consistent estimator of $\boldsymbol{v}_j^*$. The orthogonal property (2.14) ensures the convergence rate of $\hat{\beta}_j$ derived from (2.15) will not be affected by $\hat{\boldsymbol{\beta}}^{ini}$, i.e., $\hat{\boldsymbol{\beta}}^{ini}$ is allowed to have a slower convergence rate than $o(n^{1/2})$. However, it is difficult to solve (2.15) directly due to the existence of indicator functions inside $L'_{\tau,\gamma}$. To proceed, we resort to the idea of one-step estimation (Bickel, 1975). Define $S(\beta_j)=\mathbb{E}\big\{(-x_{i,j}+\boldsymbol{x}_{i,-j}^{\mathrm{T}}\boldsymbol{v}_j^*)L'_{\tau,\gamma}(y_i-\boldsymbol{x}_{i,-j}^{\mathrm{T}}\boldsymbol{\beta}_{-j}^*-x_{i,j}\beta_j)\big\}$, let $S'(\beta_j)$ be its derivative with respect to β_j and $\varepsilon_i^{ini}=y_i-\boldsymbol{x}_i^{\mathrm{T}}\hat{\boldsymbol{\beta}}^{ini}$. Instead of solving (2.15), we consider the one-step estimator $\hat{\beta}_j=\hat{\beta}_j^{ini}+\{S'(\beta_j^*)\}^{-1}n^{-1}\sum_{i=1}^{n}(x_{i,j}-\boldsymbol{x}_{i,-j}^{\mathrm{T}}\hat{\boldsymbol{v}}_j)L'_{\tau,\gamma}(\varepsilon_i^{ini})$, and plug-in an empirical counterpart of the unknown $S'(\beta_j^*)$ to obtain

$$\hat{\beta}_j=\hat{\beta}_j^{ini}+\frac{\sum_{i=1}^{n}(x_{i,j}-\boldsymbol{x}_{i,-j}^{\mathrm{T}}\hat{\boldsymbol{v}}_j)L'_{\tau,\gamma}(\varepsilon_i^{ini})}{\sum_{i=1}^{n}x_{i,j}(x_{i,j}-\boldsymbol{x}_{i,-j}^{\mathrm{T}}\hat{\boldsymbol{v}}_j)\times\frac{1}{n}\sum_{i=1}^{n}L''_{\tau,\gamma}(\varepsilon_i^{ini})}.$$

And the PSI confidence interval for β_j^* has the form

$$\big[\hat{\beta}_j-\Phi^{-1}(1-\alpha/2)n^{-1/2}\hat{\sigma},\hat{\beta}_j+\Phi^{-1}(1-\alpha/2)n^{-1/2}\hat{\sigma}\big],$$

where

$$\hat{\sigma}^2=\frac{\sum_{i=1}^{n}\big\{L'_{\tau,\gamma}(\varepsilon_i^{ini})\big\}^2}{\sum_{i=1}^{n}x_{i,j}(x_{i,j}-\boldsymbol{x}_{i,-j}^{\mathrm{T}}\hat{\boldsymbol{v}}_j)\times\big\{\frac{1}{n}\sum_{i=1}^{n}L''_{\tau,\gamma}(\varepsilon_i^{ini})\big\}^2}.$$

We use IRW `retire` (with SCAD) as the initial estimator $\hat{\boldsymbol{\beta}}^{ini}$, and estimate $\hat{\boldsymbol{v}}_j$ by the Lasso estimator $\hat{\boldsymbol{v}}_j\in\underset{\boldsymbol{v}_j}{\text{minimize}}\ n^{-1}(x_{i,j}-\boldsymbol{x}_{i,-j}^{\mathrm{T}}\boldsymbol{v}_j)^2+\lambda_j||\boldsymbol{v}_j||_1$ with cross-validation selected penalty level λ_j.

For both Multiple Bootstrap and Post Selection Inference, we follow the same data

generation process, model settings, and hyper-parameter selection methods as Section 2.6.1. All inference results, averaged over 100 repetitions, are reported in the following table.

Table 2.5. Inference results for Multiple Bootstrap (MB) and Post Selection Inference (PSI). Coverage rate (and the width of confidence intervals) are reported, averaged over 100 repetitions.

	CI type	Homo-model (2.7)		Quantile hetero-model (2.8)				Expectile hetero-model (2.9)	
		$\tau = 0.5$		$\tau = 0.5$		$\tau = 0.8$		$\tau = 0.8$	
		$\varepsilon \sim N(0,2)$	$\varepsilon \sim t_{2.1}$	$\varepsilon \sim N(0,2)$	$\varepsilon \sim t_{2.1}$	$\varepsilon \sim N(0,2)$	$\varepsilon \sim t_{2.1}$	$\varepsilon \sim N(0,2)$	$\varepsilon \sim t_{2.1}$
$n = 400$ $d = 200$	Percentile	0.912 (0.277)	0.932 (0.358)	0.920 (0.264)	0.927 (0.333)	0.918 (0.271)	0.907 (0.427)	0.914 (0.276)	0.906 (0.428)
	Pivotal	0.921 (0.277)	0.938 (0.358)	0.920 (0.264)	0.944 (0.333)	0.917 (0.271)	0.920 (0.427)	0.908 (0.276)	0.921 (0.428)
	MB-normal	0.923 (0.282)	0.940 (0.365)	0.927 (0.268)	0.940 (0.339)	0.922 (0.276)	0.928 (0.433)	0.920 (0.281)	0.927 (0.434)
	PSI	0.949 (0.333)	0.948 (0.358)	0.951 (0.310)	0.947 (0.342)	0.939 (0.321)	0.931 (0.389)	0.933 (0.327)	0.932 (0.391)
	Normal	0.934 (0.288)	0.929 (0.351)	0.929 (0.272)	0.933 (0.327)	0.932 (0.284)	0.916 (0.420)	0.927 (0.290)	0.917 (0.420)
$n = 400$ $d = 500$	Percentile	0.907 (0.277)	0.914 (0.351)	0.916 (0.263)	0.919 (0.325)	0.898 (0.269)	0.908 (0.407)	0.897 (0.275)	0.911 (0.407)
	Pivotal	0.913 (0.277)	0.926 (0.351)	0.919 (0.263)	0.928 (0.325)	0.898 (0.269)	0.920 (0.407)	0.895 (0.275)	0.919 (0.407)
	MB-normal	0.923 (0.282)	0.925 (0.355)	0.925 (0.267)	0.929 (0.330)	0.912 (0.274)	0.928 (0.412)	0.917 (0.280)	0.928 (0.412)
	PSI	0.949 (0.330)	0.951 (0.373)	0.946 (0.306)	0.949 (0.330)	0.935 (0.317)	0.938 (0.363)	0.936 (0.324)	0.938 (0.364)
	Normal	0.927 (0.287)	0.919 (0.343)	0.931 (0.271)	0.925 (0.320)	0.916 (0.281)	0.921 (0.402)	0.915 (0.287)	0.922 (0.403)

From Table 2.5 we see that all methods except PSI have coverage rates slightly lower than 95%. It is not surprising since those methods require a majority vote process to select estimated active sets from the whole dataset, and such selection may miss some of the true signals β_j^*, resulting in lower coverage rates for confidence intervals. Other than this, all types of confidence intervals perform similarly.

2.6.3 Data Application: NCI-60 Cancer Cell Lines Data

In this subsection, we apply the proposed method to the NCI-60 dataset, a panel of 60 diverse human cancer cell lines. We use two NCI-60 transcript profile datasets, the gene expression dataset and the protein profile dataset. Both datasets can be obtained via the CellMiner database and query tool (Reinhold et al., 2012, Shankavaram et al., 2009). The gene expression data are obtained on Affymetrix HG-U133A/B chips, $\log_2$-transformed, and normalized using the guanine cytosine robust multi-array analysis as in Hansen et al. (2012). It measures 17992 gene expression levels for 60 human cancer cell lines. The protein profile data of 162 antibody (protein) expression levels are obtained on reverse-phase protein lysate arrays for a total of 60 human cancer cell lines. We remove one observation since all values are missing, reducing the number of observations to $n = 59$.

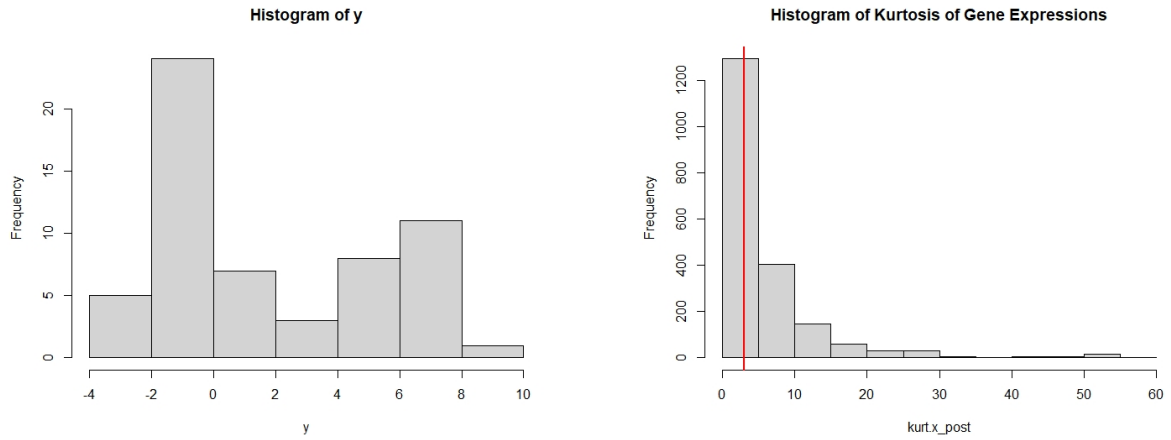

Figure 2.1. Histograms of the KRT19 antibody expression levels and the kurtosis of gene expression levels. The red line at 3 is the kurtosis of a standard normal distribution.

We center the gene expression for each gene to have mean zero, and select the first 2000 genes with largest standard deviations. In our analysis, we take the protein expression based on KRT19 antibody as the response variable since it has the largest standard deviation. The KRT19 antibody is a type I keratin, also known as Cyfra 21-1, encoded by the KRT19 gene. Due to its high sensitivity, the KRT19 antibody is the most used biomarker for the tumor cells disseminated in lymph nodes, peripheral blood, and bone marrow of breast cancer patients (Nakata et al., 2004).

We first plot the histograms for the KRT19 antibody (protein) expression levels, and kurtosis of the 2000 selected gene expression levels in Figure 2.1. The left panel of Figure 2.1 shows that the distribution of the response variable is asymmetric and bimodal. The right panel shows that 61.65% of the gene expressions have kurtosis larger than three, and 17.1% of the gene expressions have kurtosis larger than nine. In other words, 61.65% of the genes have gene expressions that are heavier tails than the normal distribution, and 17.1% of them have heavier tails than t_5, the t-distribution with five degrees of freedom. This suggests that even after performing normalization, the data can still exhibit heavy-tailedness (Purdom and Holmes, 2005).

We now estimate the conditional distribution of the protein expressions based on KRT19

antibody at the expectile levels $\tau = \{0.25, 0.5, 0.75\}$. We are interested in the solution paths for the first 15 genes that are included in the model for the different expectile levels. Selected genes in solution paths might provide insights for further biological investigations. In particular, we start with a large value of tuning parameter λ and incrementally decrease λ to obtain the first 15 covariates with non-zero regression coefficients. The list of genes are presented in Table 2.6.

Table 2.6. Solution path for NCI-60 dataset

τ	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**
0.25	*ARHGAP29*	*C19orf33*	*BAMBI*	*NRN1*	*TFF3*	*CA2*	*CEMIP*	*IGFBP2*
0.5	*C19orf33*	*ANXA3*	*MAL2*	*BAMBI*	*MALL*	*CA2*	*ARHGAP29*	*NRN1*
0.75	*C19orf33*	*MAL2*	*KRT8*	*ANXA3*	*MALL*	*VAMP8*	*MAGEA12*	*KRT19*
τ	**9**	**10**	**11**	**12**	**13**	**14**	**15**	
0.25	*ANAX3*	*SPARC*	*ALDH1A1*	*BEX1*	*EMP3*	*F3*	*ALDH1A3*	
0.5	*HOXC10*	*CEMIP*	*BEX1*	*TFF3*	*SPARC*	*G0S2*	*PDLIM1*	
0.75	*SPARC*	*CA2*	*NRN1*	*GDA*	*BAMBI*	*AKR1B10*	*GPX3*	

From Table 2.6, we see that six genes are commonly selected across the three different expectile levels: *C19orf33*, *ANXA3*, *SPARC*, *CA2*, *NRN1*, and *BAMBI*. Interestingly, most of the six genes are found to be associated with breast cancer patients' survival time. The gene *ANXA3* is shown to be upregulated, i.e., the cell increases the quantity of the component in response to an external stimulus, in breast cancer tissues and is positively correlated with poor overall survival (Du et al., 2018). Zhou et al. (2017) suggested that silencing of *ANXA3* expression by RNA interference inhibits the proliferation and invasion of breast cancer cells. Fritzmann et al. (2009) showed that the transforming growth factor-β inhibitor *BAMBI* was highly expressed in metastatic primary tumors and metastases, and observed an inverse correlation between level of *BAMBI* expression and metastasis-free survival time of patients. A very recent study in Wen et al. (2020) found that immortalization-upregulated protein, also known as *C19orf33*, was upregulated significantly in breast cancer tissues compared with noncancerous tissue. Watkins et al. (2005) reported that the transcript levels of *SPARC* were found to be significantly higher in tumor tissue when compared to normal background breast tissue, and concluded that *SPARC* plays a crucial role in tumor development in breast cancer and as such has a significant bearing on patient prognosis and long-term survival.

Chapter 3

Extension to Various Penalties

3.1 Introduction to Various Penalties

In this section, we extend the LAMM algorithm introduced in Section 2.4 to a broad class of convex penalties that still inherit the two essential features of the standard lasso, namely the shrinkage and the selection of (groups of) variables.

Consider the general optimization problem

$$\hat{\boldsymbol{\beta}} \in \underset{\boldsymbol{\beta}\in\mathbb{R}^d}{\text{minimize}} \left\{ \frac{1}{n}\sum_{i=1}^{n} L_{\tau,\gamma}(y_i - \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}) + P(\boldsymbol{\beta}) \right\},$$

where $P(\boldsymbol{\beta})$ is a generic convex penalty function and $L_{\tau,\gamma}(\cdot)$ is the `retire` loss that satisfies Condition 1. In this chapter, we focus on the following four widely used convex penalty functions.

1. Weighted lasso (Tibshirani, 1996): $P(\boldsymbol{\beta}) = \sum_{j=1}^{d} \lambda_j |\beta_j|$, where $\lambda_j \geq 0$ for $j = 1, \ldots, d$.

2. Elastic net (Zou and Hastie, 2005): $P(\boldsymbol{\beta}) = \lambda\alpha||\boldsymbol{\beta}||_1 + \lambda(1-\alpha)||\boldsymbol{\beta}||_2^2$, where $\lambda > 0$ is a sparsity-inducing parameter and $\alpha \in (0,1)$ is a user-specified constant that controls the trade-off between the ℓ_1 penalty and the ridge penalty.

3. Group lasso (Yuan and Lin, 2006): $P(\boldsymbol{\beta}) = \lambda \sum_{g=1}^{G} w_g ||\boldsymbol{\beta}_g||_2$, where $\boldsymbol{\beta} = (\boldsymbol{\beta}_1^{\mathrm{T}}, \ldots, \boldsymbol{\beta}_G^{\mathrm{T}})^{\mathrm{T}}$ and $\boldsymbol{\beta}_G$ is a sub-vector of $\boldsymbol{\beta}$ corresponding to the g-th group of coefficients, and $w_g > 0$ are predetermined weights.

4. Sparse group lasso (Simon et al., 2013): $P(\boldsymbol{\beta}) = \lambda||\boldsymbol{\beta}||_1 + \lambda\sum_{g=1}^{G} w_g||\boldsymbol{\beta}_g||_2$.

Remark 3.1.1. *Elastic net penalty can be viewed as a hybrid of ℓ_1 and ℓ_2 penalty, and the quadratic component is beneficial when the features are highly correlated. For instance, in microarray studies, features are often found to be in correlated groups, simply performing lasso will likely end up with erratic and wild behavior of the coefficient paths. To see this, consider an extreme case where two features are identical copies of each other, i.e., $X_j = X_{j'}$. Then we have infinitely many pairs of coefficients $(\hat{\beta}_j, \hat{\beta}_{j'})$ such that $\hat{\beta}_j + \hat{\beta}_{j'}$ equals a constant that depends on the given penalty level λ. When both $\hat{\beta}_j$ and $\hat{\beta}_{j'}$ are positive, the resulting loss function $L_{\tau,\gamma}(\cdot)$ and ℓ_1 penalty remain the same for all pairs of $(\hat{\beta}_j, \hat{\beta}_{j'})$, meaning that the lasso can not differentiate these pairs, which often leads to erratic behavior of coefficient paths. A quadratic penalty, on the other hand, will differentiate these two twins so that strong within-group correlations can be better handled. Meanwhile, the quadratic component adds strict convexity for the penalty function $P(\boldsymbol{\beta})$, which facilitates gradient based algorithms.*

Remark 3.1.2. *Group lasso penalty is designed for selecting (or omitting) all coefficients within a group simultaneously. A leading example is when we have qualitative factors among our predictors. We typically code their levels using a set of dummy variables or contrasts, and would want to include or exclude this group of variables together. It can be checked that the group lasso penalty reduces to the lasso penalty when all the groups are singletons, i.e., $||\beta_j||_2 = |\beta_j|$. Consequently, the group lasso penalty can be viewed as an extension of the lasso penalty when group structures are presented.*

Remark 3.1.3. *Even though the group lasso penalty achieves between-group sparsity, by the nature of ℓ_2-norm, all coefficients in a group are nonzero simultaneously if that group is selected by a group lasso fit. However, sometimes we would emphasize within-group sparsity since not all features in a group are indeed significant. For example, a biological pathway may be implicated in the progression of a particular type of cancer, but not all genes in the pathway need to be active. The sparse group lasso penalty is designed to achieve such within-group sparsity by adding extra*

ℓ_1 penalty to the group lasso penalty. Note that if we restrain on a specific sub-vector $\boldsymbol{\beta}_g$, the sparse group lasso penalty $P(\boldsymbol{\beta}_g) = \lambda||\boldsymbol{\beta}_g||_1 + \lambda w_g||\boldsymbol{\beta}_g||_2$ reduces to a variant of the elastic net penalty. Consequently within-group sparsity is introduced in the same way as the elastic net penalty.

3.2 Computational Mehtods

In this section we extend the LAMM algorithm mentioned in Section 2.4 to accommodate the aforementioned convex penalties. Recall that our target is to solve the general optimization problem

$$\hat{\boldsymbol{\beta}} \in \underset{\boldsymbol{\beta}\in\mathbb{R}^d}{\text{minimize}} \left\{ \frac{1}{n}\sum_{i=1}^{n} L_{\tau,\gamma}(y_i - \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}) + P(\boldsymbol{\beta}) \right\}. \tag{3.1}$$

Let $\mathscr{R}_n(\boldsymbol{\beta}) = n^{-1}\sum_{i=1}^{n} L_{\tau,\gamma}(y_i - \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta})$ and $\nabla\mathscr{R}_n(\boldsymbol{\beta})$ be its gradient. Following the principal of the LAMM algorithm, we locally majorize $\mathscr{R}_n(\boldsymbol{\beta})$ at $\hat{\boldsymbol{\beta}}^{(k-1)}$ by constructing an isotropic quadratic function $G_n(\cdot)$ of the form

$$G_n(\boldsymbol{\beta}|\phi_k, \hat{\boldsymbol{\beta}}^{(k-1)}) = \mathscr{R}_n(\hat{\boldsymbol{\beta}}^{(k-1)}) + \langle \nabla\mathscr{R}_n(\hat{\boldsymbol{\beta}}^{(k-1)}), \boldsymbol{\beta} - \hat{\boldsymbol{\beta}}^{(k-1)} \rangle + \frac{\phi_k}{2}||\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}^{(k-1)}||_2^2,$$

where $\phi_k > 0$ is a quadratic parameter to be determined at the k-th iteration. Then define the k-th iterate $\hat{\boldsymbol{\beta}}^{(k)}$ as the solution to

$$\underset{\boldsymbol{\beta}\in\mathbb{R}^d}{\text{minimize}}\ G_n(\boldsymbol{\beta}|\phi_k, \hat{\boldsymbol{\beta}}^{(k-1)}) + P(\boldsymbol{\beta}).$$

By the first-order optimization condition, $\hat{\boldsymbol{\beta}}^{(k)}$ satisfies

$$\mathbf{0} \in \nabla\mathscr{R}_n(\hat{\boldsymbol{\beta}}^{(k-1)}) + \phi_k(\hat{\boldsymbol{\beta}}^{(k)} - \hat{\boldsymbol{\beta}}^{(k-1)}) + \partial P(\boldsymbol{\beta})\big|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}^{(k)}},$$

where ∂P denotes the subdifferential of $P : \mathbb{R}^d \to [0,\infty)$. In practice, we often leave the intercept

term unpenalized, and its update rule takes a simple form $\hat{\beta}_1^{(k)} = \hat{\beta}_1^{(k-1)} - \phi_k^{-1}\nabla_{\beta_1}\mathscr{R}_n(\hat{\boldsymbol{\beta}}^{(k-1)})$. The update rules for coefficients other than intercept can be derived similarly as in Section 2.4, and we postpone their derivations to the Appendix C.1. For all four types of penalties, we obtain explicit formulas for each iteration, thus $\hat{\boldsymbol{\beta}}^{(k)}$ can be updated efficiently by vector-matrix multiplications. Recall that $S(a,b) = \text{sign}(a)\cdot(|a|-b)_+$ denotes the shrinkage operator, and $\text{sign}(\cdot)$ is the sign function and $(c)_+ = \max(c,0)$. We summarize the whole procedure in the following Algorithm 6.

Algorithm 6. Local Adaptive Majorize-minimization (LAMM) Algorithm for Solving (3.1) with various convex penalties.

Input: regularization parameters λ_j, expectile level τ, Huber loss tuning parameter γ, inflation factor $\Gamma = 1.25$ and convergence criterion ε.

Input(optional): hybrid level α, group structure $(1,\ldots,G)$ and group weight $(w_1,\ldots,w_G)$.

Initialization: $\hat{\boldsymbol{\beta}}^{(0)} = 0$, $\phi_0 = 0.01$.

Iterate: the following until the stopping criterion $\|\hat{\boldsymbol{\beta}}^{(k)} - \hat{\boldsymbol{\beta}}^{(k-1)}\|_2 \le \varepsilon$ is met, where $\hat{\boldsymbol{\beta}}^{(k)}$ is the value of $\boldsymbol{\beta}$ obtained at the k-th iteration.

1. Set $\phi_k \leftarrow \max(\phi_0, \phi_{k-1}/\Gamma)$.
2. **repeat**
3. $\hat{\beta}_1^{(k)} \leftarrow \hat{\beta}_1^{(k-1)} - \phi_k^{-1}\nabla_{\beta_1}\mathscr{R}_n(\hat{\boldsymbol{\beta}}^{(k-1)})$.
4. for $j = 2,\ldots,d$ (or $g = 2,\ldots,G$), update $\hat{\beta}_j^{(k)}$ (or $\hat{\boldsymbol{\beta}}_g^{(k)}$) as follows

weighted lasso	$\hat{\beta}_j^{(k)} \leftarrow S\{\hat{\beta}_j^{(k-1)} - \phi_k^{-1}\nabla_{\beta_j}\mathscr{R}_n(\hat{\boldsymbol{\beta}}^{(k-1)}), \phi_k^{-1}\lambda_j\}$.
elastic net	$\hat{\beta}_j^{(k)} \leftarrow \frac{1}{1+2\phi_k^{-1}\lambda(1-\alpha)} S\{\hat{\beta}_j^{(k-1)} - \phi_k^{-1}\nabla_{\beta_j}\mathscr{R}_n(\hat{\boldsymbol{\beta}}^{(k-1)}), \phi_k^{-1}\lambda\alpha\}$.
group lasso	$\hat{\boldsymbol{\beta}}_g^{(k)} \leftarrow \left\{\hat{\boldsymbol{\beta}}_g^{(k-1)} - \phi_k^{-1}\nabla_{\boldsymbol{\beta}_g}\mathscr{R}_n(\hat{\boldsymbol{\beta}}^{(k-1)})\right\}\cdot\left(1 - \frac{\lambda w_g}{\phi_k\|\hat{\boldsymbol{\beta}}_g^{(k-1)} - \phi_k^{-1}\nabla_{\boldsymbol{\beta}_g}\mathscr{R}_n(\hat{\boldsymbol{\beta}}^{(k-1)})\|_2}\right)_+$.
sparse group lasso	$\hat{\boldsymbol{\beta}}_g^{(k)} \leftarrow S\left\{\hat{\boldsymbol{\beta}}_g^{(k-1)} - \phi_k^{-1}\nabla_{\boldsymbol{\beta}_g}\mathscr{R}_n(\hat{\boldsymbol{\beta}}^{(k-1)}), \phi_k^{-1}\lambda\right\}\cdot\left(1 - \frac{\lambda w_g}{\phi_k\|S\{\hat{\boldsymbol{\beta}}_g^{(k-1)} - \phi_k^{-1}\nabla_{\boldsymbol{\beta}_g}\mathscr{R}_n(\hat{\boldsymbol{\beta}}^{(k-1)}), \phi_k^{-1}\lambda\}\|_2}\right)_+$.

5. **if** $\mathscr{R}_n(\hat{\boldsymbol{\beta}}^{(k)}) > G_n(\hat{\boldsymbol{\beta}}^{(k)}|\phi_k, \hat{\boldsymbol{\beta}}^{(k-1)})$, set $\phi_k \leftarrow \Gamma\phi_k$.
6. **until** $\mathscr{R}_n(\hat{\boldsymbol{\beta}}^{(k)}) \le G_n(\hat{\boldsymbol{\beta}}^{(k)}|\phi_k, \hat{\boldsymbol{\beta}}^{(k-1)})$.

Output: the final iterate $\hat{\boldsymbol{\beta}}^{(k)}$.

3.3 Numerical Experiments

In this section, we perform extensive numerical studies to access the performance of the proposed penalized `retire` estimator with four convex penalties, the lasso, elastic net, group lasso and sparse group lasso. For all of the numerical studies, we generate the covariates $\boldsymbol{x}_i$ from a multivariate normal distribution $N(\mathbf{0}, \boldsymbol{\Sigma} = (\sigma_{jk})_{1\leq j,k\leq d})$ with correlated (or block correlated) $\boldsymbol{\Sigma}$. And we generate the response variable y_i from the following model:

$$\text{Expectile heteroscedastic model: } y_i = \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}^* + (0.5|x_{id}| + 0.5)\{\varepsilon_i - e_\tau(\varepsilon_i)\}, \tag{3.2}$$

where ε_i is the random noise, and $e_\tau(\varepsilon_i)$ denotes the inverse of the expectile function of ε_i. The random noise is generated from either a Gaussian distribution, $N(0,2)$, or a t distribution with 2.1 degrees of freedom. Moreover, we consider two expectile levels $\tau = \{0.5, 0.8\}$ to access the performance under asymmetry data. Lastly, we adaptively update the robustification parameter γ using a heuristic tuning method as detailed in Section 2.6

$$\gamma^k = \mathrm{mad}(\tilde{\mathbf{r}}^k) \cdot \sqrt{\frac{n}{\log(nd)}}. \tag{3.3}$$

This heuristic approach works well across different scenarios throughout the section. Our computational results are reproducible using codes available from https://github.com/ZianWang0128/Retire.

In Subsection 3.3.1, we fit the penalized `retire` estimator with the ℓ_1 and elastic net penalties on simulated data with two types of non-grouped regression coefficients, the sparse $\boldsymbol{\beta}^*$ and the dense $\boldsymbol{\beta}^*$. In Subsection 3.3.2, we fit the penalized `retire` estimator with the group lasso and spares group lasso penalties on simulated data with two types of grouped regression coefficients, the grouped $\boldsymbol{\beta}^*$ and the sparse grouped $\boldsymbol{\beta}^*$.

3.3.1 Simulated Data with Non-grouped Regression Coefficients

In this subsection we consider the covariates $\boldsymbol{x}_i \in \mathbb{R}^d$ from a multivariate normal distribution $N(\mathbf{0}, \boldsymbol{\Sigma} = (\sigma_{jk})_{1 \le j,k \le d})$ with $\sigma_{jk} = 0.7^{|j-k|}$. And we consider two types of regression coefficients $\boldsymbol{\beta}^* = (\beta_1^*, \beta_2^*, \ldots, \beta_d^*)^{\mathrm{T}}$.

1. Sparse $\boldsymbol{\beta}^*$: $\beta_1^* = 4$ (intercept), $\beta_j^* = \{1.8, 1.6, 1.4, 1.2, 1, -1, -1.2, -1.4, -1.6, -1.8\}$ for $j = 2, 4, \ldots, 20$, and 0 otherwise.

2. Dense $\boldsymbol{\beta}^*$: $\beta_1^* = 4$ (intercept), $\beta_j^* = 0.8$ for $j = 2, \ldots, 100$, and 0 otherwise.

We implement the ℓ_1-penalized `retire` and the elastic net-penalized `retire` with three α levels ($\alpha \in \{0.2, 0.5, 0.8\}$). We add the ℓ_1-penalized asymmetric least squares regression (`sales`) proposed by Gu and Zou (2016) as comparison. To assess the performance across different methods, we report the estimation error under the ℓ_2-norm, i.e., $\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2$, the true positive rate (TPR), and the false positive rate (FPR). Here, TPR is defined as the proportion of the number of correctly identified non-zeros and the false positive rate is calculated as the proportion of the number of incorrectly identified nonzeros. For all methods, the sparsity inducing tuning parameter λ is selected via ten-fold cross-validation. Specifically, we select the largest tuning parameter that yields a value of the asymmetric least squares loss that is less than the minimum of the asymmetric least squares loss plus one standard error. Also note that both ℓ_1-penalized `retire` and elastic net-penalized `retire` require tuning an additional robustness parameter γ. We select γ using the heuristic tuning method (3.3) to update γ at the beginning of each iteration in Algorithm 6.

The results for the sparse $\boldsymbol{\beta}^*$ and dense $\boldsymbol{\beta}^*$ under both the moderate- ($n = 400,\ d = 200$) and high-dimensional ($n = 400,\ d = 500$) settings, averaged over 100 repetitions, are reported in Table 3.1 and Table 3.2, respectively.

Table 3.1 contains results for the Expectile heteroscedastic model (3.2) under sparse $\boldsymbol{\beta}^*$. We see that both ℓ_1 `retire` and ℓ_1 `sales` outperform the elastic net-penalized `retire` in all

Table 3.1. Expectile heteroscedastic model (3.2) under Sparse $\boldsymbol{\beta}^*$. Estimation error under ℓ_2-norm (and its standard deviation), true positive rate (TPR) and false positive rate (FPR), averaged over 100 repetitions, are reported.

Expectile heteroscedastic model (3.2) with Sparse $\boldsymbol{\beta}^*$ and $\tau = 0.5$							
		$n = 400, d = 200$			$n = 400, d = 500$		
Noise	Method	ℓ_2 error	TPR	FPR	ℓ_2 error	TPR	FPR
$N(0,2)$	ℓ_1 retire	0.656 (0.107)	1.000 (0.000)	0.028 (0.015)	0.706 (0.124)	1.000 (0.000)	0.012 (0.007)
	ℓ_1 sales	0.663 (0.108)	1.000 (0.000)	0.029(0.015)	0.716 (0.129)	1.000 (0.000)	0.012 (0.007)
	elastic net retire ($\alpha = 0.2$)	1.897 (0.161)	1.000 (0.000)	0.631 (0.061)	2.433 (0.103)	1.000 (0.000)	0.446 (0.045)
	elastic net retire ($\alpha = 0.5$)	1.503 (0.170)	1.000 (0.000)	0.272 (0.060)	1.816 (0.139)	1.000 (0.000)	0.157 (0.035)
	elastic net retire ($\alpha = 0.8$)	1.026 (0.152)	1.000 (0.000)	0.101 (0.030)	1.151 (0.147)	1.000 (0.000)	0.049 (0.019)
$t_{2.1}$	ℓ_1 retire	1.317 (0.336)	0.962 (0.069)	0.015 (0.008)	1.320 (0.364)	0.961 (0.071)	0.006 (0.003)
	ℓ_1 sales	1.442 (0.361)	0.949 (0.073)	0.017 (0.008)	1.465 (0.397)	0.952 (0.072)	0.007 (0.004)
	elastic net retire ($\alpha = 0.2$)	2.575 (0.275)	1.000 (0.000)	0.415 (0.107)	2.816 (0.168)	1.000 (0.000)	0.313 (0.073)
	elastic net retire ($\alpha = 0.5$)	2.283 (0.331)	1.000 (0.000)	0.122 (0.054)	2.412 (0.274)	1.000 (0.000)	0.066 (0.038)
	elastic net retire ($\alpha = 0.8$)	1.812 (0.385)	0.989 (0.035)	0.052 (0.013)	1.849 (0.375)	0.991 (0.029)	0.022 (0.007)

Expectile heteroscedastic model (3.2) with Sparse $\boldsymbol{\beta}^*$ and $\tau = 0.8$							
		$n = 400, d = 200$			$n = 400, d = 500$		
Noise	Method	ℓ_2 error	TPR	FPR	ℓ_2 error	TPR	FPR
$N(0,2)$	ℓ_1 retire	0.708 (0.124)	1.000 (0.000)	0.031 (0.015)	0.769 (0.143)	1.000 (0.000)	0.014 (0.009)
	ℓ_1 sales	0.707 (0.124)	1.000 (0.000)	0.031 (0.015)	0.776 (0.139)	1.000 (0.000)	0.013 (0.008)
	elastic net retire ($\alpha = 0.2$)	2.012 (0.174)	1.000 (0.000)	0.622 (0.064)	2.565 (0.109)	1.000 (0.000)	0.421 (0.038)
	elastic net retire ($\alpha = 0.5$)	1.593 (0.182)	1.000 (0.000)	0.276 (0.056)	1.941 (0.155)	1.000 (0.000)	0.152 (0.029)
	elastic net retire ($\alpha = 0.8$)	1.091 (0.161)	1.000 (0.000)	0.101 (0.029)	1.232 (0.172)	1.000 (0.000)	0.052 (0.020)
$t_{2.1}$	ℓ_1 retire	1.635 (0.548)	0.928 (0.090)	0.014 (0.008)	1.553 (0.534)	0.936 (0.089)	0.005 (0.003)
	ℓ_1 sales	1.791 (0.554)	0.913 (0.093)	0.017 (0.008)	1.708 (0.542)	0.927 (0.089)	0.008 (0.004)
	elastic net retire ($\alpha = 0.2$)	2.841 (0.359)	1.000 (0.000)	0.376 (0.012)	3.025 (0.265)	1.000 (0.000)	0.295 (0.084)
	elastic net retire ($\alpha = 0.5$)	2.581 (0.471)	0.991 (0.029)	0.111 (0.059)	2.655 (0.389)	0.993 (0.029)	0.064 (0.038)
	elastic net retire ($\alpha = 0.8$)	2.122 (0.565)	0.964 (0.069)	0.051 (0.015)	2.095 (0.523)	0.965 (0.066)	0.022 (0.010)

three metrics. ℓ_1 `retire` and ℓ_1 `sales` perform almost identically under normal noise, while the former gains a little advantage over the latter under t-distributed noise, which is probably due to the extra robustness of `retire` loss over the asymmetric square loss. The performance of elastic net-penalized `retire` deteriorates as α decreases due to the large number of zeros in the sparse $\boldsymbol{\beta}^*$. In conclusion, the simulation suggests that ℓ_1-penalized `retire` is the most suitable method under spare $\boldsymbol{\beta}^*$ setting.

Table 3.2 contains results for the Expectile heteroscedastic model (3.2) under dense $\boldsymbol{\beta}^*$. The performance of elastic net-penalized `retire` deteriorates as α decreases. When $\alpha = 0.8$, the elastic net-penalized `retire` admits lower ℓ_2 errors while maintaining the highest TPR and comparable FPR over ℓ_1-penalized methods, suggesting that the elastic net penalty may be beneficial when the true signals are dense and the signal-to-noise ratio is relatively low.

Table 3.2. Expectile heteroscedastic model (3.2) under Dense $\boldsymbol{\beta}^*$. Estimation error under ℓ_2-norm (and its standard deviation), true positive rate (TPR) and false positive rate (FPR), averaged over 100 repetitions, are reported.

Expectile heteroscedastic model (3.2) with Dense $\boldsymbol{\beta}^*$ and $\tau = 0.5$							
		$n = 400, d = 200$			$n = 400, d = 500$		
Noise	Method	ℓ_2 error	TPR	FPR	ℓ_2 error	TPR	FPR
$N(0,2)$	ℓ_1 retire	1.352 (0.108)	1.000 (0.000)	0.097 (0.051)	1.482 (0.162)	1.000 (0.000)	0.049 (0.025)
	ℓ_1 sales	1.352 (0.107)	1.000 (0.000)	0.097 (0.051)	1.484 (0.164)	1.000 (0.000)	0.049 (0.024)
	elastic net retire ($\alpha = 0.2$)	1.043 (0.076)	1.000 (0.000)	0.677 (0.049)	1.469 (0.095)	1.000 (0.000)	0.614 (0.040)
	elastic net retire ($\alpha = 0.5$)	1.004 (0.073)	1.000 (0.000)	0.305 (0.060)	1.141 (0.087)	1.000 (0.000)	0.228 (0.033)
	elastic net retire ($\alpha = 0.8$)	1.144 (0.081)	1.000 (0.000)	0.136 (0.057)	1.226 (0.114)	1.000 (0.000)	0.84 (0.028)
$t_{2.1}$	ℓ_1 retire	2.156 (0.499)	0.998 (0.008)	0.036 (0.038)	2.269 (0.568)	0.995 (0.020)	0.018 (0.016)
	ℓ_1 sales	2.574 (0.765)	0.991 (0.026)	0.053 (0.040)	2.679 (0.845)	0.986 (0.048)	0.028 (0.018)
	elastic net retire ($\alpha = 0.2$)	1.276 (0.193)	1.000 (0.000)	0.576 (0.087)	1.696 (0.223)	1.000 (0.000)	0.545 (0.061)
	elastic net retire ($\alpha = 0.5$)	1.282 (0.207)	1.000 (0.000)	0.198 (0.069)	1.409 (0.265)	1.000 (0.000)	0.158 (0.049)
	elastic net retire ($\alpha = 0.8$)	1.553 (0.248)	1.000 (0.000)	0.058 (0.043)	1.611 (0.310)	1.000 (0.000)	0.037 (0.025)

Expectile heteroscedastic model (3.2) with Dense $\boldsymbol{\beta}^*$ and $\tau = 0.8$							
		$n = 400, d = 200$			$n = 400, d = 500$		
Noise	Method	ℓ_2 error	TPR	FPR	ℓ_2 error	TPR	FPR
$N(0,2)$	ℓ_1 retire	1.429 (0.121)	1.000 (0.000)	0.109 (0.048)	1.574 (0.167)	1.000 (0.000)	0.048 (0.024)
	ℓ_1 sales	1.422 (0.122)	1.000 (0.000)	0.109 (0.048)	1.566 (0.167)	1.000 (0.000)	0.049 (0.024)
	elastic net retire ($\alpha = 0.2$)	1.083 (0.072)	1.000 (0.000)	0.679 (0.051)	1.472 (0.086)	1.000 (0.000)	0.594 (0.037)
	elastic net retire ($\alpha = 0.5$)	1.056 (0.072)	1.000 (0.000)	0.317 (0.065)	1.182 (0.086)	1.000 (0.000)	0.217 (0.030)
	elastic net retire ($\alpha = 0.8$)	1.210 (0.085)	1.000 (0.000)	0.152 (0.053)	1.294 (0.115)	1.000 (0.000)	0.078 (0.024)
$t_{2.1}$	ℓ_1 retire	2.596 (0.636)	0.994 (0.017)	0.044 (0.041)	2.574 (0.584)	0.995 (0.012)	0.024 (0.020)
	ℓ_1 sales	2.992 (0.940)	0.981 (0.041)	0.060 (0.046)	2.943 (0.843)	0.985 (0.031)	0.024 (0.020)
	elastic net retire ($\alpha = 0.2$)	1.520 (0.389)	1.000 (0.000)	0.584 (0.085)	1.820 (0.262)	1.000 (0.000)	0.551 (0.050)
	elastic net retire ($\alpha = 0.5$)	1.551 (0.404)	1.000 (0.000)	0.200 (0.069)	1.588 (0.307)	1.000 (0.000)	0.171 (0.045)
	elastic net retire ($\alpha = 0.8$)	1.871 (0.405)	1.000 (0.000)	0.065 (0.049)	1.838 (0.350)	1.000 (0.000)	0.043 (0.025)

3.3.2 Simulated Data with Grouped Regression Coefficients

In this subsection we consider the covariates $\boldsymbol{x}_i \in \mathbb{R}^d$ from a multivariate normal distribution $N(\mathbf{0}, \boldsymbol{\Sigma})$ with block diagonal covariance matrix $\boldsymbol{\Sigma} \in \mathbb{R}^{d \times d}$. More specifically, $\boldsymbol{\Sigma} = \text{diag}(\boldsymbol{\Sigma}_1, \ldots, \boldsymbol{\Sigma}_{15})$, where $\boldsymbol{\Sigma}_1, \boldsymbol{\Sigma}_2 \in R^{5\times 5}$, $\boldsymbol{\Sigma}_3, \boldsymbol{\Sigma}_4, \boldsymbol{\Sigma}_5 \in R^{10\times 10}$, and $\boldsymbol{\Sigma}_6, \ldots, \boldsymbol{\Sigma}_{15} \in R^{\frac{d-40}{10} \times \frac{d-40}{10}}$. Each block is an exchangeable covariance matrix with diagonal 1 and off-diagonal elements 0.6, meaning that coefficients within each group is highly correlated, but coefficients between groups are uncorrelated. And we consider two types of regression coefficients $\boldsymbol{\beta}^* = (\boldsymbol{\beta}_1^{*\mathrm{T}}, \boldsymbol{\beta}_2^{*\mathrm{T}}, \ldots, \boldsymbol{\beta}_G^{*\mathrm{T}})^{\mathrm{T}}$.

1. Group $\boldsymbol{\beta}^*$: $\beta_0^* = 4$ (intercept), $\boldsymbol{\beta}_1^* = \mathbf{2} \in \mathbb{R}^5, \boldsymbol{\beta}_2^* = \mathbf{1.6} \in \mathbb{R}^5, \boldsymbol{\beta}_3^* = -\mathbf{2} \in \mathbb{R}^{10}, \boldsymbol{\beta}_4^* = \mathbf{1} \in \mathbb{R}^{10}, \boldsymbol{\beta}_5^* = \mathbf{0.6} \in \mathbb{R}^{10}$ and $\boldsymbol{\beta}_6^* = \cdots = \boldsymbol{\beta}_{15}^* = \mathbf{0}$.

2. Sparse group $\boldsymbol{\beta}^*$: $\beta_0^* = 4$ (intercept), $\boldsymbol{\beta}_1^* = (2,2,0,0,0)^{\mathrm{T}} \in \mathbb{R}^5, \boldsymbol{\beta}_2^* = (1.6,1.6,0,0,0)^{\mathrm{T}} \in \mathbb{R}^5, \boldsymbol{\beta}_3^* = (-2,\ldots,-2,0,\ldots,0)^{\mathrm{T}} \in \mathbb{R}^{10}, \boldsymbol{\beta}_4^* = (1,\ldots,1,0,\ldots,0)^{\mathrm{T}} \in \mathbb{R}^{10}, \boldsymbol{\beta}_5^* = (0.6,\ldots, 0.6,0,\ldots,0)^{\mathrm{T}} \in \mathbb{R}^{10}$ and $\boldsymbol{\beta}_6^* = \cdots = \boldsymbol{\beta}_{15}^* = \mathbf{0}$. The first half of the signals in each group

are nonzeros, while the other half of signals in each group are zeros.

We implement the group lasso-penalized `retire` and the sparse group lasso-penalized `retire` with the weights $w_g = \sqrt{|\boldsymbol{\beta}_g|}$, where $|\boldsymbol{\beta}_g|$ is the dimension of the sub-vector $\boldsymbol{\beta}_g$. As a comparison, we add the sparse group lasso-penalized least square regression estimator computed by the R package `SGL`. Also, we compute the ℓ_1-penalized `retire` estimator that utilizes no group structure information as the benchmark.

To assess the performance across different methods, we report the estimation error under the ℓ_2-norm, i.e., $\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2$, the group true positive rate (group TPR), and the group false positive rate (group FPR). Here, group TPR is defined as the proportion of groups that are correctly estimated to contain non-zeros, and the group FPR is calculated as the proportion of groups that are incorrectly estimated to contain non-zeros. For all methods, the sparsity inducing tuning parameter λ is selected via ten-fold cross-validation. Specifically, for ℓ_1/group lasso/sparse group lasso-penalized `retire`, we select the largest tuning parameter that yields a value of the asymmetric least squares loss that is less than the minimum of the asymmetric least squares loss plus one standard error. For the estimator from `SGL` package, we select the largest tuning parameter that yields a value of the least squares loss that is less than the minimum of the asymmetric least squares loss plus one standard error. Also note that all penalized `retire` regressions require tuning an additional robustness parameter γ. We select γ using the heuristic tuning method (3.3) to update γ at the beginning of each iteration in Algorithm 6.

The results for the sparse $\boldsymbol{\beta}^*$ and dense $\boldsymbol{\beta}^*$ under both the moderate- ($n = 400,\ d = 200$) and high-dimensional ($n = 400,\ d = 500$) settings, averaged over 100 repetitions, are reported in Table 3.3 and Table 3.4, respectively.

Table 3.3 contains results for the Expectile heteroscedastic model (3.2) under group $\boldsymbol{\beta}^*$. We see that as a benchmark, ℓ_1 `retire` fails to control group FPR, indicating its inappropriateness when the true signal $\boldsymbol{\beta}^*$ possesses group structures. Also, we see that group lasso-penalized `retire` has the lowest ℓ_2 errors across all scenario. The other two sparse group lasso-penalized

Table 3.3. Expectile heteroscedastic model (3.2) under Group $\boldsymbol{\beta}^*$. Estimation error under ℓ_2-norm (and its standard deviation), group true positive rate (group TPR) and group false positive rate (group FPR), averaged over 100 repetitions, are reported.

Expectile heteroscedastic model (3.2) with Group $\boldsymbol{\beta}^*$ and $\tau = 0.5$							
		$n = 400, d = 200$			$n = 400, d = 500$		
Noise	Method	ℓ_2 error	Group TPR	Group FPR	ℓ_2 error	Group TPR	Group FPR
$N(0,2)$	ℓ_1 retire	0.716 (0.094)	1.000 (0.000)	0.182 (0.065)	0.729 (0.088)	1.000 (0.000)	0.248 (0.183)
	group lasso retire	0.536 (0.063)	1.000 (0.000)	0.010 (0.036)	0.536 (0.060)	1.000 (0.000)	0.007 (0.036)
	sparse group lasso retire	0.614 (0.072)	1.000 (0.000)	0.007 (0.029)	0.615 (0.068)	1.000 (0.000)	0.000 (0.000)
	SGL package	1.075 (0.525)	1.000 (0.000)	0.006 (0.024)	1.062 (0.512)	1.000 (0.000)	0.006 (0.024)
$t_{2.1}$	ℓ_1 retire	1.256 (0.489)	1.000 (0.000)	0.023 (0.068)	1.188 (0.310)	1.000 (0.000)	0.028 (0.073)
	group lasso retire	0.856 (0.332)	1.000 (0.000)	0.001 (0.010)	0.807 (0.209)	1.000 (0.000)	0.001 (0.010)
	sparse group lasso retire	0.966 (0.342)	1.000 (0.000)	0.001 (0.010)	0.912 (0.215)	1.000 (0.000)	0.001 (0.010)
	SGL package	1.495 (0.672)	1.000 (0.000)	0.000 (0.000)	1.493 (0.653)	0.998 (0.020)	0.000 (0.000)

Expectile heteroscedastic model (3.2) with group $\boldsymbol{\beta}^*$ and $\tau = 0.8$							
		$n = 400, d = 200$			$n = 400, d = 500$		
Noise	Method	ℓ_2 error	Group TPR	Group FPR	ℓ_2 error	Group TPR	Group FPR
$N(0,2)$	ℓ_1 retire	0.778 (0.104)	1.000 (0.000)	0.176 (0.148)	0.797 (0.106)	1.000 (0.000)	0.229 (0.183)
	group lasso retire	0.574 (0.066)	1.000 (0.000)	0.012 (0.038)	0.573 (0.070)	1.000 (0.000)	0.006 (0.031)
	sparse group lasso retire	0.680 (0.077)	1.000 (0.000)	0.001 (0.010)	0.680 (0.078)	1.000 (0.000)	0.001 (0.010)
	SGL package	1.262 (0.691)	1.000 (0.000)	0.005 (0.022)	1.299 (0.651)	1.000 (0.000)	0.005 (0.022)
$t_{2.1}$	ℓ_1 retire	1.660 (0.924)	1.000 (0.000)	0.039 (0.087)	1.434 (0.377)	1.000 (0.000)	0.033 (0.073)
	group lasso retire	1.166 (0.827)	1.000 (0.000)	0.001 (0.010)	0.968 (0.330)	1.000 (0.000)	0.001 (0.010)
	sparse group lasso retire	1.294 (0.794)	1.000 (0.000)	0.001 (0.010)	1.101 (0.309)	1.000 (0.000)	0.000 (0.000)
	SGL package	1.746 (0.802)	1.000 (0.000)	0.000 (0.000)	1.777 (0.784)	0.998 (0.020)	0.000 (0.000)

methods (sparse group lasso `retire` and `SGL` package) tend to have slightly lower group FPRs, but the gain over group lasso `retire` is barely marginal. When $\tau = 0.8$, the performance of estimators from `SGL` package deteriorates since it implicitly assumes $\tau = 0.5$ and there is a non-negligible bias when $\tau = 0.8$. The simulation results suggest that when the signal $\boldsymbol{\beta}^*$ is genuinely group structured, group lasso-penalized `retire` may be the most suitable method.

Table 3.4 contains results for the Expectile heteroscedastic model (3.2) under sparse group $\boldsymbol{\beta}^*$. We see that both the ℓ_1-penalized `retire` and the group lasso-penalized `retire` fail to control group FPR when within-group sparsity presents. Both the sparse group lasso-penalized `retire` and the `SGL` package perform fairly well, while the former has slight advantages over the latter in all three facets, especially when under t-distributed noises. This is not surprising since `retire` loss is a robust modification of the asymmetric square loss used in the `SGL` package. When $\tau = 0.8$, the performance of estimators from `SGL` package deteriorates since it implicitly assumes $\tau = 0.5$ and there is a non-negligible bias when $\tau = 0.8$. The simulation results suggest that when the signal $\boldsymbol{\beta}^*$ possesses within-group sparsity, sparse group lasso-penalized `retire`

Table 3.4. Expectile heteroscedastic model (3.2) under Sparse Group $\boldsymbol{\beta}^*$. Estimation error under ℓ_2-norm (and its standard deviation), group true positive rate (group TPR) and group false positive rate (group FPR), averaged over 100 repetitions, are reported.

Expectile heteroscedastic model (3.2) with Sparse Group $\boldsymbol{\beta}^*$ and $\tau = 0.5$							
		$n = 400, d = 200$			$n = 400, d = 500$		
Noise	Method	ℓ_2 error	Group TPR	Group FPR	ℓ_2 error	Group TPR	Group FPR
$N(0,2)$	ℓ_1 retire	0.599 (0.087)	1.000 (0.000)	0.117 (0.129)	0.618 (0.082)	1.000 (0.000)	0.158 (0.158)
	group lasso retire	1.024 (0.122)	1.000 (0.000)	0.227 (0.156)	1.134 (0.126)	1.000 (0.000)	0.167 (0.121)
	sparse group lasso retire	0.892 (0.106)	1.000 (0.000)	0.059 (0.077)	0.938 (0.114)	1.000 (0.000)	0.035 (0.058)
	SGL package	0.982 (0.200)	1.000 (0.000)	0.106 (0.112)	0.998 (0.176)	1.000 (0.000)	0.098 (0.092)
$t_{2.1}$	ℓ_1 retire	1.134 (0.452)	1.000 (0.000)	0.007 (0.036)	1.080 (0.375)	1.000 (0.000)	0.008 (0.031)
	group lasso retire	2.048 (0.680)	1.000 (0.000)	0.027 (0.058)	2.000 (0.607)	1.000 (0.000)	0.028 (0.064)
	sparse group lasso retire	1.758 (0.636)	0.998 (0.020)	0.003 (0.022)	1.689 (0.534)	1.000 (0.000)	0.003 (0.017)
	SGL package	2.063 (0.891)	0.980 (0.119)	0.011 (0.035)	2.015 (0.792)	0.990 (0.100)	0.008 (0.034)

Expectile heteroscedastic model (3.2) with Sparse Group $\boldsymbol{\beta}^*$ and $\tau = 0.8$							
		$n = 400, d = 200$			$n = 400, d = 500$		
Noise	Method	ℓ_2 error	Group TPR	Group FPR	ℓ_2 error	Group TPR	Group FPR
$N(0,2)$	ℓ_1 retire	0.658 (0.096)	1.000 (0.000)	0.126 (0.143)	0.678 (0.098)	1.000 (0.000)	0.165 (0.161)
	group lasso retire	1.124 (0.121)	1.000 (0.000)	0.215 (0.153)	1.235 (0.134)	1.000 (0.000)	0.148 (0.111)
	sparse group lasso retire	1.011 (0.114)	1.000 (0.000)	0.040 (0.070)	1.049 (0.125)	1.000 (0.000)	0.028 (0.049)
	SGL package	1.209 (0.327)	1.000 (0.000)	0.108 (0.114)	1.247 (0.306)	1.000 (0.000)	0.092 (0.095)
$t_{2.1}$	ℓ_1 retire	1.491 (0.845)	0.990 (0.059)	0.015 (0.044)	1.301 (0.472)	1.000 (0.000)	0.014 (0.043)
	group lasso retire	2.427 (0.977)	0.988 (0.074)	0.028 (0.062)	2.263 (0.657)	1.000 (0.000)	0.030 (0.069)
	sparse group lasso retire	2.136 (1.027)	0.984 (0.101)	0.009 (0.032)	1.935 (0.595)	1.000 (0.000)	0.005 (0.026)
	SGL package	2.289 (0.858)	0.978 (0.127)	0.009 (0.032)	2.249 (0.781)	0.990 (0.100)	0.005 (0.033)

may be more favorable when group FPR is more emphasized than ℓ_2 error.

Appendix A

Supplementary Material for Chapter 1

A.1 Preliminary Results

Given $\tau \in (0,1)$, let $\{(y_i, \boldsymbol{x}_i)\}_{i=1}^n$ be a sample of independent data vectors from the linear regression model in (1.1), $y_i = \boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{\beta}^*(\tau) + \varepsilon_i(\tau)$, where $\varepsilon_i(\tau)$ satisfies $e_\tau(\varepsilon_i | \boldsymbol{x}_i) = 0$. In other words, the conditional τ-mean of y_i give $\boldsymbol{x}_i$ is a linear combination of $\boldsymbol{x}_i$. We suppress the dependency of $\boldsymbol{\beta}^*(\tau)$ and $\varepsilon(\tau)$ on τ throughout the Appendix. Let $w_\tau(u) := |\tau - \mathbb{1}(u < 0)|$ and let $\ell_\gamma(u) = \gamma^2 \ell(u/\gamma)$. Recall from (1.5) that $L(u) := L_{\tau,\gamma}(u) = w_\tau(u)\ell_\gamma(u)$ and let

$$\mathscr{R}_n(\boldsymbol{\beta}) = \frac{1}{n}\sum_{i=1}^n L(y_i - \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}) \quad \text{and} \quad \nabla\mathscr{R}_n(\boldsymbol{\beta}) = -\frac{1}{n}\sum_{i=1}^n L'(y_i - \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta})\boldsymbol{x}_i,$$

where $L'(u) = \gamma w_\tau(u)\ell'(u/\gamma)$ is the first-order derivative of $L(u)$.

For $\boldsymbol{\beta} \in \mathbb{R}^d$, let $\boldsymbol{w}(\boldsymbol{\beta}) = \nabla\mathscr{R}_n(\boldsymbol{\beta}) - \nabla\mathscr{R}(\boldsymbol{\beta})$, where $\mathscr{R}(\boldsymbol{\beta}) = \mathbb{E}\{\mathscr{R}_n(\boldsymbol{\beta})\}$ is the population loss. Moreover, we define the quantity $\boldsymbol{w}^* = \nabla\mathscr{R}_n(\boldsymbol{\beta}^*) - \nabla\mathscr{R}(\boldsymbol{\beta}^*)$ as the centered score function. Recall that $\mathbb{C}(L) = \{\boldsymbol{\delta} : \|\boldsymbol{\delta}\|_1 \le L\|\boldsymbol{\delta}\|_2\}$, and let $\mathbb{C}_1 := \{\boldsymbol{\delta} : \|\boldsymbol{\delta}_{\mathscr{S}^{\mathrm{c}}}\|_1 \le 3\|\boldsymbol{\delta}_{\mathscr{S}}\|_1\}$. Furthermore, define the symmetrized Bregman divergence $\mathscr{B} : \mathbb{R}^p \times \mathbb{R}^p \to [0,\infty)$ associated with the convex function $\mathscr{R}_n(\cdot)$ evaluated at $\boldsymbol{\beta}_1, \boldsymbol{\beta}_2$ as

$$\mathscr{B}(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2) = \langle \nabla\mathscr{R}_n(\boldsymbol{\beta}_1) - \nabla\mathscr{R}_n(\boldsymbol{\beta}_2), \boldsymbol{\beta}_1 - \boldsymbol{\beta}_2 \rangle. \tag{A.1}$$

Lastly, it can be checked that we have $\lambda_u \ge \lambda_{\max}(\boldsymbol{\Sigma})$ where $\boldsymbol{\Sigma} = \mathbb{E}(\boldsymbol{x}\boldsymbol{x}^{\mathrm{T}})$ from Condition 2, and

$\mathbb{E}(\varepsilon^2|\boldsymbol{x}) \leq \sigma_\varepsilon^2$ from Condition 3.

We first present two technical lemmas that are useful for proving theoretical results in the low-dimensional setting for the non-penalized `retire` estimator in Section 1.4, i.e.,

$$\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}_\gamma = \underset{\boldsymbol{\beta}\in\mathbb{R}^d}{\operatorname{argmin}} \frac{1}{n}\sum_{i=1}^{n} L_{\tau,\gamma}(y_i - \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}). \tag{A.2}$$

Lemma A.1.1. *Under Conditions 1, 2, and 3, we have* $\|\boldsymbol{\Sigma}^{-1/2}\nabla\mathscr{R}(\boldsymbol{\beta}^*)\|_2 \leq \gamma^{-1}\bar{\tau}\sigma_\varepsilon^2$. *Moreover, for any* $t > 0$,

$$\left|\left|\boldsymbol{\Sigma}^{-1/2}\left\{\nabla\mathscr{R}_n(\boldsymbol{\beta}^*) - \nabla\mathscr{R}(\boldsymbol{\beta}^*)\right\}\right|\right|_2 \leq 3\bar{\tau}v_0\left(\sigma_\varepsilon\sqrt{\frac{2d+t}{n}} + \gamma\frac{2d+t}{2n}\right)$$

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

Lemma A.1.2. *Let* ε *be a real-valued random variable with* $\mathbb{E}(\varepsilon) = \sigma_\varepsilon^2$, $\mathbb{E}|\varepsilon|^3 = v_3 < \infty$, *and* $\mathbb{E}\{w_\tau(\varepsilon)\varepsilon\} = 0$ *with* $w_\tau(u) = |\tau - \mathbb{1}(u < 0)|$. *Let* $\ell_\gamma(\cdot)$ *be the Huber loss with parameter* γ. *We have*

$$\left|\mathbb{E}\{w_\tau(\varepsilon)\ell'_\gamma(\varepsilon)\}\right| \leq \bar{\tau}v_3/\gamma^2 \;\text{ and }\; \underline{\tau}^2\left(\sigma_\varepsilon^2 - v_3/\gamma\right) \leq \mathbb{E}\left\{w_\tau(\varepsilon)\ell'_\gamma(\varepsilon)\right\}^2 \leq \bar{\tau}^2\sigma_\varepsilon^2.$$

The proofs of all of the technical lemmas are deferred to Appendix A.5.

A.2 Proof of Theorems

A.2.1 Proof of Theorem 1.6.1

Proof. Recall from (A.2) that $\hat{\boldsymbol{\beta}} = \operatorname{argmin}\ \mathscr{R}_n(\boldsymbol{\beta})$ and from (A.1) that $\mathscr{B}(\boldsymbol{\beta}, \boldsymbol{\beta}^*) = \langle\nabla\mathscr{R}_n(\boldsymbol{\beta}) - \nabla\mathscr{R}_n(\boldsymbol{\beta}^*), \boldsymbol{\beta} - \boldsymbol{\beta}^*\rangle$ is the symmetric Bregman divergence. The main idea is to establish lower and upper bounds for $\mathscr{B}(\hat{\boldsymbol{\beta}}, \boldsymbol{\beta}^*)$.

We start with obtaining a lower bound for $\mathscr{B}(\hat{\boldsymbol{\beta}}, \boldsymbol{\beta}^*)$. Let $r_{\mathrm{loc}} = \gamma/(8\sqrt{2}A_1^2)$ and define an intermediate quantity $\hat{\boldsymbol{\beta}}_\eta = \eta\hat{\boldsymbol{\beta}} + (1-\eta)\boldsymbol{\beta}^*$, where $\eta = \sup\{\eta \in [0,1] : \hat{\boldsymbol{\beta}}_\eta \in \boldsymbol{\beta}^* + \mathbb{B}_{\boldsymbol{\Sigma}}(r_{\mathrm{loc}})\}$.

Then $\hat{\boldsymbol{\beta}}_\eta \in \boldsymbol{\beta}^* + \partial\mathbb{B}_{\boldsymbol{\Sigma}}(r_{\mathrm{loc}})$ whenever $\hat{\boldsymbol{\beta}} \notin \boldsymbol{\beta}^* + \mathbb{B}_{\boldsymbol{\Sigma}}(r_{\mathrm{loc}})$, where $\partial\mathbb{B}_{\boldsymbol{\Sigma}}(r_{\mathrm{loc}})$ is the boundary of $\mathbb{B}_{\boldsymbol{\Sigma}}(r_{\mathrm{loc}})$. On the other hand, $\hat{\boldsymbol{\beta}}_\eta = \hat{\boldsymbol{\beta}}$ whenever $\hat{\boldsymbol{\beta}} \in \boldsymbol{\beta}^* + \mathbb{B}_{\boldsymbol{\Sigma}}(r_{\mathrm{loc}})$. By an application of Lemma 1.6.1, provided that $\gamma \geq 4\sqrt{2}\sigma_\varepsilon$ and $n \gtrsim d+t$, we obtain

$$\mathscr{B}(\hat{\boldsymbol{\beta}}_\eta, \boldsymbol{\beta}^*) \geq \frac{1}{2}\kappa_1 \underline{\tau} \|\hat{\boldsymbol{\beta}}_\eta - \boldsymbol{\beta}^*\|_{\boldsymbol{\Sigma}}^2, \tag{A.3}$$

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

Next, we proceed to obtain an upper bound of $\mathscr{B}(\hat{\boldsymbol{\beta}}, \boldsymbol{\beta}^*)$. By an application of Lemma C.1 in Sun, Zhou and Fan (2020) and the first order condition $\nabla\mathscr{R}_n(\hat{\boldsymbol{\beta}}) = \mathbf{0}$, we have

$$\begin{aligned}
\mathscr{B}(\hat{\boldsymbol{\beta}}_\eta, \boldsymbol{\beta}^*) &\leq \eta\, \mathscr{B}(\hat{\boldsymbol{\beta}}, \boldsymbol{\beta}^*) && \text{(A.4)} \\
&= \eta \langle -\nabla\mathscr{R}_n(\boldsymbol{\beta}^*), \hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^* \rangle \\
&\leq ||\boldsymbol{\Sigma}^{-1/2}\nabla\mathscr{R}_n(\boldsymbol{\beta}^*)||_2 \cdot \|\hat{\boldsymbol{\beta}}_\eta - \boldsymbol{\beta}^*\|_{\boldsymbol{\Sigma}} \\
&\leq \left[\left|\left|\boldsymbol{\Sigma}^{-1/2}\nabla\mathscr{R}(\boldsymbol{\beta}^*)\right|\right|_2 + \left|\left|\boldsymbol{\Sigma}^{-1/2}\left\{\nabla\mathscr{R}_n(\boldsymbol{\beta}^*) - \nabla\mathscr{R}(\boldsymbol{\beta}^*)\right\}\right|\right|_2 \right] \cdot \|\hat{\boldsymbol{\beta}}_\eta - \boldsymbol{\beta}^*\|_{\boldsymbol{\Sigma}}. && \text{(A.5)}
\end{aligned}$$

Combining the above upper and lower bounds in (A.3) and (A.4), applying Lemma A.1.1, and picking $\gamma = \sigma_\varepsilon\sqrt{n/(d+t)}$, we have

$$\|\hat{\boldsymbol{\beta}}_\eta - \boldsymbol{\beta}^*\|_{\boldsymbol{\Sigma}} \leq C(\bar{\tau}/\underline{\tau})\kappa_1^{-1}\sigma_\varepsilon v_0 \sqrt{\frac{d+t}{n}}, \tag{A.6}$$

with probability at least $1 - 2e^{-t}$ as long as $n \gtrsim d+t$, where C is an absolute constant.

Lastly, it can be checked that with our proper choice of γ and r_{loc}, we have $\|\hat{\boldsymbol{\beta}}_\eta - \boldsymbol{\beta}^*\|_{\boldsymbol{\Sigma}} \lesssim \sigma_\varepsilon\sqrt{(d+t)/n} < \sigma_\varepsilon\sqrt{n/(d+t)} \asymp r_{\mathrm{loc}}$. It immediately implies $\hat{\boldsymbol{\beta}}_\eta \in \boldsymbol{\beta}^* + \mathbb{B}_{\boldsymbol{\Sigma}}(r_{\mathrm{loc}})$ and $\hat{\boldsymbol{\beta}}_\eta = \hat{\boldsymbol{\beta}}$ by construction. Thus (A.6) also holds when replacing $\hat{\boldsymbol{\beta}}_\eta$ by $\hat{\boldsymbol{\beta}}$.

□

A.2.2 Proof of Theorem 1.6.2

Proof. We consider the following vector-valued random process

$$\boldsymbol{B}(\boldsymbol{\beta}) = \boldsymbol{\Sigma}^{-1/2}\{\nabla\mathscr{R}_n(\boldsymbol{\beta}) - \nabla\mathscr{R}_n(\boldsymbol{\beta}^*)\} - \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{\Sigma}^{-1/2}\mathbb{E}w_\tau(\varepsilon_i)\boldsymbol{x}_i\boldsymbol{x}_i^{\mathrm{T}}(\boldsymbol{\beta}-\boldsymbol{\beta}^*).$$

By the first order condition $\nabla\mathscr{R}_n(\hat{\boldsymbol{\beta}}) = \mathbf{0}$, it can be shown that the nonasymptotic Bahadur representation in (1.8) takes the form $\|\boldsymbol{B}(\hat{\boldsymbol{\beta}})\|_2$. By the triangle inequality, we have

$$\|\boldsymbol{B}(\hat{\boldsymbol{\beta}})\|_2 \le \sup_{\boldsymbol{\beta}\in\boldsymbol{\beta}^*+\mathbb{B}_{\boldsymbol{\Sigma}}(r)}\|\mathbb{E}\{\boldsymbol{B}(\boldsymbol{\beta})\}\|_2 + \sup_{\boldsymbol{\beta}\in\boldsymbol{\beta}^*+\mathbb{B}_{\boldsymbol{\Sigma}}(r)}\|\boldsymbol{B}(\boldsymbol{\beta}) - \mathbb{E}\{\boldsymbol{B}(\boldsymbol{\beta})\}\|_2$$

for radius r that satisfies $\hat{\boldsymbol{\beta}} \in \boldsymbol{\beta}^* + \mathbb{B}_{\boldsymbol{\Sigma}}(r)$ with high probability. It suffices to obtain upper bounds for the two terms separately.

We start with an upper bound on $\sup_{\boldsymbol{\beta}\in\boldsymbol{\beta}^*+\mathbb{B}_{\boldsymbol{\Sigma}}(r)}\|\mathbb{E}\{\boldsymbol{B}(\boldsymbol{\beta})\}\|_2$. By the mean value theorem for vector-valued functions (Theorem 12 in Section 2 of Pugh (2015)), we obtain

$$\begin{aligned}
\mathbb{E}\{\boldsymbol{B}(\boldsymbol{\beta})\} &= \boldsymbol{\Sigma}^{-1/2}\mathbb{E}\int_0^1 \nabla^2\mathscr{R}_n(\boldsymbol{\beta}_t^*)dt(\boldsymbol{\beta}-\boldsymbol{\beta}^*) - \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{\Sigma}^{-1/2}\mathbb{E}w_\tau(\varepsilon_i)\boldsymbol{x}_i\boldsymbol{x}_i^{\mathrm{T}}(\boldsymbol{\beta}-\boldsymbol{\beta}^*)\\
&= \Big\langle\int_0^1\Big\{\boldsymbol{\Sigma}^{-1/2}\mathbb{E}\nabla^2\mathscr{R}_n(\boldsymbol{\beta}_t^*)\boldsymbol{\Sigma}^{-1/2} - \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}w_\tau(\varepsilon_i)\boldsymbol{z}_i\boldsymbol{z}_i^{\mathrm{T}}\Big\}dt, \boldsymbol{\Sigma}^{1/2}(\boldsymbol{\beta}-\boldsymbol{\beta}^*)\Big\rangle,
\end{aligned}$$

where $\boldsymbol{\beta}_t^* = (1-t)\boldsymbol{\beta}^* + t\boldsymbol{\beta}$ for $(0 \le t \le 1)$ and $\boldsymbol{z}_i = \boldsymbol{\Sigma}^{-1/2}\boldsymbol{x}_i$. Let $\boldsymbol{\delta}_t = \boldsymbol{\Sigma}^{1/2}(\boldsymbol{\beta}_t^* - \boldsymbol{\beta}^*)$. Since $\boldsymbol{\beta} \in \boldsymbol{\beta}^* + \mathbb{B}_{\boldsymbol{\Sigma}}(r)$, we have $||\boldsymbol{\delta}_t||_2 \le r$ and $y_i - \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}_t^* = \varepsilon_i - \boldsymbol{\delta}_t^{\mathrm{T}}\boldsymbol{z}_i$. For all $\boldsymbol{u} \in \mathbb{S}^{d-1}$, we obtain

$$\begin{aligned}
&\Big|\boldsymbol{u}^{\mathrm{T}}\Big\{\boldsymbol{\Sigma}^{-1/2}\mathbb{E}\nabla^2\mathscr{R}_n(\boldsymbol{\beta}_t^*)\boldsymbol{\Sigma}^{-1/2} - \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}w_\tau(\varepsilon_i)\boldsymbol{z}_i\boldsymbol{z}_i^{\mathrm{T}}\Big\}\boldsymbol{u}\Big|\\
&= \Big|\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{u}^{\mathrm{T}}\Big[\mathbb{E}w_\tau(\varepsilon_i - \boldsymbol{\delta}_t^{\mathrm{T}}\boldsymbol{z}_i)\Big\{1 - \mathbb{1}(|\varepsilon_i - \boldsymbol{\delta}_t^{\mathrm{T}}\boldsymbol{z}_i| > \gamma)\Big\}\boldsymbol{z}_i\boldsymbol{z}_i^{\mathrm{T}} - \mathbb{E}w_\tau(\varepsilon_i)\boldsymbol{z}_i\boldsymbol{z}_i^{\mathrm{T}}\Big]\boldsymbol{u}\Big|\\
&\le \Big|\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\Big[(\boldsymbol{u}^{\mathrm{T}}\boldsymbol{z}_i)^2\mathbb{E}\Big\{w_\tau(\varepsilon_i - \boldsymbol{\delta}_t^{\mathrm{T}}\boldsymbol{z}_i) - w_\tau(\varepsilon_i)|\boldsymbol{z}_i\Big\}\Big]\Big| + \Big|\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}(\boldsymbol{u}^{\mathrm{T}}\boldsymbol{z}_i)^2 w_\tau(\varepsilon_i - \boldsymbol{\delta}_t^{\mathrm{T}}\boldsymbol{z}_i)\mathbb{1}(|\varepsilon_i - \boldsymbol{\delta}_t^{\mathrm{T}}\boldsymbol{z}_i| > \gamma)\Big|\\
&:= \Pi_1 + \Pi_2.
\end{aligned}$$

For Π_1, let $f_{\varepsilon|\boldsymbol{x}}$ be the conditional density function of ε given $\boldsymbol{x}$, and recall that it is upper bounded by $\bar{f}_{\varepsilon|\boldsymbol{x}}$. Moreover, let $m_3 > 0$ be a constant that satisfies $\sup_{\boldsymbol{u}\in\mathbb{S}^{d-1}} \mathbb{E}|\boldsymbol{u}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1/2}\boldsymbol{x}|^3 \leq m_3$. We have $\mathbb{E}\big\{w_\tau(\varepsilon_i - \boldsymbol{\delta}_t^{\mathrm{T}}\boldsymbol{z}_i) - w_\tau(\varepsilon_i)|\boldsymbol{z}_i\big\} = \int_{-\infty}^{\infty}\big\{w_\tau(u-\boldsymbol{\delta}_t^{\mathrm{T}}\boldsymbol{z}_i) - w_\tau(u)\big\}f_{\varepsilon|\boldsymbol{x}}(u)du \leq (\bar{\tau} - \underline{\tau})\bar{f}_{\varepsilon|\boldsymbol{x}}|\boldsymbol{\delta}_t^{\mathrm{T}}\boldsymbol{z}_i|$. Consequently,

$$\Pi_1 \leq \frac{1}{n}\sum_{i=1}^{n}(\bar{\tau}-\underline{\tau})\bar{f}_{\varepsilon|\boldsymbol{x}}\mathbb{E}\Big\{(\boldsymbol{u}^{\mathrm{T}}\boldsymbol{z}_i)^2|\boldsymbol{\delta}_t^{\mathrm{T}}\boldsymbol{z}_i|\Big\} \leq (\bar{\tau}-\underline{\tau})\bar{f}_{\varepsilon|\boldsymbol{x}}m_3 rt. \tag{A.7}$$

For Π_2, we first note that $\mathbb{1}(|\varepsilon_i - \boldsymbol{\delta}_t^{\mathrm{T}}\boldsymbol{z}_i| > \gamma) \leq \mathbb{1}(|\varepsilon_i| > \gamma/2) + \mathbb{1}(|\boldsymbol{\delta}_t^{\mathrm{T}}\boldsymbol{z}_i| > \gamma/2)$. By an application of the Markov's inequality, we obtain

$$\begin{aligned}
\Pi_2 &\leq \bigg|\bar{\tau}\mathbb{E}(\boldsymbol{u}^{\mathrm{T}}\boldsymbol{z})^2\Big\{\mathbb{1}(|\varepsilon| > \gamma/2) + \mathbb{1}(|\boldsymbol{\delta}_t^{\mathrm{T}}\boldsymbol{z}| > \gamma/2)\Big\}\bigg| \\
&\leq \bar{\tau}\bigg|\mathbb{E}\bigg(\frac{|\varepsilon|}{\gamma/2}\bigg)^2(\boldsymbol{u}^{\mathrm{T}}\boldsymbol{z})^2\bigg| + \bar{\tau}\bigg|\mathbb{E}\frac{|\boldsymbol{\delta}_t^{\mathrm{T}}\boldsymbol{z}|}{\gamma/2}(\boldsymbol{u}^{\mathrm{T}}\boldsymbol{z})^2\bigg| \\
&\leq \frac{4\bar{\tau}\sigma_\varepsilon^2}{\gamma^2} + \frac{2\bar{\tau}m_3 r}{\gamma}.
\end{aligned} \tag{A.8}$$

Combining (A.7) and (A.8), we have

$$\sup_{\boldsymbol{\beta}\in\boldsymbol{\beta}^*+\mathbb{B}_{\boldsymbol{\Sigma}}(r)} \|\mathbb{E}\{\boldsymbol{B}(\boldsymbol{\beta})\}\|_2 \leq \delta(r)r := \bigg\{(\bar{\tau}-\underline{\tau})\bar{f}_{\varepsilon|\boldsymbol{x}}m_3 rt + \frac{4\bar{\tau}\sigma_\varepsilon^2}{\gamma^2} + \frac{2\bar{\tau}m_3 r}{\gamma}\bigg\}r. \tag{A.9}$$

Next, we obtain an upper bound for $\sup_{\boldsymbol{\beta}\in\boldsymbol{\beta}^*+\mathbb{B}_{\boldsymbol{\Sigma}}(r)} \|\boldsymbol{B}(\boldsymbol{\beta}) - \mathbb{E}\{\boldsymbol{B}(\boldsymbol{\beta})\}\|_2$. With some abuse of notation, let $\bar{\boldsymbol{B}}(\boldsymbol{\delta}) = \boldsymbol{B}(\boldsymbol{\beta}) - \mathbb{E}\{\boldsymbol{B}(\boldsymbol{\beta})\}$ where $\boldsymbol{\delta} = \boldsymbol{\Sigma}^{1/2}(\boldsymbol{\beta}-\boldsymbol{\beta}^*) \in \mathbb{B}(r)$. It can be checked that $\bar{\boldsymbol{B}}(\mathbf{0}) = \mathbf{0}$, $\mathbb{E}\{\bar{\boldsymbol{B}}(\boldsymbol{\delta})\} = \mathbf{0}$, and

$$\begin{aligned}
\nabla_{\boldsymbol{\delta}}\bar{\boldsymbol{B}}(\boldsymbol{\delta}) &= \frac{1}{n}\sum_{i=1}^{n}\Big[w_\tau(\varepsilon_i - \boldsymbol{\delta}^{\mathrm{T}}\boldsymbol{z}_i)\ell_\gamma''(\varepsilon_i - \boldsymbol{\delta}^{\mathrm{T}}\boldsymbol{z}_i)\boldsymbol{z}_i\boldsymbol{z}_i^{\mathrm{T}} - \mathbb{E}\Big\{w_\tau(\varepsilon_i - \boldsymbol{\delta}^{\mathrm{T}}\boldsymbol{z}_i)\ell_\gamma''(\varepsilon_i - \boldsymbol{\delta}^{\mathrm{T}}\boldsymbol{z}_i)\boldsymbol{z}_i\boldsymbol{z}_i^{\mathrm{T}}\Big\}\Big] \\
&:= \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{A}_i.
\end{aligned}$$

For all $\boldsymbol{u},\boldsymbol{v}\in\mathbb{S}^{d-1}$ and $\lambda\in\mathbb{R}$, with careful calculation, we see that $\mathbb{E}\boldsymbol{u}^{\mathrm{T}}\boldsymbol{A}_i\boldsymbol{v}=0$, $|\boldsymbol{u}^{\mathrm{T}}\boldsymbol{A}_i\boldsymbol{v}|\le \bar{\tau}|\boldsymbol{u}^{\mathrm{T}}\boldsymbol{z}_i\boldsymbol{v}^{\mathrm{T}}\boldsymbol{z}_i|+\bar{\tau}\mathbb{E}|\boldsymbol{u}^{\mathrm{T}}\boldsymbol{z}_i\boldsymbol{v}^{\mathrm{T}}\boldsymbol{z}_i|$ and $|\boldsymbol{u}^{\mathrm{T}}\boldsymbol{A}_i\boldsymbol{v}|^2\le 2\bar{\tau}^2\big(|\boldsymbol{u}^{\mathrm{T}}\boldsymbol{z}_i\boldsymbol{v}^{\mathrm{T}}\boldsymbol{z}_i|^2+\mathbb{E}^2|\boldsymbol{u}^{\mathrm{T}}\boldsymbol{z}_i\boldsymbol{v}^{\mathrm{T}}\boldsymbol{z}_i|\big)$. It then follows from the elementary inequality $|e^z-1-z|\le z^2e^{|z|}/2$ and bound

$$\mathbb{E}|\boldsymbol{u}^{\mathrm{T}}\boldsymbol{z}_i\boldsymbol{v}^{\mathrm{T}}\boldsymbol{z}_i|\le\big\{\mathbb{E}(\boldsymbol{u}^{\mathrm{T}}\boldsymbol{z}_i)^2\big\}^{1/2}\big\{\mathbb{E}(\boldsymbol{v}^{\mathrm{T}}\boldsymbol{z}_i)^2\big\}^{1/2}\le 1$$

that

$$\begin{aligned}\mathbb{E}\exp\Big\{\lambda\sqrt{n}\boldsymbol{u}^{\mathrm{T}}\nabla_{\boldsymbol{\delta}}\bar{\boldsymbol{B}}(\boldsymbol{\delta})\boldsymbol{v}\Big\}&=\prod_{i=1}^n\mathbb{E}\exp\left\{\frac{\lambda}{\sqrt{n}}\boldsymbol{u}^{\mathrm{T}}\boldsymbol{A}_i\boldsymbol{v}\right\}\\&\le\prod_{i=1}^n\mathbb{E}\left\{1+\frac{\lambda}{\sqrt{n}}\boldsymbol{u}^{\mathrm{T}}\boldsymbol{A}_i\boldsymbol{v}+\left(\frac{\lambda}{\sqrt{n}}\boldsymbol{u}^{\mathrm{T}}\boldsymbol{A}_i\boldsymbol{v}\right)^2e^{\left|\frac{\lambda}{\sqrt{n}}\boldsymbol{u}^{\mathrm{T}}\boldsymbol{A}_i\boldsymbol{v}\right|}/2\right\}\\&\le\prod_{i=1}^n\mathbb{E}\left\{1+\frac{\lambda^2\bar{\tau}^2}{n}e^{\frac{|\lambda|\bar{\tau}}{\sqrt{n}}}\Big(e^{\frac{|\lambda|\bar{\tau}}{\sqrt{n}}|\boldsymbol{u}^{\mathrm{T}}\boldsymbol{z}_i\boldsymbol{v}^{\mathrm{T}}\boldsymbol{z}_i|}+|\boldsymbol{u}^{\mathrm{T}}\boldsymbol{z}_i\boldsymbol{v}^{\mathrm{T}}\boldsymbol{z}_i|^2e^{\frac{|\lambda|\bar{\tau}}{\sqrt{n}}|\boldsymbol{u}^{\mathrm{T}}\boldsymbol{z}_i\boldsymbol{v}^{\mathrm{T}}\boldsymbol{z}_i|}\Big)\right\}.\end{aligned}\tag{A.10}$$

Here we upper-bound the components appeared in the right-hand side of (A.10). For all $t>0$, it follows from Cauchy-Schwarz inequality and the elementary inequality $ab\le a^2/2+b^2/2$ that

$$\begin{aligned}\mathbb{E}|\boldsymbol{u}^{\mathrm{T}}\boldsymbol{z}_i\boldsymbol{v}^{\mathrm{T}}\boldsymbol{z}_i|^2e^{t|\boldsymbol{u}^{\mathrm{T}}\boldsymbol{z}_i\boldsymbol{v}^{\mathrm{T}}\boldsymbol{z}_i|}&\le\mathbb{E}(\boldsymbol{u}^{\mathrm{T}}\boldsymbol{z}_i)^2(\boldsymbol{v}^{\mathrm{T}}\boldsymbol{z}_i)^2e^{t(\boldsymbol{u}^{\mathrm{T}}\boldsymbol{z}_i)^2/2+t(\boldsymbol{v}^{\mathrm{T}}\boldsymbol{z}_i)^2/2}\\&\le\Big\{\mathbb{E}(\boldsymbol{u}^{\mathrm{T}}\boldsymbol{z}_i)^4e^{t(\boldsymbol{u}^{\mathrm{T}}\boldsymbol{z}_i)^2}\Big\}^{1/2}\Big\{\mathbb{E}(\boldsymbol{v}^{\mathrm{T}}\boldsymbol{z}_i)^4e^{t(\boldsymbol{v}^{\mathrm{T}}\boldsymbol{z}_i)^2}\Big\}^{1/2}.\end{aligned}$$

Consequently $\mathbb{E}|\boldsymbol{u}^{\mathrm{T}}\boldsymbol{z}_i\boldsymbol{v}^{\mathrm{T}}\boldsymbol{z}_i|^2e^{t|\boldsymbol{u}^{\mathrm{T}}\boldsymbol{z}_i\boldsymbol{v}^{\mathrm{T}}\boldsymbol{z}_i|}\le\sup_{\boldsymbol{u}\in\mathbb{S}^{d-1}}\mathbb{E}(\boldsymbol{u}^{\mathrm{T}}\boldsymbol{z}_i)^4e^{t(\boldsymbol{u}^{\mathrm{T}}\boldsymbol{z}_i)^2}$, and similarly $\mathbb{E}e^{t|\boldsymbol{u}^{\mathrm{T}}\boldsymbol{z}_i\boldsymbol{v}^{\mathrm{T}}\boldsymbol{z}_i|}\le\sup_{\boldsymbol{u}\in\mathbb{S}^{d-1}}\mathbb{E}e^{t(\boldsymbol{u}^{\mathrm{T}}\boldsymbol{z}_i)^2}$. To further upper-bound these supremums, let $\chi:=(\boldsymbol{u}^{\mathrm{T}}\boldsymbol{z})^2/(2v_1)^2$. Recall the sub-Gaussian condition $\mathbb{P}(|\langle\boldsymbol{u},\boldsymbol{\Sigma}^{-1/2}\boldsymbol{x}\rangle|\ge v_1||\boldsymbol{u}||_2t)\le 2e^{-t^2/2}$, we have $\mathbb{P}(\chi\ge t)\le 2e^{-2t}$ (i.e., χ is sub-Exponential). It follows that $\mathbb{E}e^{\chi}=1+\int_0^\infty e^t\mathbb{P}(\chi\ge t)dt\le 1+2\int_0^\infty e^{-t}dt=3$, and

$$\mathbb{E}(\chi^2e^{\chi})=\int_0^\infty(t^2+2t)e^t\mathbb{P}(\chi\ge t)dt\le 2\int_0^\infty(t^2+2t)e^{-t}dt=8.$$

Along with the monotonicity of exponential function, we conclude both $\mathbb{E}|\boldsymbol{u}^{\mathrm{T}}\boldsymbol{z}_i\boldsymbol{v}^{\mathrm{T}}\boldsymbol{z}_i|^2 e^{\frac{|\lambda|\bar{\tau}}{\sqrt{n}}|\boldsymbol{u}^{\mathrm{T}}\boldsymbol{z}_i\boldsymbol{v}^{\mathrm{T}}\boldsymbol{z}_i|}$ and $\mathbb{E}e^{\frac{|\lambda|\bar{\tau}}{\sqrt{n}}|\boldsymbol{u}^{\mathrm{T}}\boldsymbol{z}_i\boldsymbol{v}^{\mathrm{T}}\boldsymbol{z}_i|}$ can be upper-bounded by some constants C_1, C_2 respectively, uniformly over $\boldsymbol{u},\boldsymbol{v}\in\mathbb{S}^{d-1}$, as long as $|\lambda|\le\sqrt{n}/(4v_1^2\bar{\tau})$. Substituting the above bounds into (A.10) yields,

$$\begin{aligned}\mathbb{E}\exp\left\{\lambda\sqrt{n}\boldsymbol{u}^{\mathrm{T}}\nabla_{\boldsymbol{\delta}}\bar{\boldsymbol{B}}(\boldsymbol{\delta})\boldsymbol{v}\right\} &\le \prod_{i=1}^{n}\left[1+\frac{\lambda^2\bar{\tau}^2}{n}e^{\frac{|\lambda|\bar{\tau}}{\sqrt{n}}}\left\{\sup_{\boldsymbol{u}\in\mathbb{S}^{d-1}}\mathbb{E}e^{\frac{|\lambda|\bar{\tau}}{\sqrt{n}}(\boldsymbol{u}^{\mathrm{T}}\boldsymbol{z}_i)^2}+\sup_{\boldsymbol{u}\in\mathbb{S}^{d-1}}\mathbb{E}(\boldsymbol{u}^{\mathrm{T}}\boldsymbol{z}_i)^4 e^{\frac{|\lambda|\bar{\tau}}{\sqrt{n}}(\boldsymbol{u}^{\mathrm{T}}\boldsymbol{z}_i)^2}\right\}\right]\\ &\le \exp\left\{\lambda^2\bar{\tau}^2 e^{\frac{|\lambda|\bar{\tau}}{\sqrt{n}}}(C_1+C_2)\right\}\\ &\le \exp\left\{2(C_1+C_2)\bar{\tau}^2e^{-4v_1^2}\cdot\frac{\lambda^2}{2}\right\}\quad\text{valid for all}\quad \lambda^2\le 2\cdot\frac{n}{32\bar{\tau}^2v_1^4}.\end{aligned}$$

With the above preparations, we apply Theorem A.3 in Spokoiny (2013) with $v_0^2 = 2(C_1+C_2)\bar{\tau}^2e^{-4v_1^2}$ and $g^2 = n/(32\bar{\tau}^2v_1^4)$ to yield

$$\sup_{\boldsymbol{\beta}\in\boldsymbol{\beta}^*+\mathbb{B}_{\boldsymbol{\Sigma}}(r)}||\boldsymbol{B}(\boldsymbol{\beta})-\mathbb{E}\boldsymbol{B}(\boldsymbol{\beta})||_2 \le 12\sqrt{C_1+C_2}\bar{\tau}e^{-2v_1^2}\sqrt{\frac{2d+t}{n}}\cdot r \tag{A.11}$$

with probability at least $1-e^{-t}$, as long as $n\ge 64\bar{\tau}^2v_1^4(2d+t)$.

Lastly, combining (A.9) and (A.11), we have

$$\sup_{\boldsymbol{\beta}\in\boldsymbol{\beta}^*+\mathbb{B}_{\boldsymbol{\Sigma}}(r)}||\boldsymbol{B}(\boldsymbol{\beta})||_2 \le \left\{\delta(r)+12\sqrt{C_1+C_2}\bar{\tau}e^{-2v_1^2}\sqrt{\frac{2d+t}{n}}\right\}r \tag{A.12}$$

with probability at least $1-e^{-t}$, as long as $n\ge 64\bar{\tau}^2v_1^4(2d+t)$. Recall from the proof of Theorem 1.6.1 that we have $\hat{\boldsymbol{\beta}}\in\boldsymbol{\beta}^*+\mathbb{B}_{\boldsymbol{\Sigma}}(r_0)$ with probability at least $1-2e^{-t}$ for some $r_0\asymp\sigma_\varepsilon\sqrt{(d+t)/n}$. Taking $r=r_0$ in (A.12) and $\gamma=\sigma_\varepsilon\sqrt{n/(d+t)}$ finishes the proof.

□

A.2.3 Proof of Theorem 1.6.3

Proof. Let $\boldsymbol{u}\in\mathbb{R}^d$ be an arbitrary vector, and $\mathbf{J}=\mathbb{E}\{w_\tau(\varepsilon)\boldsymbol{x}\boldsymbol{x}^{\mathrm{T}}\}$ be the Hessian matrix. Define $S_n = n^{-1/2}\sum_{i=1}^n a_ib_i$ and its centered version $S_n^0 = S_n-\mathbb{E}(S_n)$, where $a_i = w_\tau(\varepsilon_i)\ell'_\gamma(\varepsilon_i)$ and $b_i=\langle\mathbf{J}^{-1}\boldsymbol{u},\boldsymbol{x}_i\rangle$. We first show that the centered partial sum S_n^0 is close to the quantity of interest

$n^{1/2}\langle \boldsymbol{u}, \hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^* \rangle$. By an application of Theorem 1.6.2 and Lemma A.1.2 with $\gamma = \sigma_\varepsilon \sqrt{n/(d+t)}$ and $t = \log n$, we obtain

$$
\begin{aligned}
&|n^{1/2}\langle \boldsymbol{u}, \hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^* \rangle - S_n^0| \\
&\leq n^{1/2} \Big| \Big\langle \boldsymbol{\Sigma}^{1/2} \mathbf{J}^{-1} \boldsymbol{u}, \boldsymbol{\Sigma}^{-1/2} \mathbf{J} (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*) - \frac{1}{n} \sum_{i=1}^n w_\tau(\varepsilon_i) \ell_\gamma'(\varepsilon_i) \boldsymbol{\Sigma}^{-1/2} \boldsymbol{x}_i \Big\rangle \Big| + |\mathbb{E} S_n| \\
&\leq n^{1/2} \big| |\mathbf{J}^{-1} \boldsymbol{u}| \big|_{\boldsymbol{\Sigma}} \cdot \big| \big| \boldsymbol{\Sigma}^{-1/2} \mathbf{J} (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*) - \frac{1}{n} \sum_{i=1}^n w_\tau(\varepsilon_i) \ell_\gamma'(\varepsilon_i) \boldsymbol{\Sigma}^{-1/2} \boldsymbol{x}_i \big| \big|_2 + n^{1/2} \Big| \mathbb{E} \Big[\langle \mathbf{J}^{-1} \boldsymbol{u}, \boldsymbol{x} \rangle \mathbb{E} \Big\{ w_\tau(\varepsilon) \ell_\gamma'(\varepsilon) | \boldsymbol{x} \Big\} \Big] \Big| \\
&\leq n^{1/2} \big| |\mathbf{J}^{-1} \boldsymbol{u}| \big|_{\boldsymbol{\Sigma}} C \cdot \frac{d + \log n}{n} + n^{1/2} \frac{\bar{\tau} v_3}{\gamma^2} \Big(\mathbb{E} \langle \mathbf{J}^{-1} \boldsymbol{u}, \boldsymbol{x} \rangle^2 \Big)^{1/2} \\
&\leq C_1 \big| |\mathbf{J}^{-1} \boldsymbol{u}| \big|_{\boldsymbol{\Sigma}} \frac{d + \log n}{\sqrt{n}}, \qquad \text{(A.13)}
\end{aligned}
$$

with probability at least $1 - 3n^{-1}$, where $C_1 = C + \bar{\tau} v_3 / \sigma_\varepsilon^2$.

Next, we show that the centered partial sum $S_n^0 = n^{-1/2} \sum_{i=1}^n (1 - \mathbb{E}) a_i b_i$ is approximately normally distributed. It follows from Berry-Esseen inequality (e.g., see Tyurin (2011)) that

$$
\sup_{x \in \mathbb{R}} \big| \mathbb{P}(S_n^0 \leq \mathrm{var}(S_n^0)^{1/2} x) - \Phi(x) \big| \leq \frac{\mathbb{E} |a_i b_i - \mathbb{E} a_i b_i|^3}{2 \mathrm{var}(S_n^0)^{3/2} \sqrt{n}}. \qquad \text{(A.14)}
$$

Thus, it suffices to obtain a lower bound for $\mathrm{var}(S_n^0)$ and an upper bound for $\mathbb{E}(|a_i b_i - \mathbb{E} a_i b_i|^3)$. By an application of Lemma A.1.2, we have $\mathbb{E}(a_i b_i) \leq \bar{\tau} v_3 ||\mathbf{J}^{-1} \boldsymbol{u}||_{\boldsymbol{\Sigma}} / \gamma^2$ and $\mathbb{E}(a_i b_i)^2 \geq \underline{\tau}^2 (\sigma_\varepsilon^2 - 2 v_3 / \gamma) \|\mathbf{J}^{-1} \boldsymbol{u}\|_{\boldsymbol{\Sigma}}^2$. Thus, $\mathrm{var}(S_n^0) = \mathbb{E}(a_i b_i)^2 - (\mathbb{E} a_i b_i)^2 \geq \|\mathbf{J}^{-1} \boldsymbol{u}\|_{\boldsymbol{\Sigma}}^2 (\underline{\tau}^2 \sigma_\varepsilon^2 - 2 \underline{\tau}^2 v_3 / \gamma - \bar{\tau}^2 v_3^2 / \gamma^4)$. For sufficiently large γ (i.e., $n \gtrsim d$), we obtain the lower bound $\mathrm{var}(S_n^0)^{3/2} \geq ||\mathbf{J}^{-1} \boldsymbol{u}||_{\boldsymbol{\Sigma}}^3 (\underline{\tau}^3 \sigma_\varepsilon^3 / 2)$. Next, we proceed to obtain an upper bound for the centered third moment $\mathbb{E}|a_i b_i - \mathbb{E}(a_i b_i)|^3$. Recall that $m_3 = \sup_{\boldsymbol{u} \in \mathbb{S}^{d-1}} \mathbb{E} |\langle \boldsymbol{u}, \boldsymbol{\Sigma}^{-1/2} \boldsymbol{x} \rangle|^3$, we have

$$
\mathbb{E} |a_i b_i|^3 \leq \mathbb{E} \big[|\langle \mathbf{J}^{-1} \boldsymbol{u}, \boldsymbol{x}_i \rangle|^3 \mathbb{E} \big\{ |w_\tau(\varepsilon_i) \ell_\gamma'(\varepsilon_i)|^3 | \boldsymbol{x}_i \big\} \big] \leq \bar{\tau}^3 v_3 m_3 ||\mathbf{J}^{-1} \boldsymbol{u}||_{\boldsymbol{\Sigma}}^3.
$$

Along with Minkowski's inequality $|a+b|^p \leq 2^{p-1} (|a|^p + |b|^p)$, we obtain $\mathbb{E}|a_i b_i - \mathbb{E} a_i b_i|^3 \leq 4 \bar{\tau}^3 v_3 m_3 (1 + v_3^2 / m_3 \gamma^6) ||\mathbf{J}^{-1} \boldsymbol{u}||_{\boldsymbol{\Sigma}}^3$. Therefore $\mathbb{E}|a_i b_i - \mathbb{E} a_i b_i|^3 \leq 8 \bar{\tau}^3 v_3 m_3 ||\mathbf{J}^{-1} \boldsymbol{u}||_{\boldsymbol{\Sigma}}^3$ provided that

$n \gtrsim d$. Substituting the above inequalities into (A.14), we have

$$\sup_{x\in\mathbb{R}} \left|\mathbb{P}(S_n^0 \le \mathrm{var}(S_n^0)^{1/2}x) - \Phi(x)\right| \le C_2 n^{-1/2}, \tag{A.15}$$

where $C_2 = 8v_3m_3\bar{\tau}^3/(\underline{\tau}\sigma_\varepsilon)^3$.

Let $\sigma^2 = \mathbb{E}(a_ib_i)^2 = \boldsymbol{u}^{\mathrm{T}}\mathbf{J}^{-1}\mathbb{E}\big[\big\{w_\tau(\varepsilon)\ell'_\gamma(\varepsilon)\big\}^2\boldsymbol{x}\boldsymbol{x}^{\mathrm{T}}\big]\mathbf{J}^{-1}\boldsymbol{u}$. An application of Lemma A.1.2 indicates that $\underline{\tau}^2(\sigma_\varepsilon^2 - 2v_3/\gamma)\|\mathbf{J}^{-1}\boldsymbol{u}\|_{\boldsymbol{\Sigma}}^2 \le \sigma^2 \le \bar{\tau}^2\sigma_\varepsilon^2||\mathbf{J}^{-1}\boldsymbol{u}||_{\boldsymbol{\Sigma}}^2$. Moreover, $|\mathrm{var}(S_n^0) - \sigma^2| = |\mathbb{E}a_ib_i|^2 \le \bar{\tau}^2v_3^2||\mathbf{J}^{-1}\boldsymbol{u}||_{\boldsymbol{\Sigma}}^2/\gamma^4$. Provided that $n \gtrsim d$, we obtain

$$\left|\frac{\mathrm{var}(S_n^0)}{\sigma^2} - 1\right| \le \left(1 - \frac{2v_3}{\sigma_\varepsilon^2\gamma}\right)^{-1}\cdot\frac{\bar{\tau}^2v_3^2}{\underline{\tau}^2\sigma_\varepsilon^2}\cdot\frac{1}{\gamma^4} \le \frac{2\bar{\tau}^2v_3^2}{\underline{\tau}^2\sigma_\varepsilon^2}\cdot\frac{1}{\gamma^4}.$$

An application of Lemma A.7 in the supplement of Spokoiny and Zhilova (2015) indicates that

$$\sup_{x\in\mathbb{R}} \left|\Phi(x/\mathrm{var}(S_n^0)^{1/2}) - \Phi(x/\sigma)\right| \le C_3\gamma^{-4}, \tag{A.16}$$

where $C_3 = (\bar{\tau}v_3/\underline{\tau}\sigma_\varepsilon)^2$.

Let $G \sim \mathscr{N}(0,1)$. Applying the inequalities in (A.13), (A.15), and (A.16), along with the fact that for all $a < b$ and $\sigma > 0$, $\Phi(b/\sigma) - \Phi(a/\sigma) \le (2\pi)^{-1/2}(b-a)/\sigma$, we obtain that for any $x \in \mathbb{R}$ and $\boldsymbol{u} \in \mathbb{R}^d$,

$$\begin{aligned}
\mathbb{P}(n^{1/2}\langle\boldsymbol{u}, \hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\rangle \le x) &\le \mathbb{P}\left(S_n^0 \le x + C_1\big|\big|\mathbf{J}^{-1}\boldsymbol{u}\big|\big|_{\boldsymbol{\Sigma}}\frac{d+\log n}{\sqrt{n}}\right) + \frac{3}{n} \\
&\le \mathbb{P}\left(\mathrm{var}(S_n^0)^{1/2}G \le x + C_1\big|\big|\mathbf{J}^{-1}\boldsymbol{u}\big|\big|_{\boldsymbol{\Sigma}}\frac{d+\log n}{\sqrt{n}}\right) + \frac{3}{n} + \frac{C_2}{\sqrt{n}} \\
&\le \mathbb{P}\left(\sigma G \le x + C_1\big|\big|\mathbf{J}^{-1}\boldsymbol{u}\big|\big|_{\boldsymbol{\Sigma}}\frac{d+\log n}{\sqrt{n}}\right) + \frac{3}{n} + \frac{C_2}{\sqrt{n}} + \frac{C_3}{\gamma^4} \\
&\le \mathbb{P}\left(\sigma G \le x\right) + \frac{C_1||\mathbf{J}^{-1}\boldsymbol{u}||_{\boldsymbol{\Sigma}}}{\sqrt{2\pi}\sigma}\frac{d+\log n}{\sqrt{n}} + \frac{3}{n} + \frac{C_2}{\sqrt{n}} + \frac{C_3}{\gamma^4} \\
&\lesssim \mathbb{P}\left(\sigma G \le x\right) + \frac{d+\log n}{\sqrt{n}} + \frac{1}{n} + \frac{1}{\sqrt{n}} + \frac{(d+\log n)^2}{n^2},
\end{aligned}$$

where the last inequality follows from $\sigma \asymp ||\mathbf{J}^{-1}\boldsymbol{u}||_{\boldsymbol{\Sigma}}$ and taking $\gamma = \sigma_\varepsilon\sqrt{n/(d+\log n)}$. A similar argument leads to a series of reverse inequalities. Since the above bounds are independent of x and $\boldsymbol{u}$, they hold uniformly over $x \in \mathbb{R}$ and $\boldsymbol{u} \in \mathbb{R}^d$.

Putting together all the pieces, we conclude that by taking $\gamma = \sigma_\varepsilon\sqrt{n/(d+\log n)}$, we have

$$\sup_{\boldsymbol{u}\in\mathbb{R}^d, x\in\mathbb{R}} \left|\mathbb{P}(n^{1/2}\langle \boldsymbol{u}, \hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\rangle \le \sigma x) - \Phi(x)\right| \lesssim \frac{d+\log n}{\sqrt{n}},$$

as long as $n \gtrsim d$.

□

A.3 Proof of Lemmas

A.3.1 Proof of Lemma 1.6.1

Proof. The proof is a simplified version of the proof of Lemma 2.5.1, which can be found in Appendix B.3.1. In the following, we outline the slight difference of the two proofs. Let $\boldsymbol{\delta} = \boldsymbol{\Sigma}^{1/2}(\boldsymbol{\beta} - \boldsymbol{\beta}^*)$ and $\boldsymbol{z}_i = \boldsymbol{\Sigma}^{-1/2}\boldsymbol{x}_i$. Using the arguments from the beginning of the proof of Lemma 2.5.1 to (B.20), it can be shown that $\mathbb{E}\{\mathscr{B}(\alpha)\} \ge 3/4$ provided that $\gamma \ge 4\sqrt{2}\max\{\sigma_\varepsilon, 2A_1^2 r\}$, where $\mathscr{B}(\alpha)$ is as defined in (B.19). Moreover, since $d < n$, by Cauchy-Schwarz inequality, we have

$$\begin{aligned}
\mathbb{E}(\Delta) &\le \frac{\gamma}{r}\mathbb{E}\left\{\sup_{\alpha\in\mathbb{B}(r)} \frac{1}{n}\sum_{i=1}^n \langle e_i \boldsymbol{z}_i, \alpha\rangle\right\} \\
&\le \frac{\gamma}{rn}\mathbb{E}\sup_{\alpha\in\mathbb{B}(r)} \Big|\Big|\sum_{i=1}^n e_i \boldsymbol{z}_i\Big|\Big|_2 \cdot ||\alpha||_2 \\
&\le \frac{\gamma}{r}\sqrt{\frac{d}{n}}.
\end{aligned}$$

Consequently, we have $\Delta \le 1/4$ with high probability provided that $n \gtrsim (\gamma/r)^2(d+t)$. Combining the above pieces finishes the proof. □

A.4 Proof of Propositions

A.4.1 Proof of Proposition 1.6.1

Proof. Let $\boldsymbol{\delta} = \boldsymbol{\beta}^* - \boldsymbol{\beta}^*_\gamma$. The optimality of $\boldsymbol{\beta}^*_\gamma$ and the mean value theorem indicate respectively that $\nabla \mathscr{R}(\boldsymbol{\beta}^*_\gamma) = \mathbf{0}$, and

$$\begin{aligned}\boldsymbol{\delta}^{\mathrm{T}} \nabla^2 \mathscr{R}(\tilde{\boldsymbol{\beta}}^*_\gamma)\boldsymbol{\delta} &= \langle \nabla \mathscr{R}(\boldsymbol{\beta}^*) - \nabla \mathscr{R}(\boldsymbol{\beta}^*_\gamma), \boldsymbol{\delta} \rangle \\ &= \langle \nabla \mathscr{R}(\boldsymbol{\beta}^*), \boldsymbol{\delta} \rangle = -\frac{1}{n}\sum_{i=1}^n \mathbb{E}\{w_\tau(\varepsilon_i)\ell'_\gamma(\varepsilon_i)\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\delta}\}, \end{aligned} \tag{A.17}$$

where $\tilde{\boldsymbol{\beta}}^*_\gamma = \lambda\boldsymbol{\beta}^* + (1-\lambda)\boldsymbol{\beta}^*_\gamma$ for some $0 \le \lambda \le 1$.

We start with an upper bound on the right-hand side of (A.17). By the fact that $\mathbb{E}\{w_\tau(\varepsilon)\varepsilon|\boldsymbol{x}\} = 0$ and $|\ell'(u) - u| \le u^2$, we have

$$\mathbb{E}\{w_\tau(\varepsilon)\ell'_\gamma(\varepsilon)|\boldsymbol{x}\} \le \mathbb{E}[\gamma w_\tau(\varepsilon)\{\ell'(\varepsilon/\gamma) - \varepsilon/\gamma\}|\boldsymbol{x}] \le \bar{\tau}\sigma_\varepsilon^2/\gamma.$$

Consequently

$$\mathbb{E}\{w_\tau(\varepsilon_i)\ell'_\gamma(\varepsilon_i)\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\delta}\} \le \mathbb{E}|\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\delta}| \cdot \bar{\tau}\sigma_\varepsilon^2/\gamma \le ||\boldsymbol{\Sigma}^{1/2}\boldsymbol{\delta}||_2 \cdot \bar{\tau}\sigma_\varepsilon^2/\gamma. \tag{A.18}$$

Next, we obtain a lower bound for $\boldsymbol{\delta}^{\mathrm{T}}\nabla^2\mathscr{R}(\tilde{\boldsymbol{\beta}}^*_\gamma)\boldsymbol{\delta}$. Let $L_{\tau,\infty}(\cdot)$ be the resulting asymmetric ℓ_2 loss when taking $\gamma = \infty$ in $L_{\tau,\gamma}(\cdot)$. Moreover, let $\mathscr{R}_\infty(\boldsymbol{\beta}) = \mathbb{E}\{L_{\tau,\infty}(y - \boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta})\}$. Since $\mathscr{R}(\cdot)$ is convex and minimized at $\boldsymbol{\beta}^*_\gamma$, we have

$$\mathscr{R}(\tilde{\boldsymbol{\beta}}^*_\gamma) \le \lambda\mathscr{R}(\boldsymbol{\beta}^*) + (1-\lambda)\mathscr{R}(\boldsymbol{\beta}^*_\gamma) \le \mathscr{R}(\boldsymbol{\beta}^*) \le \mathscr{R}_\infty(\boldsymbol{\beta}^*) \le \bar{\tau}\sigma_\varepsilon^2/2.$$

On the other hand, by the definition of Huber loss, for all $\boldsymbol{\beta} \in \mathbb{R}^d$, we have

$$\mathscr{R}(\boldsymbol{\beta}) \ge n^{-1}\sum_{i=1}^n \mathbb{E}w_\tau(y_i - \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta})(\gamma|y_i - \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}| - \gamma^2/2)\mathbb{1}(|y_i - \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}| > \gamma).$$

Let $\tilde{\varepsilon}_i = y_i - \boldsymbol{x}_i^{\mathrm{T}}\tilde{\boldsymbol{\beta}}_\gamma^*$. Combining the above inequalities, we have

$$\begin{aligned}\frac{\gamma}{n}\sum_{i=1}^{n}\mathbb{E}\{w_\tau(\tilde{\varepsilon}_i)|\tilde{\varepsilon}_i|\mathbb{1}(|\tilde{\varepsilon}_i|>\gamma)\} &\le \frac{\gamma^2}{2n}\sum_{i=1}^{n}\mathbb{E}\{w_\tau(\tilde{\varepsilon}_i)\mathbb{1}(|\tilde{\varepsilon}_i|>\gamma)\}+\frac{\bar{\tau}\sigma_\varepsilon^2}{2}\\ &\le \frac{\gamma}{2n}\sum_{i=1}^{n}\mathbb{E}\{w_\tau(\tilde{\varepsilon}_i)|\tilde{\varepsilon}_i|\mathbb{1}(|\tilde{\varepsilon}_i|>\gamma)\}+\frac{\bar{\tau}\sigma_\varepsilon^2}{2},\end{aligned}$$

which further implies that

$$\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\{w_\tau(\tilde{\varepsilon}_i)\mathbb{1}(|\tilde{\varepsilon}_i|>\gamma)\} \le \frac{1}{n\gamma}\sum_{i=1}^{n}\mathbb{E}\{w_\tau(\tilde{\varepsilon}_i)|\tilde{\varepsilon}_i|\mathbb{1}(|\tilde{\varepsilon}_i|>\gamma)\} \le \frac{\bar{\tau}\sigma_\varepsilon^2}{\gamma^2}. \tag{A.19}$$

Moreover, note that $\nabla^2\mathscr{R}(\tilde{\boldsymbol{\beta}}_\gamma^*) = n^{-1}\sum_{i=1}^n\mathbb{E}\{w_\tau(\tilde{\varepsilon}_i)\boldsymbol{x}_i\boldsymbol{x}_i^{\mathrm{T}}\} - n^{-1}\sum_{i=1}^n\mathbb{E}\{w_\tau(\tilde{\varepsilon}_i)\mathbb{1}(|\tilde{\varepsilon}_i|>\gamma)\boldsymbol{x}_i\boldsymbol{x}_i^{\mathrm{T}}\}$. It then follows from the Cauchy–Schwarz inequality and (A.19) that

$$\begin{aligned}\boldsymbol{\delta}^{\mathrm{T}}\nabla^2\mathscr{R}(\tilde{\boldsymbol{\beta}}_\gamma^*)\boldsymbol{\delta} &= \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\{w_\tau(\tilde{\varepsilon}_i)(\boldsymbol{\delta}^{\mathrm{T}}\boldsymbol{x}_i)^2\} - \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\{w_\tau(\tilde{\varepsilon}_i)\mathbb{1}(|\tilde{\varepsilon}_i|>\gamma)(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\delta})^2\}\\ &\ge \underline{\tau}||\boldsymbol{\Sigma}^{1/2}\boldsymbol{\delta}||_2^2 - \left\{\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}w_\tau^2(\tilde{\varepsilon}_i)\mathbb{1}^2(|\tilde{\varepsilon}_i|>\gamma)\right\}^{1/2}\left\{\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\delta})^4\right\}^{1/2}\\ &\ge \underline{\tau}||\boldsymbol{\Sigma}^{1/2}\boldsymbol{\delta}||_2^2 - \left\{\frac{1}{n}\sum_{i=1}^{n}\bar{\tau}\mathbb{E}w_\tau(\tilde{\varepsilon}_i)\mathbb{1}(|\tilde{\varepsilon}_i|>\gamma)\right\}^{1/2}\left\{\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\langle\boldsymbol{\Sigma}^{1/2}\boldsymbol{\delta},\boldsymbol{\Sigma}^{-1/2}\boldsymbol{x}_i\rangle^4\right\}^{1/2}\\ &\ge \underline{\tau}||\boldsymbol{\Sigma}^{1/2}\boldsymbol{\delta}||_2^2 - \frac{\bar{\tau}\sigma_\varepsilon}{\gamma}A_1^2||\boldsymbol{\Sigma}^{1/2}\boldsymbol{\delta}||_2^2.\end{aligned}$$

Picking $\gamma \ge 2\sigma_\varepsilon A_1^2\bar{\tau}/\underline{\tau}$, we have

$$\boldsymbol{\delta}^{\mathrm{T}}\nabla^2\mathscr{R}(\tilde{\boldsymbol{\beta}}_\gamma^*)\boldsymbol{\delta} \ge \underline{\tau}||\boldsymbol{\Sigma}^{1/2}\boldsymbol{\delta}||_2^2/2. \tag{A.20}$$

Putting together (A.17), (A.18), and (A.20) completes the proof. □

A.5 Proof of Technical Lemmas A.1.1–A.1.2

A.5.1 Proof of Lemma A.1.1

Proof. The proof is a simple combination of the proofs of Lemmas B.1.1 and B.1.2, and is thus omitted. Detailed proofs of Lemmas B.1.1 and B.1.2 can be found in Appendix B.5.

□

A.5.2 Proof of Lemma A.1.2

Proof. We start with obtaining an upper bound for $\mathbb{E}\{w_\tau(\varepsilon)\ell'_\gamma(\varepsilon)\}$. Denote $\ell(\cdot)$ as the Huber loss with $\gamma = 1$. By the fact that $|\ell'(u) - u| \leq |u|^3$ for all $u \in \mathbb{R}$, we have

$$\left|\mathbb{E}\{w_\tau(\varepsilon)\ell'_\gamma(\varepsilon)\}\right| = \left|\mathbb{E}\gamma w_\tau(\varepsilon)\left\{\ell'(\varepsilon/\gamma) - (\varepsilon/\gamma)\right\}\right| \leq \bar{\tau}\gamma^{-2}\mathbb{E}|\varepsilon|^3 = \bar{\tau}\gamma^{-2}\nu_3.$$

Turning to $\mathbb{E}\left\{w_\tau(\varepsilon)\ell'_\gamma(\varepsilon)\right\}^2$, note that $\mathbb{E}\ell'_\gamma(\varepsilon)^2 = \sigma_\varepsilon^2 - \mathbb{E}\varepsilon^2\mathbb{1}(|\varepsilon| > \gamma) + \gamma^2\mathbb{P}(|\varepsilon| > \gamma)$. By Markov's inequality, $\mathbb{E}(\varepsilon^2 - \gamma^2)\mathbb{1}(|\varepsilon| > \gamma) \leq \gamma^{-1}\mathbb{E}|\varepsilon|^3 = \gamma^{-1}\nu_3$. Combining this with the fact that $\underline{\tau} \leq w_\tau(\varepsilon) \leq \bar{\tau}$ and $|\ell'_\gamma(\varepsilon)| \leq |\varepsilon|$ completes the proof.

□

Appendix B

Supplementary Material for Chapter 2

B.1 Preliminary Results

Given $\tau \in (0,1)$, let $\{(y_i, \boldsymbol{x}_i)\}_{i=1}^n$ be a sample of independent data vectors from the linear regression model in (1.1), $y_i = \boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{\beta}^*(\tau) + \varepsilon_i(\tau)$, where $\varepsilon_i(\tau)$ satisfies $e_\tau(\varepsilon_i | \boldsymbol{x}_i) = 0$. In other words, the conditional τ-mean of y_i give $\boldsymbol{x}_i$ is a linear combination of $\boldsymbol{x}_i$. We suppress the dependency of $\boldsymbol{\beta}^*(\tau)$ and $\varepsilon(\tau)$ on τ throughout the Appendix. Let $w_\tau(u) := |\tau - \mathbb{1}(u < 0)|$ and let $\ell_\gamma(u) = \gamma^2 \ell(u/\gamma)$. Recall from (1.5) that $L(u) := L_{\tau,\gamma}(u) = w_\tau(u)\ell_\gamma(u)$ and let

$$\mathscr{R}_n(\boldsymbol{\beta}) = \frac{1}{n}\sum_{i=1}^n L(y_i - \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}) \quad \text{and} \quad \nabla\mathscr{R}_n(\boldsymbol{\beta}) = -\frac{1}{n}\sum_{i=1}^n L'(y_i - \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta})\boldsymbol{x}_i,$$

where $L'(u) = \gamma w_\tau(u)\ell'(u/\gamma)$ is the first-order derivative of $L(u)$.

For $\boldsymbol{\beta} \in \mathbb{R}^d$, let $\boldsymbol{w}(\boldsymbol{\beta}) = \nabla\mathscr{R}_n(\boldsymbol{\beta}) - \nabla\mathscr{R}(\boldsymbol{\beta})$, where $\mathscr{R}(\boldsymbol{\beta}) = \mathbb{E}\{\mathscr{R}_n(\boldsymbol{\beta})\}$ is the population loss. Moreover, we define the quantity $\boldsymbol{w}^* = \nabla\mathscr{R}_n(\boldsymbol{\beta}^*) - \nabla\mathscr{R}(\boldsymbol{\beta}^*)$ as the centered score function. Recall from Definition 2.5.1 that $\mathbb{C}(L) = \{\boldsymbol{\delta} : \|\boldsymbol{\delta}\|_1 \le L\|\boldsymbol{\delta}\|_2\}$. Let $\mathbb{C}_1 := \{\boldsymbol{\delta} : \|\boldsymbol{\delta}_{\mathscr{S}^{\mathrm{c}}}\|_1 \le 3\|\boldsymbol{\delta}_{\mathscr{S}}\|_1\}$. Moreover, define the symmetrized Bregman divergence $\mathscr{B} : \mathbb{R}^p \times \mathbb{R}^p \to [0,\infty)$ associated with the convex function $\mathscr{R}_n(\cdot)$ evaluated at $\boldsymbol{\beta}_1, \boldsymbol{\beta}_2$ as

$$\mathscr{B}(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2) = \langle \nabla\mathscr{R}_n(\boldsymbol{\beta}_1) - \nabla\mathscr{R}_n(\boldsymbol{\beta}_2), \boldsymbol{\beta}_1 - \boldsymbol{\beta}_2 \rangle. \tag{B.1}$$

Recall from Condition 2 that $\lambda_u \ge \lambda_{\max}(\boldsymbol{\Sigma})$, where $\boldsymbol{\Sigma} = \mathbb{E}(\boldsymbol{x}\boldsymbol{x}^{\mathrm{T}})$. Also recall from Condition 3

that $\mathbb{E}(\varepsilon^2|\boldsymbol{x}) \leq \sigma_\varepsilon^2$.

We present some technical lemmas that are useful for analyzing the high-dimensional penalized `retire` estimator. Recall that the penalized `retire` estimator is obtain by solving optimization problem (2.1). For notational convenience, throughout the Appendix, we define the minimizer of (2.1) as

$$\hat{\boldsymbol{\beta}}^{(t)} \in \underset{\boldsymbol{\beta}\in\mathbb{R}^d}{\operatorname{argmin}} \left\{ \mathscr{R}_n(\boldsymbol{\beta}) + \|\boldsymbol{\lambda}^{(t)} \circ \boldsymbol{\beta}\|_1 \right\}, \tag{B.2}$$

where $\boldsymbol{\lambda}^{(t)} = (\lambda_1^{(t)}, \ldots, \lambda_d^{(t)})^{\mathrm{T}}$ is a d-dimensional vector of tuning parameters with $\lambda_j^{(t)} = p'_\lambda(|\hat{\beta}_j^{(t-1)}|)$, and $\circ$ is the Hadamard product. Throughout the proof, we drop the superscript from $\hat{\boldsymbol{\beta}}^{(t)}$ and $\boldsymbol{\lambda}^{(t)}$ when the context is clear.

The proofs of all of the technical lemmas are deferred to Appendix B.5.

Lemma B.1.1. *Under Conditions 1, 2, and 3, we have*

$$\|\nabla\mathscr{R}(\boldsymbol{\beta}^*)\|_2 \leq \gamma^{-1}\bar{\tau}\lambda_u^{1/2}\sigma_\varepsilon^2 \;\; \textit{and} \;\; \|\nabla\mathscr{R}(\boldsymbol{\beta}^*)\|_\infty \leq \gamma^{-1}\bar{\tau}\sigma_{\boldsymbol{x}}\sigma_\varepsilon^2.$$

Moreover, for any $t \geq 0$,

$$\|\boldsymbol{w}^*\|_\infty = \|\nabla\mathscr{R}_n(\boldsymbol{\beta}^*) - \nabla\mathscr{R}(\boldsymbol{\beta}^*)\|_\infty \leq \nu_0\sigma_{\boldsymbol{x}}\bar{\tau}\left(2\sigma_\varepsilon\sqrt{\frac{\log d + t}{n}} + \gamma\frac{\log d + t}{n}\right)$$

holds with probability at least $1 - 2e^{-t}$.

Lemma B.1.1 reveals the proper range for the penalty level λ so that event $\{\lambda \geq 2\|\nabla\mathscr{R}_n(\boldsymbol{\beta}^*)\|_\infty\}$ occurs with high probability. Let $\mathscr{S}$ be the active set of the true regression parameter $\boldsymbol{\beta}^*$, and $\mathbf{S} = \mathbb{E}(\boldsymbol{x}_{\mathscr{S}}\boldsymbol{x}_{\mathscr{S}}^{\mathrm{T}})$ be the $s \times s$ principal submatrix of $\boldsymbol{\Sigma}$. Denote by $\lambda_{\max}(\mathbf{S})$ the maximal eigenvalue of $\mathbf{S}$. Write $\boldsymbol{w}^* = \nabla\mathscr{R}_n(\boldsymbol{\beta}^*) - \nabla\mathscr{R}(\boldsymbol{\beta}^*)$. The next lemma provides an upper bound for the centered score $\boldsymbol{w}^*$, projected on the true support $\mathscr{S}$.

Lemma B.1.2. *Under Conditions 2–3, for any $t > 0$, we have*

$$\|\boldsymbol{w}^*_{\mathscr{S}}\|_2 \le 3\bar{\tau}\nu_0\lambda_{\max}^{1/2}(\mathbf{S})\left(\sigma_\varepsilon\sqrt{\frac{2s+t}{n}} + \gamma\frac{2s+t}{2n}\right),$$

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

The following two lemmas contain some results for the solution of (B.2). Both lemmas are essential for the proof of Proposition B.2.1, which is the key to the proof of Theorem 2.5.2.

Lemma B.1.3. *Let $\mathscr{A}$ be a set such that $\mathscr{S} \subseteq \mathscr{A} \subseteq [d]$. For any $\boldsymbol{\beta} \in \mathbb{R}^d$, let $\boldsymbol{\beta}_{\mathscr{A}^c} = \mathbf{0}$. Assume that $\|\boldsymbol{\lambda}_{\mathscr{A}^c}\|_{\min} > \|\boldsymbol{w}(\boldsymbol{\beta})\|_\infty$. Then, any solution $\hat{\boldsymbol{\beta}}$ to the optimization problem* (B.2) *satisfies*

$$\|(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})_{\mathscr{A}^c}\|_1 \le \frac{\{\|\boldsymbol{\lambda}\|_\infty + \|\boldsymbol{w}(\boldsymbol{\beta})\|_\infty\}\|(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})_{\mathscr{A}}\|_1 + \|\nabla\mathscr{R}(\boldsymbol{\beta})\|_2\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_2}{\|\boldsymbol{\lambda}_{\mathscr{A}^c}\|_{\min} - \|\boldsymbol{w}(\boldsymbol{\beta})\|_\infty}. \tag{B.3}$$

Lemma B.1.4. *Let $\mathscr{A}$ be a set such that $\mathscr{S} \subseteq \mathscr{A} \subseteq [d]$ and $|\mathscr{A}| = k$. Let $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_d)^{\mathrm{T}}$ be a vector of tuning parameters that satisfies $\|\boldsymbol{\lambda}\|_\infty \le \lambda$ and $\|\boldsymbol{\lambda}_{\mathscr{A}^c}\|_{\min} \ge a\lambda$ for some constant $a \in (0,1]$ and $\lambda \ge s^{-1/2}\|\nabla\mathscr{R}(\boldsymbol{\beta}^*)\|_2$. Then, under the event $\{a\lambda \ge 2\|\boldsymbol{w}^*\|_\infty\}$, any solution $\hat{\boldsymbol{\beta}}$ to* (B.2) *satisfies $\hat{\boldsymbol{\beta}} \in \boldsymbol{\beta}^* + \mathbb{C}(L)$ with $L = (2+2/a)k^{1/2} + 2s^{1/2}/a$. In addition, let $\kappa, r > 0$ satisfy $r > \kappa^{-1}(2s^{1/2} + k^{1/2}a/2)\lambda$. Then, under the event $\mathscr{E}_{\mathrm{rsc}}(r, L, \kappa)$, we have*

$$\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2 \le \kappa^{-1}\{(2s^{1/2} + k^{1/2}a/2)\lambda\} < r.$$

B.2 Proof of Theorems

B.2.1 Proof of Theorem 2.5.1

Proof. Let $\hat{\boldsymbol{\beta}} := \hat{\boldsymbol{\beta}}^{(1)}$ be a minimizer of (2.1) with $p'_\lambda(0) = \lambda$, i.e., optimization problem (2.1) reduces to the ℓ_1-penalized robustified expectile regression, i.e.,

$$\hat{\boldsymbol{\beta}} \in \underset{\boldsymbol{\beta}\in\mathbb{R}^d}{\text{minimize}} \left\{\frac{1}{n}\sum_{i=1}^n L_{\tau,\gamma}(y_i - \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}) + \lambda\sum_{j=2}^d |\beta_j|\right\}. \tag{B.4}$$

Let $\mathscr{S} = \{1, \ldots, d\}$ be the active set of $\boldsymbol{\beta}^*$, i.e., the index set $\mathscr{S}$ contains indices for which $\boldsymbol{\beta}^*_j \neq 0$. Let $s = |\mathscr{S}|$ be the cardinality of $\mathscr{S}$. Recall the definition of the symmetric Bregman divergence in (B.1). The main crux of the proof of Theorem 2.5.1 involves establishing upper and lower bounds for $\mathscr{B}(\hat{\boldsymbol{\beta}}, \boldsymbol{\beta}^*)$. We start with deriving an upper bound for $\mathscr{B}(\hat{\boldsymbol{\beta}}, \boldsymbol{\beta}^*)$. Throughout the proof, we write $\hat{\boldsymbol{\delta}} = \hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*$.

Since $\hat{\boldsymbol{\beta}}$ is a minimizer of (B.4), we have

$$\begin{aligned} \mathscr{R}_n(\hat{\boldsymbol{\beta}}) - \mathscr{R}_n(\boldsymbol{\beta}^*) &\leq \lambda(\|\boldsymbol{\beta}^*\|_1 - \|\hat{\boldsymbol{\beta}}\|_1) && \text{(B.5)} \\ &\leq \lambda(\|\boldsymbol{\beta}^*_{\mathscr{S}}\|_1 - \|\boldsymbol{\beta}^*_{\mathscr{S}} + \hat{\boldsymbol{\delta}}_{\mathscr{S}}\|_1 - \|\hat{\boldsymbol{\delta}}_{\mathscr{S}^{\mathrm{c}}}\|_1) \\ &\leq \lambda(\|\hat{\boldsymbol{\delta}}_{\mathscr{S}}\|_1 - \|\hat{\boldsymbol{\delta}}_{\mathscr{S}^{\mathrm{c}}}\|_1). && \text{(B.6)} \end{aligned}$$

By the optimality condition of $\hat{\boldsymbol{\beta}}$, we have $\langle \nabla \mathscr{R}_n(\hat{\boldsymbol{\beta}}) + \lambda \hat{\boldsymbol{z}}, \hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^* \rangle \leq 0$, where $\hat{\boldsymbol{z}} \in \partial \|\hat{\boldsymbol{\beta}}\|_1$ and $\langle \hat{\boldsymbol{z}}, \hat{\boldsymbol{\beta}} \rangle = \|\hat{\boldsymbol{\beta}}\|_1$. Thus, conditioned on the event $\mathscr{E}_{\text{score}} := \{\lambda \geq 2\|\nabla \mathscr{R}_n(\boldsymbol{\beta}^*)\|_\infty\}$, $\mathscr{B}(\hat{\boldsymbol{\beta}}, \boldsymbol{\beta}^*)$ can be upper bounded by

$$\begin{aligned} \mathscr{B}(\hat{\boldsymbol{\beta}}, \boldsymbol{\beta}^*) &= \langle \nabla \mathscr{R}_n(\hat{\boldsymbol{\beta}}) - \nabla \mathscr{R}_n(\boldsymbol{\beta}^*), \hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^* \rangle && \text{(B.7)} \\ &= \langle \nabla \mathscr{R}_n(\hat{\boldsymbol{\beta}}) + \lambda \hat{\boldsymbol{z}}, \hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^* \rangle + \langle -\lambda \hat{\boldsymbol{z}} - \nabla \mathscr{R}_n(\boldsymbol{\beta}^*), \hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^* \rangle \\ &\leq 0 + \lambda(\|\boldsymbol{\beta}^*\|_1 - \|\hat{\boldsymbol{\beta}}\|_1) + \|\nabla \mathscr{R}_n(\boldsymbol{\beta}^*)\|_\infty \cdot \|\hat{\boldsymbol{\delta}}\|_1 \\ &\leq \lambda(\|\hat{\boldsymbol{\delta}}_{\mathscr{S}}\|_1 - \|\hat{\boldsymbol{\delta}}_{\mathscr{S}^{\mathrm{c}}}\|_1) + \frac{\lambda}{2}(\|\hat{\boldsymbol{\delta}}_{\mathscr{S}}\|_1 + \|\hat{\boldsymbol{\delta}}_{\mathscr{S}^{\mathrm{c}}}\|_1) \\ &\leq \frac{3}{2}\lambda s^{1/2}\|\hat{\boldsymbol{\delta}}\|_2. && \text{(B.8)} \end{aligned}$$

We now obtain a lower bound for $\mathscr{B}(\hat{\boldsymbol{\beta}}, \boldsymbol{\beta}^*)$. To this end, we apply the restricted strong convexity result in Lemma 2.5.1. First, from the proof of Lemma 2.5.1, we know that the result in Lemma 2.5.1 is applicable for any $\boldsymbol{\beta} \in \boldsymbol{\beta}^* + \mathbb{B}(r) \cap \mathbb{C}_1$, for which $\hat{\boldsymbol{\beta}}$ does not necessarily satisfies. To this end, we define an intermediate quantity to help facilitate the proof. Let A_1 be a constant that satisfies $\mathbb{E}(\boldsymbol{u}^{\mathrm{T}} \boldsymbol{\Sigma}^{-1/2} \boldsymbol{x})^4 \leq A_1^4 \|\boldsymbol{u}\|_2^4$ for all $\boldsymbol{u} \in \mathbb{R}^d$ and let $r_{\text{loc}} = \gamma/(8\sqrt{2}\lambda_u A_1^2)$. Consider

$\hat{\boldsymbol{\beta}}_\eta = \eta\hat{\boldsymbol{\beta}} + (1-\eta)\boldsymbol{\beta}^*$, where $\eta = \sup\{u \in [0,1] : (1-u)\boldsymbol{\beta}^* + u\hat{\boldsymbol{\beta}} \in \boldsymbol{\beta}^* + \mathbb{B}(r_{\text{loc}})\}$. Then, $\hat{\boldsymbol{\beta}}_\eta \in \boldsymbol{\beta}^* + \partial\mathbb{B}(r_{\text{loc}})$ whenever $\hat{\boldsymbol{\beta}} \notin \boldsymbol{\beta}^* + \mathbb{B}(r_{\text{loc}})$ where $\partial\mathbb{B}(r_{\text{loc}})$ is the boundary of $\mathbb{B}(r_{\text{loc}})$, and $\hat{\boldsymbol{\beta}}_\eta = \hat{\boldsymbol{\beta}}$ whenever $\hat{\boldsymbol{\beta}} \in \boldsymbol{\beta}^* + \mathbb{B}(r_{\text{loc}})$. Let $\hat{\boldsymbol{\delta}}_\eta = \hat{\boldsymbol{\beta}}_\eta - \boldsymbol{\beta}^*$. It remains to show that $\hat{\boldsymbol{\beta}}_\eta \in \boldsymbol{\beta}^* + \mathbb{C}_1$, i.e., $\|(\hat{\boldsymbol{\delta}}_\eta)_{S^c}\|_1 \le 3\|(\hat{\boldsymbol{\delta}}_\eta)_S\|_1$. By convexity of $\mathscr{R}_n(\cdot)$, we have

$$\mathscr{R}_n(\hat{\boldsymbol{\beta}}) - \mathscr{R}_n(\boldsymbol{\beta}^*) \ge \langle \nabla\mathscr{R}_n(\boldsymbol{\beta}^*), \hat{\boldsymbol{\delta}}\rangle \ge -\|\nabla\mathscr{R}_n(\boldsymbol{\beta}^*)\|_\infty \cdot \|\hat{\boldsymbol{\delta}}\|_1 \ge -\frac{\lambda}{2}(\|\hat{\boldsymbol{\delta}}_{\mathscr{S}}\|_1 + \|\hat{\boldsymbol{\delta}}_{\mathscr{S}^c}\|_1) \quad \text{(B.9)}$$

Combining (B.6) and (B.9), we have $\|\hat{\boldsymbol{\delta}}_{\mathscr{S}^c}\|_1 \le 3\|\hat{\boldsymbol{\delta}}_{\mathscr{S}}\|_1$, conditioned on the event $\mathscr{E}_{\text{score}}$. Since $\eta\hat{\boldsymbol{\delta}} = \hat{\boldsymbol{\delta}}_\eta$, we have verified that $\hat{\boldsymbol{\beta}}_\eta \in \boldsymbol{\beta}^* + \mathbb{B}(r_{\text{loc}}) \cap \mathbb{C}_1$, conditioned on $\mathscr{E}_{\text{score}}$. Applying Lemma 2.5.1 with $\boldsymbol{\beta} = \hat{\boldsymbol{\beta}}_\eta$, the following bound holds with probability at least $1 - e^{-t}$

$$\mathscr{B}(\hat{\boldsymbol{\beta}}_\eta, \boldsymbol{\beta}^*) = \langle \nabla\mathscr{R}_n(\hat{\boldsymbol{\beta}}_\eta) - \nabla\mathscr{R}_n(\boldsymbol{\beta}^*), \hat{\boldsymbol{\beta}}_\eta - \boldsymbol{\beta}^*\rangle \ge \frac{1}{2}\kappa_1\underline{\tau}\|\hat{\boldsymbol{\delta}}_\eta\|_2^2, \quad \text{(B.10)}$$

as long as (γ, n, d) satisfies $\gamma \ge 4\sqrt{2}\lambda_u\sigma_\varepsilon$ and $n \gtrsim s\log d + t$. For notational convenience, we denote the event at (B.10) as $\mathscr{E}_{\text{rsc}}$ with $\mathbb{P}(\mathscr{E}_{\text{rsc}}) \ge 1 - e^{-t}$.

We now combine the lower and upper bounds in (B.8) and (B.10). Since $\mathscr{R}_n(\cdot)$ is convex, by Lemma C.1 in Sun, Zhou and Fan (2020) we have $\mathscr{B}(\hat{\boldsymbol{\beta}}_\eta, \boldsymbol{\beta}^*) \le \eta\mathscr{B}(\hat{\boldsymbol{\beta}}, \boldsymbol{\beta}^*)$, and thus implies that

$$\begin{cases} \|\hat{\boldsymbol{\delta}}_\eta\|_2 \le 3(\kappa_1\underline{\tau})^{-1}s^{1/2}\lambda; \\ \|\hat{\boldsymbol{\delta}}_\eta\|_1 \le 4s^{1/2}\|\hat{\boldsymbol{\delta}}_\eta\|_2 \le 12(\kappa_1\underline{\tau})^{-1}s\lambda. \end{cases}$$

We now show that with proper choice of λ and γ, $\hat{\boldsymbol{\beta}} \in \boldsymbol{\beta}^* + \mathbb{B}(r_{\text{loc}})$, implying $\hat{\boldsymbol{\beta}}_\eta = \hat{\boldsymbol{\beta}}$. Let $\gamma = \sigma_\varepsilon\sqrt{n/(\log d + t)}$. By Lemma B.1.1,

$$\|\nabla\mathscr{R}_n(\boldsymbol{\beta}^*)\|_\infty \le \bar{\tau}\sigma_{\boldsymbol{x}}\sigma_\varepsilon(3\nu_0 + 1)\sqrt{(\log d + t)/n}$$

with probability at least $1 - 2e^{-t}$, suggesting that $\lambda = 2c\bar{\tau}\sqrt{(\log d + t)/n}$ where $c = \sigma_{\boldsymbol{x}}\sigma_\varepsilon(3\nu_0 + 1)$. Moreover, it can be verified that $\gamma \ge 4\sqrt{2}\lambda_u\sigma_\varepsilon$ under the scaling condition $n \gtrsim s\log d + t$.

Finally, we have $\|\hat{\boldsymbol{\beta}}_\eta - \boldsymbol{\beta}^*\|_2 \leq 3(\kappa_1 \underline{\tau})^{-1} s^{1/2}\lambda \lesssim \sqrt{s(\log d + t)/n} < \sqrt{n/(\log d + t)} \asymp r_{\text{loc}}$, i.e., $\hat{\boldsymbol{\beta}}_\eta \in \boldsymbol{\beta}^* + \mathbb{B}(r_{\text{loc}})$. This further implies that $\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}_\eta$ by construction. Thus, we obtain the desired results

$$\begin{cases} \|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2 \leq 3(\kappa_1 \underline{\tau})^{-1} s^{1/2}\lambda; \\ \|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_1 \leq 12(\kappa_1 \underline{\tau})^{-1} s\lambda, \end{cases}$$

with probability $\mathbb{P}\big(\mathscr{E}_{\text{score}} \cap \mathscr{E}_{\text{rsc}}\big) \geq 1 - 3e^{-t}$.

□

B.2.2 Proof of Theorem 2.5.2

Recall that $\mathscr{R}(\boldsymbol{\beta}^*) = \mathbb{E}\{\mathscr{R}_n(\boldsymbol{\beta}^*)\}$ be the population loss evaluated at $\boldsymbol{\beta}^*$ and let $\boldsymbol{w}^* = \nabla\mathscr{R}_n(\boldsymbol{\beta}^*) - \nabla\mathscr{R}(\boldsymbol{\beta}^*)$ be the centered score function. We first show that given an estimator at the $(T-1)th$ iteration, $\hat{\boldsymbol{\beta}}^{(T-1)}$, the estimation error of the subsequent estimator $\hat{\boldsymbol{\beta}}^{(T)}$ can be improved sequentially by a δ-fraction for some constant $\delta \in (0,1)$, under a beta-min condition on $\|\boldsymbol{\beta}^*_{\mathscr{S}}\|_{\min}$. We establish a deterministic claim in the following proposition, where we conditioned on events that are related to the local restricted strong convexity property and the gradient of the loss function, $\mathscr{E}_{\text{rsc}}(r, L, \kappa)$ and $\{p_0'(a_0)\lambda \geq 2\boldsymbol{w}^*\}$, respectively.

Proposition B.2.1. *Let $p_0(\cdot)$ be a penalty function that satisfies Condition 6. Given $\kappa > 0$, assume that there exists some constant $a_0 > 0$ such that $p_0'(a_0) > 0$ and $\kappa > \sqrt{5}/(2a_0)$. Let $c > 0$ be a constant that is the solution to the equation*

$$0.5p_0'(a_0)(c^2+1)^{1/2} + 2 = c\kappa a_0. \tag{B.11}$$

Assume the beta-min condition $\|\boldsymbol{\beta}^*_{\mathscr{S}}\|_{\min} \geq a_0\lambda$ and let $r^{\text{crude}} = ca_0 s^{1/2}\lambda$. Conditioned on the event $\mathscr{E}_{\text{rsc}}(r, L, \kappa) \cap \{p_0'(a_0)\lambda \geq 2\|\boldsymbol{w}^*\|_\infty\}$ with

$$L = \{2 + 2/p_0'(a_0)\}(c^2+1)^{1/2}s^{1/2} + 2s^{1/2}/p_0'(a_0),\ r > r^{\text{crude}},\ \text{and}\ \lambda \geq s^{-1/2}\|\nabla\mathscr{R}(\boldsymbol{\beta}^*)\|_2,$$

the sequence of solutions $\hat{\boldsymbol{\beta}}^{(1)}, \ldots, \hat{\boldsymbol{\beta}}^{(T)}$ *obtained from solving* (2.1) *satisfies*

$$\|\hat{\boldsymbol{\beta}}^{(T)} - \boldsymbol{\beta}^*\|_2 \leq \delta \|\hat{\boldsymbol{\beta}}^{(T-1)} - \boldsymbol{\beta}^*\|_2 + \kappa^{-1} \left\{ \|p'_\lambda\{(|\boldsymbol{\beta}^*_j| - a_0\lambda)_+\}\|_2 + \|\boldsymbol{w}^*_{\mathscr{S}}\|_2 + \|\nabla \mathscr{R}(\boldsymbol{\beta}^*)\|_2 \right\}, \quad \text{(B.12)}$$

where $\delta = \sqrt{5}/(2a_0\kappa) \in (0,1)$ *and* $z_+ = \max(z, 0)$. *Furthermore, we have*

$$\|\hat{\boldsymbol{\beta}}^{(T)} - \boldsymbol{\beta}^*\|_2 \leq \delta^{T-1} r^{\text{crude}} + \{(1-\delta)\kappa\}^{-1} \left\{ \|p'_\lambda\{(|\boldsymbol{\beta}^*_j| - a_0\lambda)_+\}\|_2 + \|\boldsymbol{w}^*_{\mathscr{S}}\|_2 + \|\nabla \mathscr{R}(\boldsymbol{\beta}^*)\|_2 \right\}. \quad \text{(B.13)}$$

Proposition B.2.1 establishes the fact that every additional iteration of the proposed iteratively reweighted method shrinks the estimation error of the solution obtained from the previous iteration by a factor of $\delta \in (0,1)$, at the cost of inducing some extra terms $\|p'_\lambda\{(|\boldsymbol{\beta}^*_j| - a_0\lambda)_+\}\|_2$, $\|\boldsymbol{w}^*_{\mathscr{S}}\|_2$, and $\|\nabla \mathscr{R}(\boldsymbol{\beta}^*)\|_2$, which can be shown to be smaller than r^{crude}. Such a phenomenon is also known as the contraction property and has been studied in different contexts (Fan et al., 2018, Pan, Sun and Zhou, 2021). We refer the reader to Pan, Sun and Zhou (2021) for a detailed discussion on the various terms that appear in (B.13). For completeness, we also provide the proof of Proposition B.2.1 in Appendix B.4.1.

The results in Proposition B.2.1 are deterministic, conditioned on some events. In the following proof of Theorem 2.5.2, we provide an appropriate choice of the set of tuning parameters (λ, γ) such that the event $\mathscr{E}_{\text{rsc}}(r, L, \kappa) \cap \{p'_0(a_0)\lambda \geq 2\|\boldsymbol{w}^*\|_\infty\}$ holds with high probability. Moreover, we will control the shrinkage bias $\|p'_\lambda\{(|\boldsymbol{\beta}^*_j| - a_0\lambda)_+\}\|_2$ in (B.13) by proposing slightly stronger conditions on the minimum signal strength $\|\boldsymbol{\beta}^*_{\mathscr{S}}\|_{\min}$ as well as the first derivative of the penalty function $p_\lambda(\cdot)$.

Proof. The proof is based on Proposition B.2.1. We will show that under the stated conditions in Theorem 2.5.2, the events $\mathscr{E}_{\text{rsc}}(r, L, \kappa)$ and $\{p'_0(a_0)\lambda \geq 2\|\boldsymbol{w}^*\|_\infty\}$ in Proposition B.2.1 hold with high probabilities. We then show that the terms $p'_\lambda\{(|\boldsymbol{\beta}^*_j| - a_0\lambda)_+\}\|_2$, $\|\boldsymbol{w}^*_{\mathscr{S}}\|_2$, and $\|\nabla \mathscr{R}(\boldsymbol{\beta}^*)\|_2$ can be upper bounded with high probabilities.

Picking $\gamma = \sigma_\varepsilon \sqrt{n/(s+\log d+t)}$ and applying Lemma B.1.1 indicates that

$$\|\nabla \mathscr{R}(\boldsymbol{\beta}^*)\|_2 \leq \bar{\tau}\sigma_\varepsilon \lambda_u^{1/2} \sqrt{(s+\log d+t)/n}.$$

and

$$\|\nabla \mathscr{R}_n(\boldsymbol{\beta}^*) - \nabla \mathscr{R}(\boldsymbol{\beta}^*)\|_\infty \leq 3\nu_0 \bar{\tau}\sigma_{\boldsymbol{x}}\sigma_\varepsilon \sqrt{(\log d+t)/n},$$

with probability at least $1-2e^{-t}$. Picking $\lambda \asymp \sigma_\varepsilon \sqrt{(\log d+t)/n}$, we have $\lambda \geq s^{-1/2}\|\nabla \mathscr{R}(\boldsymbol{\beta}^*)\|_2$ and the event $\{p_0'(a_0)\lambda \geq 2\|\boldsymbol{w}^*\|_\infty\}$ holds with probability at least $1-2e^{-t}$.

Next, we set $\kappa = 0.5\kappa_1 \underline{\tau}$ and set the constant c to be the solution to (B.11). Picking $r = \gamma/(8\sqrt{2}\lambda_u A_1^2)$, it can be shown that $r \asymp \sigma_\varepsilon \sqrt{n/(s+\log d+t)} > \sigma_\varepsilon \sqrt{s(\log d+t)/n} \asymp r^{\text{crude}}$ and $\delta = \sqrt{5}/(a_0\kappa_1\underline{\tau}) < 1$. Thus, setting $L = \{2+\frac{2}{p_0'(a_0)}\}(c^2+1)^{1/2}s^{1/2} + \frac{2}{p_0'(a_0)}s^{1/2}$, Lemma 2.5.1 indicates that the event $\mathscr{E}_{\text{rsc}}(r, L, 0.5\kappa_1\underline{\tau})$ holds with probability at least $1-e^{-t}$.

Moreover, by Lemma B.1.2 and the choice of $\gamma = \sigma_\varepsilon \sqrt{n/(s+\log d+t)}$, we obtain

$$\|\boldsymbol{w}^*_{\mathscr{S}}\|_2 \lesssim \sigma_\varepsilon \sqrt{\frac{s+t}{n}}, \tag{B.14}$$

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

Finally, we obtain an upper bound for the term $\|p_\lambda'(|\boldsymbol{\beta}^*_{\mathscr{S}}| - a_0\lambda)_+\|_2$. Since $|\beta_j^*| \geq (a_0+a_1)\lambda$ for any $j \in \mathscr{S}$, we have $p_\lambda'(|\beta_j^*| - a_0\lambda) = 0$. Combining the aforementioned inequalities to (B.13), we obtain

$$\|\hat{\boldsymbol{\beta}}^{(T)} - \boldsymbol{\beta}^*\|_2 \lesssim \delta^{T-1}\sigma_\varepsilon \sqrt{\frac{s(\log d+t)}{n}} + \frac{\sigma_\varepsilon}{1-\delta}\sqrt{\frac{s+\log d+t}{n}},$$

with probability at least $1-4e^{-t}$. Setting $T \gtrsim \frac{\log\{\log(d+t)\}}{\log(1/\delta)}$ leads to the desired results in (2.5) and (2.6). □

B.3 Proof of Lemmas

B.3.1 Proof of Lemma 2.5.1

Proof. For notational convenience, throughout the proof we let $\boldsymbol{\delta} = \boldsymbol{\beta} - \boldsymbol{\beta}^*$. Recall from Definition 2.5.1 that $\mathbb{B}(r) = \{\boldsymbol{\delta} \in \mathbb{R}^d : \|\boldsymbol{\delta}\|_2 \leq r\}$ is a ball and $\mathbb{C}(L) = \{\boldsymbol{\delta} : \|\boldsymbol{\delta}\|_1 \leq L\|\boldsymbol{\delta}\|_2\}$ is an ℓ_1-cone. In the following proof, we will provide a lower bound for the symmetrized Bregman divergence $\mathscr{B}(\boldsymbol{\beta}, \boldsymbol{\beta}^*)$ under the constraint $\boldsymbol{\beta} \in \boldsymbol{\beta}^* + \mathbb{B}(r) \cap \mathbb{C}(L)$.

We start by defining the events

$$E_i(\boldsymbol{\delta}, r, \gamma) = \{|\varepsilon_i| \leq \gamma/2\} \cap \left\{ |\boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{\delta}| \leq \frac{\gamma \|\boldsymbol{\delta}\|_2}{2r} \right\} \tag{B.15}$$

for $i = 1, \ldots, n$. The symmetrized Bregman divergence can then be low bounded by

$$\begin{aligned} \mathscr{B}(\boldsymbol{\beta}, \boldsymbol{\beta}^*) &= \frac{1}{n} \sum_{i=1}^{n} \left\{ L'(\varepsilon_i) - L'(\varepsilon_i - \boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{\delta}) \right\} \cdot \boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{\delta} \\ &\geq \frac{1}{n} \sum_{i=1}^{n} \left\{ L'(\varepsilon_i) - L'(\varepsilon_i - \boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{\delta}) \right\} \cdot \boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{\delta} \cdot \mathbb{1}_{E_i(\boldsymbol{\delta}, r, \gamma)}, \end{aligned} \tag{B.16}$$

where $\mathbb{1}_{E_i(\boldsymbol{\delta}, r, \gamma)}$ is an indicator function that takes value one when the event in (B.15) holds and zero otherwise. Thus, it suffices to obtain a lower bound on (B.16) for any $\boldsymbol{\beta} \in \boldsymbol{\beta}^* + \mathbb{B}(r) \cap \mathbb{C}(L)$.

Recall from Appendix B.1 that $L'(u) = \gamma w_\tau(u) \ell'(u/\gamma)$ with $w_\tau(u) = |\tau - I(u < 0)|$. Conditioned on the event $E_i(\boldsymbol{\delta}, r, \gamma)$, for any $\boldsymbol{\delta} \in \mathbb{B}(r)$, we have $|\varepsilon_i| \leq \gamma$ and $|\varepsilon_i - \boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{\delta}| \leq \gamma/2 + \gamma/2 = \gamma$. For notational convenience, let $u_i = \varepsilon_i$ and let $v_i = \varepsilon_i - \boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{\delta}$. Then, the term $\left\{ L'(\varepsilon_i) - L'(\varepsilon_i - \boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{\delta}) \right\} \cdot \boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{\delta}$ can be rewritten as $\{L'(u_i) - L'(v_i)\}(u_i - v_i)$. In the following, we obtain a lower bound for the term $\{L'(u_i) - L'(v_i)\}(u_i - v_i)$ for any $u_i, v_i \in [-\gamma, \gamma]$. Let $\kappa_1 = \min_{|t| \leq 1} \ell''(t)$. To this end, we consider three possible cases:

(i) $(u_i v_i = 0)$. If $v_i = 0$, we have $\{L'(u_i) - L'(v_i)\}(u_i - v_i) \geq \gamma w_\tau(u_i) \{\ell'(u_i/\gamma) - \ell'(0)\} u_i \geq \kappa_1 \underline{\tau} u_i^2$, where the last inequality hold by the mean value theorem. Similarly if $u_i = 0$, $\{L'(u_i) - L'(v_i)\}(u_i - v_i) \geq \kappa_1 \underline{\tau} v_i^2$.

(ii) $(u_i v_i > 0)$. In this case, $w_\tau(u_i) = w_\tau(v_i)$ and hence $\{L'(u_i) - L'(v_i)\}(u_i - v_i) = \gamma w_\tau(u_i)\{\ell'(u_i/\gamma) - \ell'(v_i/\gamma)\}(u_i - v_i) \geq \kappa_1 \underline{\tau}(u_i - v_i)^2$.

(iii) $(u_i v_i < 0)$. In this case, we have either $u > 0, v < 0$ or $u < 0, v > 0$. For the former, $\{L'(u_i) - L'(v_i)\}(u_i - v_i) = \gamma\{\tau\ell'(u_i/\gamma) - (1-\tau)\ell'(v_i/\gamma)\}(u_i - v_i) \geq \kappa_1 \underline{\tau}(u_i - v_i)^2$, where the last inequality holds by the mean value theorem. The latter can be shown in a similar fashion.

Combining all three cases, we conclude that $\{L'(u_i) - L'(v_i)\}(u_i - v_i) \geq \kappa_1 \underline{\tau}(u_i - v_i)^2$ for all $u_i, v_i \in [-\gamma, \gamma]$. Substituting this into (B.16) yields

$$\mathscr{B}(\boldsymbol{\beta}, \boldsymbol{\beta}^*) \geq \frac{\kappa_1 \underline{\tau}}{n} \sum_{i=1}^{n} (\boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{\delta})^2 \mathbb{1}_{E_i(\boldsymbol{\delta}, r, \gamma)} \tag{B.17}$$

for any $\boldsymbol{\delta} \in \mathbb{B}(r)$.

Next, we will derive a lower bound for $(1/n)\sum_{i=1}^{n} (\boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{\delta})^2 \mathbb{1}_{E_i(\boldsymbol{\delta}, r, \gamma)}$, uniformly over $\boldsymbol{\delta} \in \mathbb{B}(r)$. To this end, we smooth the discontinuous indicator function $\mathbb{1}_{E_i(\boldsymbol{\delta}, r, \gamma)} = \mathbb{1}_{\{|\boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{\delta}| \leq \gamma \|\boldsymbol{\delta}\|_2 / (2r)\}} \cdot \mathbb{1}_{\{|\varepsilon_i| \leq \gamma/2\}}$ by a Lipschitz continuous function. Using similar ideas from the proof of Proposition 2 in Loh (2017), for any $R \geq 0$, we define the truncated squared function as

$$\varphi_R(u) = u^2 \mathbb{1}(|u| \leq R/2) + (|u| - R)^2 \mathbb{1}(R/2 < |u| \leq R), \;\; u \in \mathbb{R}.$$

It can be verified that the function $\varphi_R(\cdot)$ is R-Lipschitz continuous and satisfies the following:

$$u^2 \mathbb{1}(|u| \leq R/2) \leq \varphi_R(u) \leq \min\big\{u^2 \mathbb{1}(|u| \leq R), (R/2)^2\big\} \quad \text{and} \quad \varphi_{cR}(cu) = c^2 \varphi_R(u) \;\; \text{for any } c \geq 0. \tag{B.18}$$

It then follows from (B.17) and (B.18) that

$$\mathscr{B}(\boldsymbol{\beta},\boldsymbol{\beta}^*) \geq \kappa_1 \underline{\tau} \|\boldsymbol{\delta}\|_2^2 \cdot \underbrace{\frac{1}{n}\sum_{i=1}^{n} \mathbb{1}(|\varepsilon_i| \leq \gamma/2) \cdot \varphi_{\gamma/(2r)}(\boldsymbol{x}_i^{\mathrm{T}}\alpha)}_{=:\mathscr{B}(\alpha)}, \quad \text{where } \alpha := \boldsymbol{\delta}/\|\boldsymbol{\delta}\|_2 \in \mathbb{S}^{d-1}. \tag{B.19}$$

Next, we bound the random quantity $\mathscr{B}(\alpha)$ from below. Let $\Delta = \sup_{\alpha\in\mathbb{S}^{d-1}} -\mathscr{B}(\alpha) + \mathbb{E}\{\mathscr{B}(\alpha)\}$. Then, we have $\mathscr{B}(\alpha) \geq \mathbb{E}\{\mathscr{B}(\alpha)\} - \Delta$. It suffices to obtain a lower bound for $\mathbb{E}\{\mathscr{B}(\alpha)\}$ and an upper bound for the random fluctuation Δ. We start with obtaining a lower bound for $\mathbb{E}\{\mathscr{B}(\alpha)\}$.

Recall that A_1 is a constant that satisfies $\mathbb{E}\{(\boldsymbol{u}^{\mathrm{T}}\boldsymbol{x})^4\} \leq A_1^4\|\boldsymbol{u}\|_{\boldsymbol{\Sigma}}^4 \leq \lambda_u^2 A_1^4 \|\boldsymbol{u}\|_2^4$ for all $\boldsymbol{u} \in \mathbb{R}^d$. Applying the inequality in (B.18), for any $\alpha \in \mathbb{S}^{d-1}$, we have

$$\begin{aligned}
\mathbb{E}\{\mathscr{B}(\alpha)\} &\geq \mathbb{E}\big\{(\boldsymbol{x}_i^{\mathrm{T}}\alpha)^2 \mathbb{1}(|\varepsilon_i| \leq \gamma/2)\mathbb{1}(|\boldsymbol{x}_i^{\mathrm{T}}\alpha| \leq \gamma/4r)\big\} \\
&\geq \mathbb{E}\Big[(\boldsymbol{x}_i^{\mathrm{T}}\alpha)^2\big\{1 - \mathbb{1}(|\boldsymbol{x}_i^{\mathrm{T}}\alpha| > \gamma/4r) - \mathbb{1}(|\varepsilon_i| > \gamma/2)\big\}\Big] \\
&\geq 1 - (4r/\gamma)^2 \mathbb{E}(\boldsymbol{x}_i^{\mathrm{T}}\alpha)^4 - \mathbb{E}\Big[(\boldsymbol{x}_i^{\mathrm{T}}\alpha)^2 \mathbb{E}\big\{(2|\varepsilon_i|/\gamma)^2 | \boldsymbol{x}_i\big\}\Big] \\
&\geq 1 - (4r/\gamma)^2 \lambda_u^2 A_1^4 - (2/\gamma)^2 \sigma_\varepsilon^2 \lambda_u
\end{aligned} \tag{B.20}$$

Provided $\gamma \geq 4\sqrt{2}\lambda_u \max\{\sigma_\varepsilon, 2A_1^2 r\}$, we obtain $\mathbb{E}\{\mathscr{B}(\alpha)\} \geq 3/4$.

Next, we obtain an upper bound for $\Delta = \sup_{\alpha\in\mathbb{S}^{d-1}} -\mathscr{B}(\alpha) + \mathbb{E}\{\mathscr{B}(\alpha)\}$. Applying the inequality in (B.18) on $\varphi_{\gamma/(2r)}(\cdot)$, we have $\mathscr{B}(\alpha) \leq (\gamma/4r)^2$. Applying Theorem 7.3 in Bousquet (2003) and the inequality $ab \leq a^2/4 + b^2$, for any $t \geq 0$, we obtain

$$\begin{aligned}
\Delta &\leq \mathbb{E}(\Delta) + \sqrt{\frac{\gamma^2 t}{4r^2 n}\mathbb{E}(\Delta)} + \lambda_u A_1^2 \sqrt{\frac{2t}{n}} + \frac{\gamma^2}{48r^2}\cdot\frac{t}{n} \\
&\leq 1.25\mathbb{E}(\Delta) + \lambda_u A_1^2 \sqrt{\frac{2t}{n}} + \frac{\gamma^2}{3r^2}\cdot\frac{t}{n},
\end{aligned} \tag{B.21}$$

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

It remains to bound $\mathbb{E}(\Delta)$. Let $\mathscr{B}_i(\alpha) = \mathbb{1}(|\varepsilon_i| \leq \gamma/2) \cdot \varphi_{\gamma/(2r)}(\boldsymbol{x}_i^{\mathrm{T}}\alpha)$ and note that

$\mathbb{E}(\Delta) = \mathbb{E}\big[\sup_{\alpha\in\mathbb{S}^{d-1}}\big\{-(1/n)\sum_{i=1}^n \mathscr{B}_i(\alpha) + (1/n)\sum_{i=1}^n \mathbb{E}\mathscr{B}_i(\alpha)\big\}\big]$. By the symmetrization inequality for empirical process, $\mathbb{E}(\Delta) \le 2\mathbb{E}\big\{\sup_{\alpha\in\mathbb{S}^{d-1}}(1/n)\sum_{i=1}^n e_i\mathscr{B}_i(\alpha)\big\}$, where $e_1,\ldots,e_n$ are independent Rademacher random variables. Recall that $\mathbb{C}_1 = \{\boldsymbol{\delta} : \|\boldsymbol{\delta}_{\mathscr{S}^c}\|_1 \le 3\|\boldsymbol{\delta}_{\mathscr{S}}\|_1\}$ where $\mathscr{S} = \mathrm{supp}(\boldsymbol{\beta}^*)$. For all $\boldsymbol{\beta} \in \boldsymbol{\beta}^* + \mathbb{B}(r)\cap\mathbb{C}_1$, we have $\|\boldsymbol{\beta}-\boldsymbol{\beta}^*\|_1 \le 4\|(\boldsymbol{\beta}-\boldsymbol{\beta}^*)_{\mathscr{S}}\|_1 \le 4s^{1/2}\|\boldsymbol{\beta}-\boldsymbol{\beta}^*\|_2$. Since $\mathscr{B}_i(\alpha)$ is $\gamma/(2r)$-Lipschitz, applying the Talagrand's contraction principle (Ledoux and Talagrand, 1991) and Holder's inequality, we have

$$\begin{aligned}\mathbb{E}(\Delta) &\le \frac{\gamma}{r}\mathbb{E}\left\{\sup_{\boldsymbol{\beta}\in\boldsymbol{\beta}^*+\mathbb{B}(r)\cap\mathbb{C}_1}\frac{1}{n}\sum_{i=1}^n\left\langle e_i\boldsymbol{x}_i, \frac{\boldsymbol{\beta}-\boldsymbol{\beta}^*}{\|\boldsymbol{\beta}-\boldsymbol{\beta}^*\|_2}\right\rangle\right\} \\ &\le \frac{\gamma}{rn}4s^{1/2}\mathbb{E}\left\|\sum_{i=1}^n e_i\boldsymbol{x}_i\right\|_\infty. \end{aligned} \tag{B.22}$$

Let $S_j = \sum_{i=1}^n e_i x_{ij}$ for $j = 1,\ldots,d$. It remains to bound $\mathbb{E}\|\sum_{i=1}^n e_i\boldsymbol{x}_i\|_\infty = \mathbb{E}(\max_j |S_j|)$. Since $\boldsymbol{x}$ is sub-exponential, by Condition 2, we have $\mathbb{P}(|x_{ij}| \ge \nu_0\sigma_{jj}^{1/2}t) \le e^{-t}$. Consequently, we obtain

$$\mathbb{E}(|e_i x_{ij}|^k) \le \int_0^\infty \mathbb{P}(|x_{ij}|^k \ge t)dt \le k!\nu_0^k\sigma_{jj}^{k/2} \quad \text{for all} \quad k \ge 2.$$

Along with the fact that $\mathbb{E}(e_i x_{ij}) = 0$, for any $0 \le \lambda \le (\nu_0\sigma_{\boldsymbol{x}})^{-1}$, the moment generating function of $e_i x_{ij}$ can be upper bounded by

$$\begin{aligned}\mathbb{E}\left(e^{\lambda e_i x_{ij}}\right) &\le 1 + \sum_{k\ge 2}\frac{\lambda^k}{k!}\mathbb{E}|e_i x_{ij}|^k \\ &\le 1 + \sum_{k\ge 2}(\nu_0\sigma_{jj}^{1/2}\lambda)^k \\ &\le 1 + \frac{\nu_0^2\sigma_{\boldsymbol{x}}^2\lambda^2}{1-\nu_0\sigma_{\boldsymbol{x}}\lambda}\end{aligned}$$

Using the inequality $\log(1+x) \le x$ for all $x > 0$, we have

$$\log\{\mathbb{E}(e^{\lambda S_j})\} \le \sum_{i=1}^n \log\{1 + \nu_0^2\sigma_{\boldsymbol{x}}^2\lambda^2/(1-\nu_0\sigma_{\boldsymbol{x}}\lambda)\} \le (2n\nu_0^2\sigma_{\boldsymbol{x}}^2\lambda^2)/\{2(1-\nu_0\sigma_{\boldsymbol{x}}\lambda)\}$$

for any $1 \le j \le d$ and $0 \le \lambda \le (v_0\sigma_{\boldsymbol{x}})^{-1}$. Consequently $S_1, ..., S_d$ are sub-gamma $\Gamma_+(v,c)$ with $v = 2nv_0^2\sigma_{\boldsymbol{x}}^2$ and $c = v_0\sigma_{\boldsymbol{x}}$. Applying Corollary 2.6 in Boucheron, Lugosi, and Massart (2013), we obtain

$$\mathbb{E}\left\|\sum_{i=1}^{n} e_i \boldsymbol{x}_i\right\|_\infty = \mathbb{E}\left(\max_j |S_j|\right) \le \sqrt{2v\log 2d} + c\log 2d = v_0\sigma_{\boldsymbol{x}}\left(2\sqrt{n\log 2d} + \log 2d\right) \quad \text{(B.23)}$$

Combining (B.21), (B.22), and (B.23), we obtain

$$\Delta \le 5s^{1/2}\frac{\gamma v_0\sigma_{\boldsymbol{x}}}{r}\left(2\sqrt{\frac{\log 2d}{n}} + \frac{\log 2d}{n}\right) + \lambda_u A_1^2\sqrt{\frac{2t}{n}} + \frac{\gamma^2}{3r^2}\frac{t}{n},$$

with probability at least $1 - e^{-t}$. Provided that $n \gtrsim (\sigma_{\boldsymbol{x}} v_0\gamma/r)^2 s(\log d + t)$, we have $\Delta \le 1/8$ with probability at least $1 - e^{-t}$. Putting all pieces together, as long as $\gamma \ge 4\sqrt{2}\lambda_u \max\{\sigma_\varepsilon, 2A_1^2 r\}$ and $n \gtrsim (\sigma_{\boldsymbol{x}} v_0\gamma/r)^2(s\log d + t)$, the following bound holds uniformly over $\boldsymbol{\beta} \in \boldsymbol{\beta}^* + \mathbb{B}(r) \cap \mathbb{C}_1$:

$$\mathscr{B}(\boldsymbol{\beta}, \boldsymbol{\beta}^*) \ge \frac{1}{2}\kappa_1 \underline{\tau}\|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_2^2,$$

with probability at least $1 - e^{-t}$. The final result is obtained by replacing $4s^{1/2}$ by L. □

B.4 Proof of Propositions

B.4.1 Proof of Proposition B.2.1

Proof. We start by obtaining an upper bound for $\hat{\boldsymbol{\beta}}^{(1)}$ obtained by solving (2.1), or equivalently, solving (B.2), with an initial estimator $\hat{\boldsymbol{\beta}}^{(0)} = \mathbf{0}$ and $\boldsymbol{\lambda}^{(0)} = p'_\lambda(\mathbf{0}) = (\lambda, \ldots, \lambda)^{\mathrm{T}}$. Conditioned on the event $\mathscr{E}_{\mathrm{rsc}}(r, L_0, \kappa) \cap \mathscr{E}_{\mathrm{score}}(\lambda)$ with $L_0 = 6s^{1/2}$, and from the proof of Lemma B.1.4 with parameters $r, \kappa, \lambda > 0$ such that $r > 2.5\kappa^{-1}s^{1/2}\lambda$, any solution $\hat{\boldsymbol{\beta}}^{(1)}$ to (B.2) satisfies

$$\|\hat{\boldsymbol{\beta}}^{(1)} - \boldsymbol{\beta}^*\|_2 \le 2.5\kappa^{-1}s^{1/2}\lambda. \quad \text{(B.24)}$$

We now continue to establish an upper bound on the estimation error for the subsequent

estimators $\hat{\boldsymbol{\beta}}^{(t)}$ for $t \geq 2$. For $t = 1, 2, \ldots$, we first construct a series of augmented sets

$$\mathscr{A}_t = \mathscr{S} \cup \big\{1 \leq j \leq d : \boldsymbol{\lambda}_j^{(t-1)} < p_0'(a_0)\lambda\big\}.$$

Let $c > 0$ be a constant such that $0.5p_0'(a_0)(c^2+1)^{1/2} + 2 = c\kappa a_0$. In the following, using mathematical induction, we will show that the cardinality of $\mathscr{A}_t$ can be upper bounded as

$$|\mathscr{A}_t| \leq (c^2+1)s. \tag{B.25}$$

For $t = 1$, the inequality holds trivially, i.e., $|\mathscr{A}_1| = |\mathscr{S}| = s \leq (c^2+1)s$. Now, assume that (B.25) holds for some integer $t \geq 2$. We aim to show that $|\mathscr{A}_{t+1}| \leq (c^2+1)s$. To this end, we first obtain an upper bound of the cardinality of the set $\mathscr{A}_{t+1} \setminus \mathscr{S}$. Since $p_\lambda'(\cdot)$ is monotonically decreasing on $\mathbb{R}^+$, by the definition of $\mathscr{A}_{t+1}$, for each $j \in \mathscr{A}_{t+1} \setminus \mathscr{S}$, we have $p_\lambda'(|\hat{\boldsymbol{\beta}}_j^{(t)}|) = \boldsymbol{\lambda}_j^{(t)} \leq p_0'(a_0)\lambda = p_\lambda'(a_0\lambda)$, which implies $|\hat{\boldsymbol{\beta}}_j^{(t)}| \geq a_0\lambda$. Moreover, the monotonicity of $p_\lambda'(\cdot)$ on $\mathbb{R}^+$ and the definition of $\mathscr{A}_t$ imply that $\|\boldsymbol{\lambda}^{(t-1)}\|_\infty = \|p_\lambda'(|\hat{\boldsymbol{\beta}}^{(t-1)}|)\|_\infty \leq \|p_\lambda'(\mathbf{0})\|_\infty = \lambda$ and $\big\|\boldsymbol{\lambda}_{\mathscr{A}_t^{\mathrm{c}}}^{(t-1)}\big\|_{\min} \geq p_0'(a_0)\lambda$, respectively.

Conditioned on the event $\mathscr{E}_{\mathrm{rsc}}(r, L, \kappa) \cap \{p_0'(a_0)\lambda \geq 2\|\nabla\mathscr{R}_n(\boldsymbol{\beta}^*) - \nabla\mathscr{R}(\boldsymbol{\beta}^*)\|_\infty\}$ with $L = \{2 + 2/p_0'(a_0)\}(c^2+1)^{1/2}s^{1/2} + 2s^{1/2}/p_0'(a_0)$, it follows from the proof of Lemma B.1.4 that

$$\begin{aligned}
\|\hat{\boldsymbol{\beta}}^{(t)} - \boldsymbol{\beta}^*\|_2 &\leq \frac{\|\boldsymbol{\lambda}_{\mathscr{S}}^{(t-1)}\|_2 + \|\boldsymbol{w}_{\mathscr{A}_t}^*\|_2 + \|\nabla\mathscr{R}(\boldsymbol{\beta}^*)\|_2}{\kappa} \qquad \text{(B.26)} \\
&\leq \frac{\{0.5p_0'(a_0)(c^2+1)^{1/2} + 2\}s^{1/2}\lambda}{\kappa} \\
&= ca_0 s^{1/2}\lambda = r^{\mathrm{crude}} < r. \qquad \text{(B.27)}
\end{aligned}$$

Along with the fact that $\boldsymbol{\beta}_j^* = 0$ for all $j \in \mathscr{A}_{t+1} \setminus \mathscr{S}$, we obtain

$$
\begin{aligned}
|\mathscr{A}_{t+1} \setminus \mathscr{S}|^{1/2} = \|\mathbf{1}_{\mathscr{A}_{t+1}\setminus\mathscr{S}}\|_2 &\le \left\| \Big(\frac{\hat{\boldsymbol{\beta}}^{(t)}}{a_0 \boldsymbol{\lambda}} \Big)_{\mathscr{A}_{t+1}\setminus\mathscr{S}} \right\|_2 \\
&\le \frac{1}{a_0 \lambda} \big\| (\hat{\boldsymbol{\beta}}^{(t)} - \boldsymbol{\beta}^*)_{\mathscr{A}_{t+1}\setminus\mathscr{S}} \big\|_2 \\
&\le c s^{1/2},
\end{aligned} \tag{B.28}
$$

where the last inequality holds by applying (B.27). Therefore $|\mathscr{A}_{\ell+1}| = |\mathscr{A}_{\ell+1} \setminus \mathscr{S}| + |\mathscr{S}| \le (c^2+1)s$. By induction, $|\mathscr{A}_t| \le (c^2+1)s$ holds for all $t \ge 1$. Consequently, (B.26) holds for all $t \ge 1$.

We note that the upper bound (B.27) is not sharp and is mainly derived for proving (B.28). We now derive a sharper upper bound for $\hat{\boldsymbol{\beta}}^{(t)}$ by controlling the terms $\|\boldsymbol{\lambda}_{\mathscr{S}}^{(t-1)}\|_2$ and $\|\boldsymbol{w}^*_{\mathscr{A}_t}\|_2$ more carefully. We start with providing a tighter upper bound for $\|\boldsymbol{\lambda}_{\mathscr{S}}^{(t-1)}\|_2$. For each $j \in \mathscr{S}$, we consider the following two cases: (i) if $|\hat{\boldsymbol{\beta}}_j^{(t-1)} - \boldsymbol{\beta}_j^*| \ge a_0\lambda$, then the inequality $\lambda_j^{(\ell-1)} \le \lambda \le a_0^{-1}|\hat{\boldsymbol{\beta}}_j^{(t-1)} - \boldsymbol{\beta}_j^*|$ holds trivially; (ii) if $|\hat{\boldsymbol{\beta}}_j^{(t-1)} - \boldsymbol{\beta}_j^*| < a_0\lambda$, then along with minimal signal strength condition $\|\boldsymbol{\beta}^*_{\mathscr{S}}\|_{\min} \ge a_0\lambda$ and the monotonicity of $p'_\lambda(\cdot)$ on $\mathbb{R}^+$, we have $0 \le |\boldsymbol{\beta}_j^*| - a_0\lambda \le |\hat{\boldsymbol{\beta}}_j^{(t-1)}|$, thus $\lambda_j^{(t-1)} = p'_\lambda(|\hat{\boldsymbol{\beta}}_j^{(t-1)}|) \le p'_\lambda\{(|\boldsymbol{\beta}_j^*| - a_0\lambda)_+\}$. Combining the two cases above, we obtain

$$
\|\boldsymbol{\lambda}_{\mathscr{S}}^{(t-1)}\|_2 \le \|p'_\lambda\{(|\boldsymbol{\beta}_j^*| - a_0\lambda)_+\}\|_2 + a_0^{-1}\|(\hat{\boldsymbol{\beta}}^{(t-1)} - \boldsymbol{\beta}^*)_{\mathscr{S}}\|_2 \tag{B.29}
$$

We now obtain an upper bound for $\|\boldsymbol{w}^*_{\mathscr{A}_t}\|_2$. Since $\mathscr{A}_t = \mathscr{S} \cup (\mathscr{A}_t \setminus \mathscr{S})$, we have

$$\begin{aligned}
\|\boldsymbol{w}^*_{\mathscr{A}_\ell}\|_2 &= \|\boldsymbol{w}^*_{\mathscr{S}}\|_2 + \|\boldsymbol{w}^*_{\mathscr{A}_t \setminus \mathscr{S}}\|_2 \\
&\leq \|\boldsymbol{w}^*_{\mathscr{S}}\|_2 + |\mathscr{A}_t \setminus \mathscr{S}|^{1/2} \|\boldsymbol{w}^*\|_\infty \\
&\leq \|\boldsymbol{w}^*_{\mathscr{S}}\|_2 + \frac{p'_0(a_0)}{2a_0} \|(\hat{\boldsymbol{\beta}}^{(t-1)} - \boldsymbol{\beta}^*)_{\mathscr{A}_t \setminus \mathscr{S}}\|_2 \qquad \text{(B.30)} \\
&\leq \|\boldsymbol{w}^*_{\mathscr{S}}\|_2 + \frac{1}{2a_0} \|(\hat{\boldsymbol{\beta}}^{(t-1)} - \boldsymbol{\beta}^*)_{\mathscr{A}_t \setminus \mathscr{S}}\|_2, \qquad \text{(B.31)}
\end{aligned}$$

where (B.30) holds from applying (B.28), and (B.31) holds from the fact that $p'_\lambda(a_0) \leq 1$.

Putting (B.26), (B.29), and (B.31) together, and applying the inequality $\sqrt{a} + \sqrt{b/4} \leq \sqrt{5(a+b)/4}$ for $a, b \geq 0$, we obtain

$$\begin{aligned}
\|\hat{\boldsymbol{\beta}}^{(t)} - \boldsymbol{\beta}^*\|_2 &\leq \frac{\|\boldsymbol{\lambda}^{(t-1)}_{\mathscr{S}}\|_2 + \|\boldsymbol{w}^*_{\mathscr{A}_\ell}\|_2 + \|\nabla \mathscr{R}(\boldsymbol{\beta}^*)\|_2}{\kappa} \\
&\leq \frac{\|p'_\lambda\{(|\boldsymbol{\beta}^*_j| - a_0\lambda)_+\}\|_2 + \|\boldsymbol{w}^*_{\mathscr{S}}\|_2 + \|\nabla \mathscr{R}(\boldsymbol{\beta}^*)\|_2}{\kappa} + \frac{\sqrt{5}}{2a_0\kappa} \|(\hat{\boldsymbol{\beta}}^{(t-1)} - \boldsymbol{\beta}^*)_{\mathscr{A}_\ell}\|_2 \\
&\leq \frac{\|p'_\lambda\{(|\boldsymbol{\beta}^*_j| - a_0\lambda)_+\} + \|\boldsymbol{w}^*_{\mathscr{S}}\|_2 + \|\nabla \mathscr{R}(\boldsymbol{\beta}^*)\|_2}{\kappa} + \delta \|\hat{\boldsymbol{\beta}}^{(t-1)} - \boldsymbol{\beta}^*\|_2, \qquad \text{(B.32)}
\end{aligned}$$

for all $t \geq 2$. The result in (B.13) can then be obtained by applying (B.32) iteratively. □

B.5 Proof of Technical Lemmas B.1.1–B.1.4

B.5.1 Proof of Lemma B.1.1

Proof. We start with an upper bound for the term $\|\nabla \mathscr{R}(\boldsymbol{\beta}^*)\|_2 = \sup_{\boldsymbol{u} \in \mathbb{S}^{d-1}} \mathbb{E}\{L'(\varepsilon)\boldsymbol{u}^{\mathrm{T}}\boldsymbol{x}\}$. Under Condition 1 on $\ell(\cdot)$ and Condition 3 on the random noise ε, we have $\mathbb{E}(\varepsilon^2|\boldsymbol{x}) \leq \sigma_\varepsilon^2$ and $|\ell'(u) - u| \leq u^2$. Since $\mathbb{E}[w_\tau(\varepsilon)\varepsilon|\boldsymbol{x}] = 0$ and $L'(\varepsilon) = \gamma w_\tau(\varepsilon)\ell'(\varepsilon/\gamma)$, we have

$$\big|\mathbb{E}\{L'(\varepsilon)|\boldsymbol{x}\}\big| \leq \big|\gamma \mathbb{E}\big[w_\tau(\varepsilon)\{\ell'(\varepsilon/\gamma) - \varepsilon/\gamma\}|\boldsymbol{x}\big]\big| \leq \big|\gamma^{-1}\mathbb{E}\big\{w_\tau(\varepsilon)\varepsilon^2|\boldsymbol{x}\big\}\big| \leq \gamma^{-1}\bar{\tau}\sigma_\varepsilon^2.$$

Therefore,

$$\mathbb{E}\big\{L'(\varepsilon)\boldsymbol{u}^{\mathrm{T}}\boldsymbol{x}\big\} = \mathbb{E}\big[\mathbb{E}\{L'(\varepsilon)|\boldsymbol{x}\}\boldsymbol{u}^{\mathrm{T}}\boldsymbol{x}\big] \le \gamma^{-1}\bar{\tau}\sigma_{\varepsilon}^2\mathbb{E}(|\boldsymbol{u}^{\mathrm{T}}\boldsymbol{x}|) \le \gamma^{-1}\bar{\tau}\sigma_{\varepsilon}^2\|\boldsymbol{u}\|_{\boldsymbol{\Sigma}}.$$

Taking the supremum over all $\boldsymbol{u} \in \mathbb{S}^{d-1}$, we have $\|\nabla\mathscr{R}(\boldsymbol{\beta}^*)\|_2 \le \gamma^{-1}\bar{\tau}\sigma_{\varepsilon}^2\lambda_u^{1/2}$, as desired.

Next, we obtain an upper bound for the centered score $\boldsymbol{w}^* = \nabla\mathscr{R}_n(\boldsymbol{\beta}^*) - \nabla\mathscr{R}(\boldsymbol{\beta}^*) = -(1/n)\sum_{i=1}^n \big[L'(\varepsilon_i)\boldsymbol{x}_i - \mathbb{E}\{L'(\varepsilon_i)\boldsymbol{x}_i\}\big]$ using the Bernstein's inequality. We start with establishing an upper bound on the kth moment of $L'(\varepsilon_i)\boldsymbol{x}_i$. Let $\mathbf{e}_j \in \mathbb{R}^d$ be the canonical basis vector, i.e., the jth entry equals one and all other entries equal zero. Setting $\boldsymbol{u} = \mathbf{e}_j$ in Condition 2 yields $\mathbb{P}\big(|x_{ij}| \ge \nu_0\sigma_{jj}^{1/2}t\big) \le e^{-t}$. Therefore,

$$\begin{aligned}
\mathbb{E}|x_{ij}|^k &= \int_0^{\infty} ku^{k-1}\mathbb{P}(|x_{ij}| \ge u)\mathrm{d}u \\
&= \int_0^{\infty} k\nu_0^k\sigma_{jj}^{k/2}\mathbb{P}\Big(|x_{ij}| \ge \nu_0\sigma_{jj}^{1/2}t\Big)t^{k-1}\mathrm{d}t \\
&\le \nu_0^k\sigma_{jj}^{k/2}k\int_0^{\infty} t^{k-1}e^{-t}\mathrm{d}t \\
&= k!\nu_0^k\sigma_{jj}^{k/2}.
\end{aligned}$$

In addition, $|\ell'(u)| \le \min(1,|u|)$ for all $u \in \mathbb{R}$, thus $|L'(\varepsilon_i)| = |\gamma w_\tau(\varepsilon_i)\ell'(\varepsilon_i/\gamma)| \le \min\big\{\bar{\tau}\gamma, \bar{\tau}|\varepsilon_i|\big\}$. Combining the above inequalities, for all $k \ge 2$ and $1 \le j \le d$, we have

$$\begin{aligned}
\mathbb{E}|L'(\varepsilon_i)x_{ij}|^k &\le \mathbb{E}\Big\{(\bar{\tau}\gamma)^{k-2}|x_{ij}|^k \cdot \mathbb{E}(\bar{\tau}^2\varepsilon_i^2|\boldsymbol{x}_i)\Big\} \\
&\le \bar{\tau}^k\gamma^{k-2}\sigma_{\varepsilon}^2\mathbb{E}|x_{ij}|^k \\
&\le \bar{\tau}^k\gamma^{k-2}\sigma_{\varepsilon}^2\nu_0^k\sigma_{jj}^{k/2}k! \\
&\le \frac{k!}{2}(2\bar{\tau}^2\sigma_{\varepsilon}^2\nu_0^2\sigma_{\boldsymbol{x}}^2)(\nu_0\bar{\tau}\sigma_{\boldsymbol{x}}\gamma)^{k-2}.
\end{aligned}$$

By Bernstein's inequality, for every $u > 0$ and $j \in \{1, \ldots, d\}$, we obtain

$$\left| \frac{1}{n} \sum_{i=1}^{n} \left[L'(\varepsilon_i) x_{ij} - \mathbb{E}\{L'(\varepsilon_i) x_{ij}\} \right] \right| \leq \nu_0 \sigma_{\boldsymbol{x}} \bar{\tau} \left(2\sigma_\varepsilon \sqrt{\frac{u}{n}} + \gamma \frac{u}{n} \right)$$

with probability at least $1 - 2e^{-u}$. Applying the union bound yields

$$\|\nabla \mathscr{R}_n(\boldsymbol{\beta}^*) - \nabla \mathscr{R}(\boldsymbol{\beta}^*)\|_\infty \leq \nu_0 \sigma_{\boldsymbol{x}} \bar{\tau} \left(2\sigma_\varepsilon \sqrt{\frac{u}{n}} + \gamma \frac{u}{n} \right)$$

with probability at least $1 - 2de^{-u}$. We then set $u = \log d + t$ to reach

$$\|\nabla \mathscr{R}_n(\boldsymbol{\beta}^*) - \nabla \mathscr{R}(\boldsymbol{\beta}^*)\|_\infty \leq \nu_0 \sigma_{\boldsymbol{x}} \bar{\tau} \left(2\sigma_\varepsilon \sqrt{\frac{\log d + t}{n}} + \gamma \frac{\log d + t}{n} \right) \tag{B.33}$$

with probability at least $1 - 2e^{-t}$.

Finally, we now obtain an upper bound for $\|\nabla \mathscr{R}_n(\boldsymbol{\beta}^*)\|_\infty$. By the triangle inequality, we have $\|\nabla \mathscr{R}_n(\boldsymbol{\beta}^*)\|_\infty \leq \|\nabla \mathscr{R}_n(\boldsymbol{\beta}^*) - \nabla \mathscr{R}(\boldsymbol{\beta}^*)\|_\infty + \|\nabla \mathscr{R}(\boldsymbol{\beta}^*)\|_\infty$. It suffices to obtain an upper bound for $\|\nabla \mathscr{R}(\boldsymbol{\beta}^*)\|_\infty$. We have

$$\|\nabla \mathscr{R}(\boldsymbol{\beta}^*)\|_\infty = \max_j \mathbb{E}\{L'(\varepsilon_i) x_{ij}\} \leq \max_j \mathbb{E}\left[x_{ij} \mathbb{E}\{L'(\varepsilon_i) | \boldsymbol{x}_i\} \right] \leq \max_j \mathbb{E}(|x_{ij}| \gamma^{-1} \bar{\tau} \sigma_\varepsilon^2) \leq \sigma_{\boldsymbol{x}} \gamma^{-1} \bar{\tau} \sigma_\varepsilon^2.$$

Combining the above and (B.33), we have

$$\|\nabla \mathscr{R}_n(\boldsymbol{\beta}^*)\|_\infty \leq \sigma_{\boldsymbol{x}} \bar{\tau} \left(2\nu_0 \sigma_\varepsilon \sqrt{\frac{\log d + t}{n}} + \nu_0 \gamma \frac{\log d + t}{n} + \gamma^{-1} \sigma_\varepsilon^2 \right)$$

with probability at least $1 - 2e^{-t}$, as desired. □

B.5.2 Proof of Lemma B.1.2

Proof. Recall that $\boldsymbol{w}^* = \nabla \mathscr{R}_n(\boldsymbol{\beta}^*) - \nabla \mathscr{R}(\boldsymbol{\beta}^*)$. The goal is to obtain an upper bound for the oracle centered loss function $\boldsymbol{w}^*_{\mathscr{J}}$ under the ℓ_2 norm. To this end, we employ a covering argument. Specifically, for any $\varepsilon \in (0, 1)$, there exists an ε-net $\mathscr{N}_\varepsilon$ of the unit sphere in $\mathbb{R}^s$ with cardinality

$|\mathscr{N}_\varepsilon| \le (1+2/\varepsilon)^s$ such that

$$\|\boldsymbol{w}^*_{\mathscr{S}}\|_2 \le \frac{1}{1-\varepsilon} \max_{\boldsymbol{u}\in\mathscr{N}_\varepsilon} \big\langle -\boldsymbol{w}^*_{\mathscr{S}}, \boldsymbol{u}\big\rangle = \frac{1}{1-\varepsilon} \max_{\boldsymbol{u}\in\mathscr{N}_\varepsilon} \frac{1}{n}\sum_{i=1}^n \Big[L'(\varepsilon_i)\boldsymbol{x}_{i\mathscr{S}}^{\mathrm{T}}\boldsymbol{u} - \mathbb{E}\left\{L'(\varepsilon_i)\boldsymbol{x}_{i\mathscr{S}}^{\mathrm{T}}\boldsymbol{u}\right\}\Big] \quad \text{(B.34)}$$

From Condition 1 on the loss function $\ell(\cdot)$, we have $|\ell'(u)| \le \min(1,|u|)$ for all $u \in \mathbb{R}$. Thus, we have $|L'(\varepsilon_i)| = |\gamma w_\tau(\varepsilon_i)\ell'(\varepsilon_i/\gamma)| \le \min\big(\bar{\tau}\gamma, \bar{\tau}|\varepsilon_i|\big)$. Since $\boldsymbol{x}$ is sub-exponential, by Condition 2, we have $\mathbb{P}\big(|\boldsymbol{u}^{\mathrm{T}}\boldsymbol{x}| \ge \nu_0\|\boldsymbol{u}\|_{\boldsymbol{\Sigma}} \cdot t\big) \le e^{-t}$ for all $t \in \mathbb{R}$ and $\boldsymbol{u} \in \mathbb{R}^d$. Thus, for all $k \ge 2$, and by a change of variable, we obtain

$$\begin{aligned}
\mathbb{E}\left(\big|L'(\varepsilon_i)\boldsymbol{x}_{i\mathscr{S}}^{\mathrm{T}}\boldsymbol{u}\big|^k\right) &\le \mathbb{E}\Big\{(\bar{\tau}\gamma)^{k-2}|\boldsymbol{x}_{i\mathscr{S}}^{\mathrm{T}}\boldsymbol{u}|^k \mathbb{E}\big(\bar{\tau}^2\varepsilon_i^2|\boldsymbol{x}_{i\mathscr{S}}\big)\Big\} \\
&\le \bar{\tau}^k\gamma^{k-2}\sigma_\varepsilon^2 \mathbb{E}\big|\boldsymbol{x}_{i\mathscr{S}}^{\mathrm{T}}\boldsymbol{u}\big|^k \\
&\le \bar{\tau}^k\gamma^{k-2}\sigma_\varepsilon^2 \int_0^\infty kt^{k-1}\mathbb{P}\big(|\boldsymbol{x}_{i\mathscr{S}}^{\mathrm{T}}\boldsymbol{u}| \ge t\big)\mathrm{d}t \\
&\le \frac{k!}{2}\Big(2\bar{\tau}^2\sigma_\varepsilon^2\nu_0^2\|\boldsymbol{u}\|_{\mathbf{S}}^2\Big)\cdot\Big(\bar{\tau}\gamma\nu_0\|\boldsymbol{u}\|_{\mathbf{S}}\Big)^{k-2}.
\end{aligned}$$

Applying the Bernstein's inequality with $a = 2\bar{\tau}^2\sigma_\varepsilon^2\nu_0^2\|\boldsymbol{u}\|_{\mathbf{S}}^2$ and $b = \bar{\tau}\gamma\nu_0\|\boldsymbol{u}\|_{\mathbf{S}}$, along with the inequality $\|\boldsymbol{u}\|_{\mathbf{S}} \le \lambda_{\max}^{1/2}(\mathbf{S})\|\boldsymbol{u}\|_2 = \lambda_{\max}^{1/2}(\mathbf{S})$, we have for all $x > 0$,

$$\frac{1}{n}\sum_{i=1}^n \Big[L'(\varepsilon_i)\boldsymbol{x}_{i\mathscr{S}}^{\mathrm{T}}\boldsymbol{u} - \mathbb{E}\{L'(\varepsilon_i)\boldsymbol{x}_{i\mathscr{S}}^{\mathrm{T}}\boldsymbol{u}\}\Big] \le \bar{\tau}\nu_0\lambda_{\max}^{1/2}(\mathbf{S})\left(2\sigma_\varepsilon\sqrt{\frac{x}{n}} + \gamma\frac{x}{n}\right), \quad \text{(B.35)}$$

with probability at least $1-e^{-x}$. Combining (B.34) and (B.35), and applying the union bound over all vectors $\boldsymbol{u} \in \mathscr{N}_\varepsilon$, we have

$$\|\boldsymbol{w}^*_{\mathscr{S}}\|_2 \le \frac{\bar{\tau}\nu_0\lambda_{\max}^{1/2}(\mathbf{S})}{1-\varepsilon}\left(2\sigma_\varepsilon\sqrt{\frac{x}{n}} + \gamma\frac{x}{n}\right)$$

with probability at least $1-(1+2/\varepsilon)^s e^{-x}$. Selecting $\varepsilon = 1/3$ and $x = 2s+t$, we obtain

$$\|\boldsymbol{w}^*_{\mathscr{S}}\|_2 \le 3\bar{\tau}\nu_0\lambda_{\max}^{1/2}(\mathbf{S})\left(\sigma_\varepsilon\sqrt{\frac{2s+t}{n}} + \gamma\frac{2s+t}{2n}\right),$$

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

B.5.3 Proof of Lemma B.1.3

Proof. Let $\hat{\boldsymbol{\beta}}$ be any solution to (B.2). Since (B.2) is convex, there exists a subgradient $\boldsymbol{\xi} \in \partial \|\hat{\boldsymbol{\beta}}\|_1$ such that $\nabla \mathscr{R}_n(\hat{\boldsymbol{\beta}}) + \boldsymbol{\lambda} \circ \boldsymbol{\xi} = \mathbf{0}$. Thus, we have

$$
\begin{aligned}
0 &= \langle \nabla \mathscr{R}_n(\hat{\boldsymbol{\beta}}) + \boldsymbol{\lambda} \circ \boldsymbol{\xi}, \hat{\boldsymbol{\beta}} - \boldsymbol{\beta} \rangle \\
&= \langle \nabla \mathscr{R}_n(\hat{\boldsymbol{\beta}}) - \nabla \mathscr{R}_n(\boldsymbol{\beta}), \hat{\boldsymbol{\beta}} - \boldsymbol{\beta} \rangle + \langle \nabla \mathscr{R}_n(\boldsymbol{\beta}) - \nabla \mathscr{R}(\boldsymbol{\beta}), \hat{\boldsymbol{\beta}} - \boldsymbol{\beta} \rangle + \langle \nabla \mathscr{R}(\boldsymbol{\beta}), \hat{\boldsymbol{\beta}} - \boldsymbol{\beta} \rangle + \langle \boldsymbol{\lambda} \circ \boldsymbol{\xi}, \hat{\boldsymbol{\beta}} - \boldsymbol{\beta} \rangle \\
&\geq 0 + \langle \boldsymbol{w}(\boldsymbol{\beta}), \hat{\boldsymbol{\beta}} - \boldsymbol{\beta} \rangle + \langle \nabla \mathscr{R}(\boldsymbol{\beta}), \hat{\boldsymbol{\beta}} - \boldsymbol{\beta} \rangle + \langle \boldsymbol{\lambda} \circ \boldsymbol{\xi}, \hat{\boldsymbol{\beta}} - \boldsymbol{\beta} \rangle \\
&\geq -\|\boldsymbol{w}(\boldsymbol{\beta})\|_\infty \|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_1 - \|\nabla \mathscr{R}(\boldsymbol{\beta})\|_2 \|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_2 + \langle \boldsymbol{\lambda} \circ \boldsymbol{\xi}, \hat{\boldsymbol{\beta}} - \boldsymbol{\beta} \rangle
\end{aligned}
$$

Since $\boldsymbol{\beta}_{\mathscr{A}^c} = \mathbf{0}, \|\boldsymbol{\xi}\|_\infty \leq \mathbf{1}$, and $\langle \boldsymbol{\xi}, \hat{\boldsymbol{\beta}} \rangle = \|\hat{\boldsymbol{\beta}}\|_1$, we can obtain a lower bound for $\langle \boldsymbol{\lambda} \circ \boldsymbol{\xi}, \hat{\boldsymbol{\beta}} - \boldsymbol{\beta} \rangle$ as

$$
\begin{aligned}
\langle \boldsymbol{\lambda} \circ \boldsymbol{\xi}, \hat{\boldsymbol{\beta}} - \boldsymbol{\beta} \rangle &= \langle (\boldsymbol{\lambda} \circ \boldsymbol{\xi})_{\mathscr{A}^c}, \hat{\boldsymbol{\beta}}_{\mathscr{A}^c} \rangle + \langle (\boldsymbol{\lambda} \circ \boldsymbol{\xi})_{\mathscr{A}}, (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})_{\mathscr{A}} \rangle \\
&\geq \|\boldsymbol{\lambda}_{\mathscr{A}^c}\|_{\min} \|\hat{\boldsymbol{\beta}}_{\mathscr{A}^c}\|_1 - \|\boldsymbol{\lambda}_{\mathscr{A}}\|_\infty \|(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})_{\mathscr{A}}\|_1 \\
&\geq \|\boldsymbol{\lambda}_{\mathscr{A}^c}\|_{\min} \|(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})_{\mathscr{A}^c}\|_1 - \|\boldsymbol{\lambda}\|_\infty \|(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})_{\mathscr{A}}\|_1.
\end{aligned}
$$

Combining the above inequalities yields

$$
\|\boldsymbol{w}(\boldsymbol{\beta})\|_\infty \|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_1 + \|\nabla \mathscr{R}(\boldsymbol{\beta})\|_2 \|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_2 \geq \|\boldsymbol{\lambda}_{\mathscr{A}^c}\|_{\min} \|(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})_{\mathscr{A}^c}\|_1 - \|\boldsymbol{\lambda}\|_\infty \|(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})_{\mathscr{A}}\|_1.
$$

The result (B.3) can then be obtained by rearranging the terms. □

B.5.4 Proof of Lemma B.1.4

Proof. The proof is similar to that of the proof of Theorem 2.5.1. For some $r > 0$ to be specified, define an intermediate quantity $\hat{\boldsymbol{\beta}}_\eta = \eta \hat{\boldsymbol{\beta}} + (1 - \eta) \boldsymbol{\beta}^*$ where $\eta = \sup\{u \in [0,1] : (1-u)\boldsymbol{\beta}^* + u\hat{\boldsymbol{\beta}} \in \boldsymbol{\beta}^* + \mathbb{B}(r)\}$. When $\hat{\boldsymbol{\beta}} \in \boldsymbol{\beta}^* + \mathbb{B}(r)$, we have $\hat{\boldsymbol{\beta}}_\eta = \hat{\boldsymbol{\beta}}$. On the other hand, when $\hat{\boldsymbol{\beta}} \notin \boldsymbol{\beta}^* + \mathbb{B}(r)$, $\hat{\boldsymbol{\beta}}_\eta$ lies on $\boldsymbol{\beta}^* + \partial \mathbb{B}(r)$ with $\eta < 1$.

We first show that $\hat{\boldsymbol{\beta}}_\eta \in \boldsymbol{\beta}^* + \mathbb{B}(r) \cap \mathbb{C}(L)$. Since $\mathscr{R}_n(\cdot)$ is convex, by an application of Lemma C.1 in Sun, Zhou and Fan (2020), we have

$$0 \le \langle \nabla \mathscr{R}_n(\hat{\boldsymbol{\beta}}_\eta) - \nabla \mathscr{R}_n(\boldsymbol{\beta}^*), \hat{\boldsymbol{\beta}}_\eta - \boldsymbol{\beta}^* \rangle \le \eta \langle \nabla \mathscr{R}_n(\hat{\boldsymbol{\beta}}) - \nabla \mathscr{R}_n(\boldsymbol{\beta}^*), \hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^* \rangle. \tag{B.36}$$

Conditioned on the event $\{a\lambda \ge 2\|\nabla \mathscr{R}_n(\boldsymbol{\beta}^*) - \nabla \mathscr{R}(\boldsymbol{\beta}^*)\|_\infty\}$ and the assumption that $\|\boldsymbol{\lambda}\|_\infty \le \lambda$ and $\|\boldsymbol{\lambda}_{\mathscr{A}^c}\|_{\min} \ge a\lambda$, applying Lemma B.1.3, we have

$$\begin{aligned}
\|(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)_{\mathscr{A}^c}\|_1 &\le \frac{\big\{\|\boldsymbol{\lambda}\|_\infty + \|\boldsymbol{w}(\boldsymbol{\beta}^*)\|_\infty\big\}\|(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)_{\mathscr{A}}\|_1 + \|\nabla \mathscr{R}(\boldsymbol{\beta}^*)\|_2 \|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2}{\|\boldsymbol{\lambda}_{\mathscr{A}^c}\|_{\min} - \|\boldsymbol{w}(\boldsymbol{\beta}^*)\|_\infty} \\
&\le \left(1 + \frac{2}{a}\right)\|(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)_{\mathscr{A}}\|_1 + \frac{2}{a\lambda}\|\nabla \mathscr{R}(\boldsymbol{\beta}^*)\|_2 \|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2
\end{aligned}$$

By the assumption that $\lambda \ge s^{-1/2}\|\nabla \mathscr{R}(\boldsymbol{\beta}^*)\|_2$, we have

$$\begin{aligned}
\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_1 &\le (2 + 2/a)\,\|(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)_{\mathscr{A}}\|_1 + \frac{2}{a\lambda}\|\nabla \mathscr{R}(\boldsymbol{\beta}^*)\|_2 \|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2 \\
&\le (2 + 2/a)\, k^{1/2}\|(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)_{\mathscr{A}}\|_2 + \frac{2}{a\lambda}\|\nabla \mathscr{R}(\boldsymbol{\beta}^*)\|_2 \|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2 \\
&\le \Big\{(2 + 2/a)\, k^{1/2} + 2s^{1/2}/a\Big\}\,\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2.
\end{aligned}$$

The above inequality implies that $\hat{\boldsymbol{\beta}} \in \boldsymbol{\beta}^* + \mathbb{C}(L)$ with $L = (2 + 2/a)k^{1/2} + 2s^{1/2}/a$. Since $\hat{\boldsymbol{\beta}}_\eta - \boldsymbol{\beta}^* = \eta(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)$ and $\hat{\boldsymbol{\beta}}_\eta \in \boldsymbol{\beta}^* + \mathbb{B}(r)$ by construction, we have $\hat{\boldsymbol{\beta}}_\eta \in \boldsymbol{\beta}^* + \mathbb{B}(r) \cap \mathbb{C}(L)$. Consequently, conditioned on the event $\mathscr{E}_{\mathrm{rsc}}(r, L, \kappa)$, we have

$$\langle \nabla \mathscr{R}_n(\hat{\boldsymbol{\beta}}_\eta) - \nabla \mathscr{R}_n(\boldsymbol{\beta}^*), \hat{\boldsymbol{\beta}}_\eta - \boldsymbol{\beta}^* \rangle \ge \kappa \|\hat{\boldsymbol{\beta}}_\eta - \boldsymbol{\beta}^*\|_2^2. \tag{B.37}$$

Next we upper bound the right-hand side of (B.36). Let Since $\hat{\boldsymbol{\beta}}$ is a solution to (B.2),

we have

$$\begin{aligned}\langle \nabla \mathscr{R}_n(\hat{\boldsymbol{\beta}}) - \nabla \mathscr{R}_n(\boldsymbol{\beta}^*), \hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^* \rangle &= \langle \nabla \mathscr{R}_n(\hat{\boldsymbol{\beta}}) + \boldsymbol{\lambda} \circ \boldsymbol{\xi}, \hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^* \rangle - \langle \boldsymbol{\lambda} \circ \boldsymbol{\xi}, \hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^* \rangle \\ &\quad - \langle \nabla \mathscr{R}_n(\boldsymbol{\beta}^*) - \nabla \mathscr{R}(\boldsymbol{\beta}^*), \hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^* \rangle - \langle \nabla \mathscr{R}(\boldsymbol{\beta}^*), \hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^* \rangle \\ &:= \Pi_1 - \Pi_2 - \Pi_3 - \Pi_4\end{aligned}$$

We now obtain bounds for the terms $\Pi_1, \ldots, \Pi_4$. For Π_1, since $\hat{\boldsymbol{\beta}}$ is a solution to (B.2), we have $\Pi_1 \leq 0$. For Π_2, since $[d] = \mathscr{S} \cup (\mathscr{A} \setminus \mathscr{S}) \cup \mathscr{A}^{\mathrm{c}}$, $\boldsymbol{\beta}^*_{\mathscr{S}^{\mathrm{c}}} = \mathbf{0}$, $\|\boldsymbol{\xi}\|_\infty \leq 1$, and $\langle \boldsymbol{\xi}, \hat{\boldsymbol{\beta}} \rangle = \|\hat{\boldsymbol{\beta}}\|_1$, we have

$$\begin{aligned}\langle \boldsymbol{\lambda} \circ \boldsymbol{\xi}, \hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^* \rangle &= \langle (\boldsymbol{\lambda} \circ \boldsymbol{\xi})_{\mathscr{S}}, (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)_{\mathscr{S}} \rangle + \langle (\boldsymbol{\lambda} \circ \boldsymbol{\xi})_{\mathscr{A} \setminus \mathscr{S}}, \hat{\boldsymbol{\beta}}_{\mathscr{A} \setminus \mathscr{S}} \rangle + \langle (\boldsymbol{\lambda} \circ \boldsymbol{\xi})_{\mathscr{A}^{\mathrm{c}}}, \hat{\boldsymbol{\beta}}_{\mathscr{A}^{\mathrm{c}}} \rangle \\ &\geq -\|\boldsymbol{\lambda}_{\mathscr{S}}\|_2 \|(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)_{\mathscr{S}}\|_2 + \langle \boldsymbol{\lambda}_{\mathscr{A} \setminus \mathscr{S}}, |\hat{\boldsymbol{\beta}}_{\mathscr{A} \setminus \mathscr{S}}| \rangle + \langle \boldsymbol{\lambda}_{\mathscr{A}^{\mathrm{c}}}, |\hat{\boldsymbol{\beta}}_{\mathscr{A}^{\mathrm{c}}}| \rangle \\ &\geq -\|\boldsymbol{\lambda}_{\mathscr{S}}\|_2 \|(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)_{\mathscr{S}}\|_2 + 0 + \|\boldsymbol{\lambda}_{\mathscr{A}^{\mathrm{c}}}\|_{\min} \|\hat{\boldsymbol{\beta}}_{\mathscr{A}^{\mathrm{c}}}\|_1 \\ &\geq -\|\boldsymbol{\lambda}_{\mathscr{S}}\|_2 \|(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)_{\mathscr{S}}\|_2 + \|\boldsymbol{\lambda}_{\mathscr{A}^{\mathrm{c}}}\|_{\min} \|(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)_{\mathscr{A}^{\mathrm{c}}}\|_1.\end{aligned}$$

For Π_3, it can be shown that

$$\begin{aligned}\langle \nabla \mathscr{R}_n(\boldsymbol{\beta}^*) - \nabla \mathscr{R}(\boldsymbol{\beta}^*), \hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^* \rangle &= \langle \boldsymbol{w}^*_{\mathscr{A}}, \hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^* \rangle + \langle \boldsymbol{w}^*_{\mathscr{A}^{\mathrm{c}}}, \hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^* \rangle \\ &\geq -\|\boldsymbol{w}^*_{\mathscr{A}}\|_2 \|(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)_{\mathscr{A}}\|_2 - \|\boldsymbol{w}^*\|_\infty \|(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)_{\mathscr{A}^{\mathrm{c}}}\|_1.\end{aligned}$$

Finally, for Π_4, we have

$$\left\langle \nabla \mathscr{R}(\boldsymbol{\beta}^*), \hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^* \right\rangle \geq -\|\nabla \mathscr{R}(\boldsymbol{\beta}^*)\|_2 \|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2.$$

Combining all of the above inequalities with $\|\boldsymbol{\lambda}_{\mathscr{A}^c}\|_{\min} \geq a\lambda \geq 2\|\boldsymbol{w}^*\|_\infty$, we obtain,

$$\begin{aligned}\langle \nabla\mathscr{R}_n(\hat{\boldsymbol{\beta}}) - \nabla\mathscr{R}_n(\boldsymbol{\beta}^*), \hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^* \rangle &\leq \Big(-\|\boldsymbol{\lambda}_{\mathscr{A}^c}\|_{\min} + \|\boldsymbol{w}^*\|_\infty \Big) \|(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)_{\mathscr{A}^c}\|_1 \\ &\quad + \|\boldsymbol{w}^*_{\mathscr{A}}\|_2 \|(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)_{\mathscr{A}}\|_2 + \|\boldsymbol{\lambda}_{\mathscr{S}}\|_2 \|(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)_{\mathscr{S}}\|_2 \\ &\quad + \|\nabla\mathscr{R}(\boldsymbol{\beta}^*)\|_2 \|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2 \\ &\leq \Big(\|\boldsymbol{\lambda}_{\mathscr{S}}\|_2 + \|\boldsymbol{w}^*_{\mathscr{A}}\|_2 + \|\nabla\mathscr{R}(\boldsymbol{\beta}^*)\|_2 \Big) \|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2. \end{aligned} \tag{B.38}$$

Putting (B.36), (B.37), and (B.38) together, and using the fact that $\eta\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2 = \|\hat{\boldsymbol{\beta}}_\eta - \boldsymbol{\beta}^*\|_2$, we obtain

$$\kappa\|\hat{\boldsymbol{\beta}}_\eta - \boldsymbol{\beta}^*\|_2 \leq \|\boldsymbol{\lambda}_{\mathscr{S}}\|_2 + \|\boldsymbol{w}^*_{\mathscr{A}}\|_2 + \|\nabla\mathscr{R}(\boldsymbol{\beta}^*)\|_2. \tag{B.39}$$

Furthermore, under the scaling conditions, we have $\|\boldsymbol{\lambda}_{\mathscr{S}}\|_2 \leq s^{1/2}\lambda$ and $\|\boldsymbol{w}^*_{\mathscr{A}}\|_2 \leq k^{1/2}a\lambda/2$. Putting these into (B.39), we obtain $\|\hat{\boldsymbol{\beta}}_\eta - \boldsymbol{\beta}^*\|_2 \leq \kappa^{-1}\big\{(2s^{1/2} + k^{1/2}a/2)\lambda\big\} < r$. Thus, $\hat{\boldsymbol{\beta}}_\eta$ falls in the interior of $\boldsymbol{\beta}^* + \mathbb{B}(r)$, implying that $\eta = 1$ and that $\hat{\boldsymbol{\beta}}_\eta = \hat{\boldsymbol{\beta}}$. This completes the proof that $\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2 \leq \kappa^{-1}\big\{(2s^{1/2} + k^{1/2}a/2)\lambda\big\}$. □

Appendix C

Supplementary Material for Chapter 3

C.1 Derivation of Algorithm 6

Recall that the penalized-`retire` amounts to solving the general optimization problem

$$\hat{\boldsymbol{\beta}} \in \underset{\boldsymbol{\beta}\in\mathbb{R}^d}{\text{minimize}} \ \{\mathscr{R}_n(\boldsymbol{\beta}) + P(\boldsymbol{\beta})\}, \tag{C.1}$$

where $\mathscr{R}_n(\boldsymbol{\beta}) = n^{-1}\sum_{i=1}^n L_{\tau,\gamma}(y_i - \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta})$ is the empirical loss function, and $P(\boldsymbol{\beta})$ is the penalty function described in Section 3.1. Let $\nabla\mathscr{R}_n(\boldsymbol{\beta})$ be the gradient of $\mathscr{R}_n(\boldsymbol{\beta})$, we locally majorize $\mathscr{R}_n(\boldsymbol{\beta})$ at $\hat{\boldsymbol{\beta}}^{(k-1)}$ by constructing an isotropic quadratic function $G_n(\cdot)$ of the form

$$G_n(\boldsymbol{\beta}|\phi_k, \hat{\boldsymbol{\beta}}^{(k-1)}) = \mathscr{R}_n(\hat{\boldsymbol{\beta}}^{(k-1)}) + \langle \nabla\mathscr{R}_n(\hat{\boldsymbol{\beta}}^{(k-1)}), \boldsymbol{\beta} - \hat{\boldsymbol{\beta}}^{(k-1)}\rangle + \frac{\phi_k}{2}||\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}^{(k-1)}||_2^2,$$

where $\phi_k > 0$ is a quadratic parameter to be determined at the k-th iteration. Then define the k-th iterate $\hat{\boldsymbol{\beta}}^{(k)}$ as the solution to

$$\underset{\boldsymbol{\beta}\in\mathbb{R}^d}{\text{minimize}} \ G_n(\boldsymbol{\beta}|\phi_k, \hat{\boldsymbol{\beta}}^{(k-1)}) + P(\boldsymbol{\beta}). \tag{C.2}$$

By the principal of the LAMM algorithm, solving the penalized-`retire` (C.1) amounts to solving (C.2) iteratively. Starting from $\hat{\boldsymbol{\beta}}^{(0)} = \mathbf{0}$, the first-order optimization condition at the

k-th iteration for (C.2) implies

$$\mathbf{0} \in \nabla \mathscr{R}_n(\hat{\boldsymbol{\beta}}^{(k-1)}) + \phi_k(\hat{\boldsymbol{\beta}}^{(k)} - \hat{\boldsymbol{\beta}}^{(k-1)}) + \partial P(\boldsymbol{\beta})\big|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}^{(k)}}, \tag{C.3}$$

where ∂P denotes the subdifferential of $P : \mathbb{R}^d \to [0,\infty)$. Moreover, let $S(a,b) = \text{sign}(a) \cdot (|a| - b)_+$ be the shrinkage operator, $\text{sign}(\cdot)$ be the sign function and $(c)_+ = \max(c,0)$. Furthermore, let $\nabla_{\boldsymbol{\beta}_g} \mathscr{R}_n(\cdot)$ be the sub-vector of the gradient $\nabla \mathscr{R}_n(\cdot)$ indexed by the g-th group. Below we derive the explicit update rules for all penalty functions described in Section 3.1.

1. Weighted lasso (Tibshirani, 1996): $P(\boldsymbol{\beta}) = \sum_{j=1}^d \lambda_j |\beta_j|$, where $\lambda_j \geq 0$ for $j = 1,\ldots,d$.

   For notational convenience, let $y = \hat{\beta}_j^{(k-1)} - \phi_k^{-1} \nabla_{\beta_j} \mathscr{R}_n(\hat{\boldsymbol{\beta}}^{(k-1)})$ be a quantity determined by the $(k-1)$-th iteration. It can be checked that the subdifferential ∂P satisfies

   $$\frac{\partial P(\boldsymbol{\beta})}{\partial \beta_j}\bigg|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}^{(k)}} = \lambda_j z, \text{ where } z = \begin{cases} \text{sign}(\hat{\beta}_j^{(k)}) & \text{if } \hat{\beta}_j^{(k)} \neq 0, \\ [-1,1] & \text{if } \hat{\beta}_j^{(k)} = 0. \end{cases}$$

   When $\hat{\beta}_j^{(k)} \neq 0$, rearranging the first-order condition (C.3) along with the fact that $u = \text{sign}(u) \cdot |u|$ for arbitrary $u \in \mathbb{R}$, we have

   $$\text{sign}(\hat{\beta}_j^{(k)}) \cdot (|\hat{\beta}_j^{(k)}| + \phi_k^{-1}\lambda_j) = \text{sign}(y) \cdot |y|.$$

   Consequently, $\text{sign}(\hat{\beta}_j^{(k)}) = \text{sign}(y)$ and $|y| = |\hat{\beta}_j^{(k)}| + \phi_k^{-1}\lambda_j > 0$. Plug-in $\text{sign}(\hat{\beta}_j^{(k)}) = \text{sign}(y)$ to obtain

   $$\begin{aligned} \hat{\beta}_j^{(k)} &= \text{sign}(y) \cdot (|y| - \phi_k^{-1}\lambda_j) \\ &= \text{sign}(y) \cdot (|y| - \phi_k^{-1}\lambda_j)_+ \\ &= S(y, \phi_k^{-1}\lambda_j), \end{aligned}$$

where the second equality comes from the fact that $|y| - \phi_k^{-1}\lambda_j = |\hat{\beta}_j^{(k)}| > 0$.

When $\hat{\beta}_j^{(k)} = 0$, rearrange the first-order condition (C.3) to yield

$$y = \phi_k^{-1}\lambda_j z.$$

Therefore, $|y| = \phi_k^{-1}\lambda_j|z| \leq \phi_k^{-1}\lambda_j$ and $\hat{\beta}_j^{(k)} = S(y, \phi_k^{-1}\lambda_j) = 0$.

Combining the two cases, $\hat{\beta}_j^{(k)}$ takes the update rule

$$\hat{\beta}_j^{(k)} \leftarrow S\{\hat{\beta}_j^{(k-1)} - \phi_k^{-1}\nabla_{\beta_j}\mathscr{R}_n(\hat{\boldsymbol{\beta}}^{(k-1)}), \phi_k^{-1}\lambda_j\}.$$

2. Elastic net (Zou and Hastie, 2005): $P(\boldsymbol{\beta}) = \lambda\alpha||\boldsymbol{\beta}||_1 + \lambda(1-\alpha)||\boldsymbol{\beta}||_2^2$, where $\lambda > 0$ and $\alpha \in (0,1)$.

   For notational convenience, let $y = \hat{\beta}_j^{(k-1)} - \phi_k^{-1}\nabla_{\beta_j}\mathscr{R}_n(\hat{\boldsymbol{\beta}}^{(k-1)})$ be a quantity determined by the $(k-1)$-th iteration. It can be checked that the subdifferential ∂P satisfies

$$\left.\frac{\partial P(\boldsymbol{\beta})}{\partial \beta_j}\right|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}^{(k)}} = \lambda\alpha z + 2\lambda(1-\alpha)\hat{\beta}_j^{(k)}, \text{ where } z = \begin{cases} \text{sign}(\hat{\beta}_j^{(k)}) & \text{if } \hat{\beta}_j^{(k)} \neq 0, \\ [-1,1] & \text{if } \hat{\beta}_j^{(k)} = 0. \end{cases}$$

   When $\hat{\beta}_j^{(k)} \neq 0$, rearranging the first-order condition (C.3) along with the fact that $u = \text{sign}(u)\cdot|u|$ for arbitrary $u \in \mathbb{R}$, we have

$$\text{sign}(\hat{\beta}_j^{(k)})\cdot\left[\{1 + 2\phi_k^{-1}\lambda(1-\alpha)\}|\hat{\beta}_j^{(k)}| + \phi_k^{-1}\lambda\alpha\right] = \text{sign}(y)\cdot|y|.$$

   Consequently, $\text{sign}(\hat{\beta}_j^{(k)}) = \text{sign}(y)$ and $|y| = \{1 + 2\phi_k^{-1}\lambda(1-\alpha)\}|\hat{\beta}_j^{(k)}| + \phi_k^{-1}\lambda\alpha > 0$.

Plug-in $\text{sign}(\hat{\beta}_j^{(k)}) = \text{sign}(y)$ to obtain

$$\begin{aligned}\{1+2\phi_k^{-1}\lambda(1-\alpha)\}\hat{\beta}_j^{(k)} &= \text{sign}(y)\cdot(|y|-\phi_k^{-1}\lambda\alpha)\\ &= \text{sign}(y)\cdot(|y|-\phi_k^{-1}\lambda\alpha)_+\\ &= S(y,\phi_k^{-1}\lambda\alpha),\end{aligned}$$

where the second equality comes from the fact that $|y|-\phi_k^{-1}\lambda\alpha = \{1+2\phi_k^{-1}\lambda(1-\alpha)\}|\hat{\beta}_j^{(k)}| > 0$. Therefore $\hat{\beta}_j^{(k)} = \{1+2\phi_k^{-1}\lambda(1-\alpha)\}^{-1}S(y,\phi_k^{-1}\lambda\alpha)$.

When $\hat{\beta}_j^{(k)} = 0$, rearrange the first-order condition (C.3) to yield

$$y = \phi_k^{-1}\lambda\alpha z.$$

Therefore, $|y| = \phi_k^{-1}\lambda\alpha|z| \le \phi_k^{-1}\lambda\alpha$, $S(y,\phi_k^{-1}\lambda\alpha) = 0$, and $\hat{\beta}_j^{(k)} = \{1+2\phi_k^{-1}\lambda(1-\alpha)\}^{-1}S(y,\phi_k^{-1}\lambda\alpha) = 0$.

Combining the two cases, $\hat{\beta}_j^{(k)}$ takes the update rule

$$\hat{\beta}_j^{(k)} \leftarrow \frac{1}{1+2\phi_k^{-1}\lambda(1-\alpha)}S\{\hat{\beta}_j^{(k-1)} - \phi_k^{-1}\nabla_{\beta_j}\mathscr{R}_n(\hat{\boldsymbol{\beta}}^{(k-1)}), \phi_k^{-1}\lambda\alpha\}.$$

3. Group lasso (Yuan and Lin, 2006): $P(\boldsymbol{\beta}) = \lambda\sum_{g=1}^{G} w_g||\boldsymbol{\beta}_g||_2$, where $\boldsymbol{\beta}_G$ is a sub-vector of $\boldsymbol{\beta}$ corresponding to the g-th group of coefficients, and $w_g > 0$.

   For notational convenience, let $\boldsymbol{y} = \hat{\boldsymbol{\beta}}_g^{(k-1)} - \phi_k^{-1}\nabla_{\boldsymbol{\beta}_g}\mathscr{R}_n(\hat{\boldsymbol{\beta}}^{(k-1)})$ be a vector determined by the $(k-1)$-th iteration. It can be checked that the subdifferential ∂P satisfies

$$\left.\frac{\partial P(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}_g}\right|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}^{(k)}} = \begin{cases}\lambda w_g \dfrac{\hat{\boldsymbol{\beta}}_g^{(k)}}{||\hat{\boldsymbol{\beta}}_g^{(k)}||_2} & \text{if } \hat{\boldsymbol{\beta}}_g^{(k)} \neq 0,\\ \lambda w_g \boldsymbol{z} & \text{if } \hat{\boldsymbol{\beta}}_g^{(k)} = 0, \text{ where } ||\boldsymbol{z}||_2 \le 1.\end{cases}$$

When $\hat{\boldsymbol{\beta}}_g^{(k)} \neq \mathbf{0}$, rearrange the first-order condition (C.3) to obtain

$$\left(1+\frac{\lambda w_g}{\phi_k||\hat{\boldsymbol{\beta}}_g^{(k)}||_2}\right)\hat{\boldsymbol{\beta}}_g^{(k)} = \boldsymbol{y}, \tag{C.4}$$

which implies (as vectors) $\hat{\boldsymbol{\beta}}_g^{(k)}$ and $\boldsymbol{y}$ have the same direction, i.e., $\hat{\boldsymbol{\beta}}_g^{(k)}/||\hat{\boldsymbol{\beta}}_g^{(k)}||_2 = \boldsymbol{y}/||\boldsymbol{y}||_2$. Plug into (C.4) to obtain $\hat{\boldsymbol{\beta}}_g^{(k)} = \big(1-\frac{\lambda w_g}{\phi_k||\boldsymbol{y}||_2}\big)\boldsymbol{y}$. The same direction statement in return implies $1-\frac{\lambda w_g}{\phi_k||\boldsymbol{y}||_2} > 0$, consequently $\hat{\boldsymbol{\beta}}_g^{(k)} = \big(1-\frac{\lambda w_g}{\phi_k||\boldsymbol{y}||_2}\big)_+ \cdot \boldsymbol{y}$.

When $\hat{\boldsymbol{\beta}}_g^{(k)} = \mathbf{0}$, rearrange the first-order condition (C.3) to yield

$$\boldsymbol{y} = \frac{\lambda w_g}{\phi_k}\boldsymbol{z}.$$

Therefore, $||\boldsymbol{y}||_2 = \frac{\lambda w_g}{\phi_k}||\boldsymbol{z}||_2 \leq \frac{\lambda w_g}{\phi_k}$. Consequently $1-\frac{\lambda w_g}{\phi_k||\boldsymbol{y}||_2} \leq 0$, and $\hat{\boldsymbol{\beta}}_g^{(k)} = \big(1-\frac{\lambda w_g}{\phi_k||\boldsymbol{y}||_2}\big)_+ \cdot \boldsymbol{y} = \mathbf{0}$.

Combining the two cases, $\hat{\beta}_j^{(k)}$ takes the update rule

$$\hat{\boldsymbol{\beta}}_g^{(k)} \leftarrow \Big\{\hat{\boldsymbol{\beta}}_g^{(k-1)} - \phi_k^{-1}\nabla_{\boldsymbol{\beta}_g}\mathscr{R}_n(\hat{\boldsymbol{\beta}}^{(k-1)})\Big\}\cdot\left(1-\frac{\lambda w_g}{\phi_k||\hat{\boldsymbol{\beta}}_g^{(k-1)} - \phi_k^{-1}\nabla_{\boldsymbol{\beta}_g}\mathscr{R}_n(\hat{\boldsymbol{\beta}}^{(k-1)})||_2}\right)_+.$$

4. Sparse group lasso (Simon et al., 2013): $P(\boldsymbol{\beta}) = \lambda||\boldsymbol{\beta}||_1 + \lambda\sum_{g=1}^G w_g||\boldsymbol{\beta}_g||_2$.

   For notational convenience, let $\boldsymbol{y} = \hat{\boldsymbol{\beta}}_g^{(k-1)} - \phi_k^{-1}\nabla_{\boldsymbol{\beta}_g}\mathscr{R}_n(\hat{\boldsymbol{\beta}}^{(k-1)})$ be a vector determined by the $(k-1)$-th iteration. It can be checked that the subdifferential ∂P satisfies

$$\left.\frac{\partial P(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}_g}\right|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}^{(k)}} = \begin{cases} \lambda\,\mathrm{sign}(\hat{\boldsymbol{\beta}}_g^{(k)}) + \lambda w_g\frac{\hat{\boldsymbol{\beta}}_g^{(k)}}{||\hat{\boldsymbol{\beta}}_g^{(k)}||_2} & \text{if } \hat{\boldsymbol{\beta}}_g^{(k)} \neq 0, \\ \lambda\boldsymbol{z}_1 + \lambda w_g\boldsymbol{z}_2 & \text{if } \hat{\boldsymbol{\beta}}_g^{(k)} = 0, \text{ where } |\boldsymbol{z}_1| \leq 1, ||\boldsymbol{z}_2||_2 \leq 1. \end{cases}$$

   When $\hat{\boldsymbol{\beta}}_g^{(k)} \neq \mathbf{0}$, rearranging the first-order condition (C.3) along with the fact that $u =$

$\text{sign}(u) \cdot |u|$ for arbitrary $u \in \mathbb{R}$, we have

$$\text{sign}(\hat{\boldsymbol{\beta}}_g^{(k)}) \cdot \left\{ \frac{\lambda}{\phi_k} + \left(1 + \frac{\lambda w_g}{\phi_k ||\hat{\boldsymbol{\beta}}_g^{(k)}||_2}\right) |\hat{\boldsymbol{\beta}}_g^{(k)}| \right\} = \text{sign}(\boldsymbol{y}) \cdot |\boldsymbol{y}|. \tag{C.5}$$

Consequently, $\text{sign}(\hat{\boldsymbol{\beta}}_g^{(k)}) = \text{sign}(\boldsymbol{y})$ and $|\boldsymbol{y}| - \phi_k^{-1}\lambda = \left(1 + \frac{\lambda w_g}{\phi_k ||\hat{\boldsymbol{\beta}}_g^{(k)}||_2}\right) |\hat{\boldsymbol{\beta}}_g^{(k)}| > 0$ (entry-wise). Plug $\text{sign}(\hat{\boldsymbol{\beta}}_g^{(k)}) = \text{sign}(\boldsymbol{y})$ into (C.5) to obtain

$$\begin{aligned}
\left(1 + \frac{\lambda w_g}{\phi_k ||\hat{\boldsymbol{\beta}}_g^{(k)}||_2}\right) \cdot \hat{\boldsymbol{\beta}}_g^{(k)} &= \text{sign}(\boldsymbol{y}) \cdot (|\boldsymbol{y}| - \phi_k^{-1}\lambda) \\
&= \text{sign}(\boldsymbol{y}) \cdot (|\boldsymbol{y}| - \phi_k^{-1}\lambda)_+ \\
&= S(\boldsymbol{y}, \phi_k^{-1}\lambda).
\end{aligned} \tag{C.6}$$

Note that (as vectors) $\hat{\boldsymbol{\beta}}_g^{(k)}$ has the same direction as $S(\boldsymbol{y}, \phi_k^{-1}\lambda)$, i.e. $\hat{\boldsymbol{\beta}}_g^{(k)} / ||\hat{\boldsymbol{\beta}}_g^{(k)}||_2 = S(\boldsymbol{y}, \phi_k^{-1}\lambda) / ||S(\boldsymbol{y}, \phi_k^{-1}\lambda)||_2$. Combine with (C.6) to obtain

$$\hat{\boldsymbol{\beta}}_g^{(k)} = \big(1 - \frac{\lambda w_g}{\phi_k ||S(\boldsymbol{y}, \phi_k^{-1}\lambda)||_2}\big) S(\boldsymbol{y}, \phi_k^{-1}\lambda).$$

The same direction statement in return implies $1 - \frac{\lambda w_g}{\phi_k ||S(\boldsymbol{y}, \phi_k^{-1}\lambda)||_2} > 0$, consequently $\hat{\boldsymbol{\beta}}_g^{(k)} = \big(1 - \frac{\lambda w_g}{\phi_k ||S(\boldsymbol{y}, \phi_k^{-1}\lambda)||_2}\big)_+ \cdot S(\boldsymbol{y}, \phi_k^{-1}\lambda)$.

When $\hat{\boldsymbol{\beta}}_g^{(k)} = \mathbf{0}$, rearrange the first-order condition (C.3) to yield

$$\frac{\lambda w_g}{\phi_k} \boldsymbol{z}_2 = \boldsymbol{y} - \phi_k^{-1}\lambda \boldsymbol{z}_1.$$

Consider arbitrary entry i in group g, it can be checked that

$$|y_i - \phi_k^{-1}\lambda z_{1,i}| \geq \max\{|y_i| - \phi_k^{-1}\lambda, 0\} \geq |S(y_i, \phi_k^{-1}\lambda)|.$$

Consequently, $\frac{\lambda w_g}{\phi_k} \geq ||\boldsymbol{y} - \phi_k^{-1}\lambda \boldsymbol{z}_1||_2 \geq ||S(\boldsymbol{y}, \phi_k^{-1}\lambda)||_2$ and $1 - \frac{\lambda w_g}{\phi_k ||S(\boldsymbol{y}, \phi_k^{-1}\lambda)||_2} \leq 0$. Therefore, $\hat{\boldsymbol{\beta}}_g^{(k)} = \big(1 - \frac{\lambda w_g}{\phi_k ||S(\boldsymbol{y}, \phi_k^{-1}\lambda)||_2}\big)_+ \cdot S(\boldsymbol{y}, \phi_k^{-1}\lambda) = \mathbf{0}$.

Combining the two cases, $\hat{\beta}_j^{(k)}$ takes the update rule

$$\hat{\boldsymbol{\beta}}_g^{(k)} \leftarrow S\Big\{\hat{\boldsymbol{\beta}}_g^{(k-1)} - \phi_k^{-1}\nabla_{\boldsymbol{\beta}_g}\mathscr{R}_n(\hat{\boldsymbol{\beta}}^{(k-1)}), \phi_k^{-1}\lambda\Big\} \cdot \left(1 - \frac{\lambda w_g}{\phi_k ||S\{\hat{\boldsymbol{\beta}}_g^{(k-1)} - \phi_k^{-1}\nabla_{\boldsymbol{\beta}_g}\mathscr{R}_n(\hat{\boldsymbol{\beta}}^{(k-1)}), \phi_k^{-1}\lambda\}||_2}\right)_+ .$$

Bibliography

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

ZHANG, T. (2010b). Analysis of multi-stage convex relaxation for sparse regularization. *J. Mach. Learn. Res.* **11** 1081–1107.

ZHANG, C.-H. and ZHANG, T. (2012). A general theory of concave regularization for high-dimensional sparse estimation problems. *Statist. Sci.* **27** 576–593.

ZOU, H. and LI, R. (2008). One-step sparse estimates in nonconcave penalized likelihood models. *Ann. Statist.* **36** 1509–1533.

PORTNOY, S. and KOENKER, R. (1997). The Gaussian hare and the Laplacian tortoise: Computability of squared-error versus absolute-error estimators. *Statist. Sci.* **12** 279–300.

NEWEY, W. K. and POWELL, J. L. (1987). Asymmetric least squares estimation and testing. *Econometrica* **55** 819–849.

EFRON, B. (1991). Regression percentiles using asymmetric squared error loss. *Statist. Sinica* **1** 93–125.

AIGNER, D.J., AMEMIYA, T. and POIRIER, D.J. (1976). On the estimation of production frontiers: maximum likelihood estimation of the parameters of a discontinuous density function. *Inter. Econ. Rev.* **17** 377–396.

ABADIE, A., ANGRIST, J. and IMBENS, G. (2002). Instrumental Variables Estimates of the Effect of Subsidized Training on the Quantiles of Trainee Earnings. *Econometrica* **70**(1): 91–117.

KOENKER, R. and BASSETT, G. (1978). Regression quantiles. *Econometrica* **46** 33–50.

JONE, M. C. (1994). Expectiles and M-quantiles are quantiles. *Statistics & Probability Letters* **20** 149–153.

HUBER, P. J. (1964). Robust estimation of a location parameter. *Ann. Math. Statist.* **35** 73–101.

LANGE, K. (1990). Convergence of image reconstruction algorithms with Gibbs smoothing. *IEEE Trans. Med. Imaging.* **9** 439–446.

NOCEDAL, J. and WRIGHT, S. J. (1999). *Numerical Optimization.* Springer, New York.

BÜHLMANN, P. and VAN DE GEER, S. (2011). *Statistics for High-Dimensional Data: Methods, Theory and Applications*. Springer, Heidelberg.

WAINWRIGHT, M. J. (2019). *High-Dimensional Statistics: A Non-Asymptotic Viewpoint.* Cambridge University Press, Cambridge.

FAN, J., LI, R., ZHANG, C.-H. and ZOU, H. (2020). *Statistical Foundations of Data Science.* CRC Press, Boca Raton.

TIBSHIRANI, R. (1996). Regression shrinkage and selection via the lasso. *J. R. Statist. Soc. B* **58** 267–288.

FAN, J. and LI, R. (2001). Variable selection via nonconcave regularized likelihood and its oracle properties. *J. Amer. Statist. Assoc.* **96** 1348–1360.

SU, W., BOGDAN, M. and CANDÉS, E. (2017). False discoveries occur early on the Lasso path. *Ann. Statist.* **45** 2133–2150.

LAHIRI, S. N. (2021). Necessary and sufficient conditions for variable selection consistency of the LASSO in high dimensions. *Ann. Statist.* **49** 820–844.

LOH, P.-L. and WAINWRIGHT, M. J. (2015). Regularized M-estimators with nonconvexity: Statistical and algorithmic theory for local optima. *J. Mach. Learn. Res.* **16** 559–616.

LOH, P. (2017). Statistical consistency and asymptotic normality for high-dimensional robust M-estimators. *Ann. Statist.* **45** 866–896.

FAN, J., XUE, L., and ZOU, H. (2014). Strong oracle optimality of folded concave regularized estimation. *Ann. Statist.* **42** 819–849.

FAN, J., LIU, H., SUN, Q. and ZHANG, T. (2018). I-LAMM for sparse learning: Simultaneous control of algorithmic complexity and statistical error. *Ann. Statist.* **46** 814–841.

PAN, X., SUN, Q. and ZHOU, W.-X. (2021). Iteratively reweighted ℓ_1-penalized robust regression. *Electron. J. Statist.* **15** 3287–3348.

BELLONI, A. and CHERNOZHUKOV, V. (2011). ℓ_1-penalized quantile regression in high-

dimensional sparse models. *Ann. Statist.* **39** 82–130.

WANG, L., WU, Y. and LI, R. (2012). Quantile regression for analyzing heterogeneity in ultra-high dimension. *J. Amer. Statist. Assoc.* **107** 214–222.

WANG, L. and HE, X. (2022). A unified analysis of global and local optima of regularized quantile regression in high dimensions: A subgradient approach. *Econom. Theory*, in press. Preprint.

GU, Y. and ZOU, H. (2016). High-dimensional generalizations of asymmetric least squares regression and their applications. *Ann. Statist.* **44** 2661–2694.

YI, C. and HUANG, J. (2017). Semismooth Newton coordinate descent algorithm for elastic-net penalized Huber loss regression and quantile regression. *J. Comp. Graph. Statist.* **26**(3): 547–557.

SUN, Q., ZHOU, W.-X. and FAN, J. (2020). Adaptive Huber regression. *J. Amer. Statist. Assoc.* **115** 254–265.

TAN, K.M., SUN, Q. and WITTEN, D. (2022). Sparse reduced rank Huber regression in high dimensions. *J. Amer. Statist. Assoc.*, in press.

FAN, J., LI, Q. and WANG, Y. (2017). Estimation of high dimensional mean regression in the absence of symmetry and light tail assumptions. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **79** 247–265.

BELLEC, P. C., LECUÉ, G., and TSYBAKOV, A. B. (2018). Slope meets Lasso: Improved oracle bounds and optimality. *Ann. Statist.* **46** 3603–3642.

HOERL, A.E. and KENNARD, R.W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics.* **12**(1) 55–67.

YAO, Q. and TONG, H. (1996). Asymmetric least squares regression estimation: a nonparametric approach. *Journal of nonparametric statistics.* **6**(2-3) 273–292.

HUNTER, D. R. and LANGE, K. (2004). A tutorial on MM algorithms. *The American Statistician.* **58**(1) 30–37.

SHERWOOD, B. and MAIDMAN, A. (2020). Package "`rqPen`", version 2.2.2. Reference manual: .

NEYMAN, J. (1959). Optimal asymptotic tests of composite hypotheses. *Probability and statsitics.* 213–234.

ZHANG, C. H. and ZHANG, S. S. (2014). Confidence intervals for low dimensional parameters in high dimensional linear models. *Journal of the Royal Statistical Society: Series B: Statistical Methodology.* 217–242.

BELLONI, A., CHERNOZHUKOV, V. and KATO, K. (2015). Uniform post-selection inference for least absolute deviation regression and other Z-estimation problems. *Biometrika.* **102**(1) 77–94.

NING, Y. and LIU, H. (2017). A general theory of hypothesis tests and confidence regions for sparse high dimensional models. *Ann. Statist.* **45**(1) 158–195.

BICKEL, P. J. (1975). One-step Huber estimates in the linear model. *J. Amer. Statist. Assoc.* **70**(350) 428–434.

REINHOLD, W., SUNSHINE, M., LIU, H., VARMA, S., KOHN, K., MORRIS, J., DOROSHOW, J. and POMMIER, Y. (2012). CellMiner: A web-based suite of genomic and pharmacologic tools to explore transcript and drug patterns in the NCI-60 cell line set. *Cancer Res.* **72** 3499–511.

SHANKAVARAM, U., VARMA, S., KANE, D., SUNSHINE, M., CHARY, K., REINHOLD, W., POMMIER, Y. and WEINSTEIN, J. (2009). CellMiner: a relational database and query tool for the NCI-60 cancel cell lines. *BMC Genomics* **10** 277.

HANSEN, K. D., IRIZARRY, R. A. and WU, Z. (2012). Removing technical variability in RNA-seq data using conditional quantile normalization. *Biostatistics* **13(2)** 204–216.

NAKATA, B., TAKASHIMA, T., OGAWA, Y., ISHIKAWA, T. and HIRAKAWA, K. (2004). Serum CYFRA 21-1 (cytokeratin-19 fragments) is a useful tumour marker for detecting disease relapse and assessing treatment efficacy in breast cancer. *British J. Cancer* **91** 873–878.

PURDOM, E. and HOLMES, S. P. (2005). Error distribution for gene expression data. *Statist. Appl. in Gen. and Mol. Biol.* **4** 16.

DU, R., LIU, B., ZHOU, L., WANG, D., HE, X., XU, X., ZHANG, L., NIU, C. and LIU, S. (2018). Downregulation of annexin A3 inhibits tumor metastasis and decreases drug resistance in breast cancer. *Cell Death & Disease* **9** 1–11.

ZHOU, T., LI, Y., YANG, L., LIU, L., JU, Y. and LI, C. (2017). Silencing of ANXA3 expression by RNA interference inhibits the proliferation and invasion of breast cancer cells. *Onc. Rep.* **37** 388–398.

FRITZMANN, J., MORKEL, M., BESSER, D., BUDCZIES, J., KOSEL, F., BREMBECK, F. H., ULRIKE, S., FICHTNER, I., SCHLAG, P. M. and BIRCHMEIER, W. (2009). A colorectal cancer expression profile that includes transforming growth factor *beta* inhibitor BAMBI predicts metastatic potential. *Gastroenterology* **137** 165–175.

WEN, J., LIN, L., LIN, B., XIA, E., QU, J. and WANG, O. (2020). Downregulation of immortalization-upregulated protein suppresses the progression of breast cancer cell lines by regulating epithelial–mesenchymal transition. *Cancer Man. Res.* **12** 8631–8642.

WATKINS, G., DOUGLAS-JONES, A., BRYCE, R., MANSEL, R. E. and JIANG, W. G. (2005). Increased levels of SPARC (osteonectin) in human breast cancer tissues and its association with clinical outcomes. *Prostag. Leuko. Essen. Fat. Acids* **72** 267–272.

NDAOUD, M. (2019). Interplay of minimax estimation and minimax support recovery under sparsity. In *Proc. Mach. Learn. Res.* **98** 647–668.

BARZILAI, J. and BORWEIN, J.M. (1988). Two point step size gradient method. *IMA J. Numer. Anal.* **8** 141–148.

BECK, A. and TEBOULLE, M. (2009). A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM journal on imaging sciences* **2**(1) 183–202.

ZOU, H. and HASTIE, T. (2005). Regularization and variable selection via the elastic net. *Journal of the royal statistical society: series B (statistical methodology)* **67**(2) 301–320.

YUAN, M. and LIN, Y. (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **68**(1) 49–67.

SIMON, N., FRIEDMAN, J., HASTIE, T. and TIBSHIRANI, R. (2013). A sparse-group lasso. *Journal of computational and graphical statistics* **22**(2) 231–245.

PUGH, C. C. (1991). *Real Mathematical Analysis,* 2nd ed. Springer-Verlag, New York.

SPOKOINY, V. (2013). Bernstein-von Mises Theorem for growing parameter dimension. arXiv:1302.3430.

TYURIN, I. S. (2011). On the convergence rate in Lyapunov's theorem. *Theory Probab. Appl.* **22** 253–270.

SPOKOINY, V. and ZHILOVA, M. (2015). Bootstrap confidence sets under model misspecification. *Ann. Statist.* **43** 2653–2675.

BOUSQUET, O. (2003). Concentration inequalities for sub-additive functions using the entropy method. *In Stochastic Inequalities and Applications. Progress in Probability* **56** 213–247. Birkhäuser, Basel.

LEDOUX, M. and TALAGRAND, M. (1991). *Probability in Banach Spaces: Isoperimetry and Processes*. Springer-Verlag, Berlin.

BOUCHERON, S., LUGOSI, G. and MASSART, P. (2013). *Concentration Inequalities: A Nonasymptotic Theory of Independence.* Oxford Univ. Press, London.

TAYLOR, J. W. (2008). Estimating value at risk and expected shortfall using expectiles. *Journal of Financial Economics* **6** 231–252.

KUAN, C, YEH, J and HSU, Y. (2009). Assessing value at risk with CARE, the conditional autoregressive expectile models. *Journal of Econometrics* **150** 261–270.

XIE, S., ZHOU, Y. and WAN, A. T. K. (2014). A varying-coefficient expectile model for estimating value at risk. *Journal of Business and Economic Statistics* **32** 576–592.

BELLINI, F. and BERNARDINO, E. D. (2017). Risk management with expectiles. *The European Journal of Finance* **23**(6): 487–506.

DAOUIA, A., GIRARD, S. and STUPFLER, G. (2018). Estimation of tail risk based on extreme expectiles. *Journal of the Royal Statistical Society: Series B* **80** 263–292.

BUSETTI, F., CAIVANO, M. and MONACHE, D. D. (2021). Domestic and global determinants of inflation: evidence from expectile regression. *Oxford Bulletin of Economics and Statistics* **83** 982–1001.

SCHNABEL, S. K. and EILERS, P. H. C. (2009). An analysis of life expectancy and economic production using expectile frontier zones. *Demographic Research* **21** 109–134.

ACERBIE, C. and TASCHE, D. (2002). On the coherence of expected shortfall. *Journal of Banking and Finance* **26** 1487–1503.

GNEITING, T. (2011). Making and evaluating point forecasts. *Journal of the American Statistical Association* **106** 746–762.

ZIEGEL, J. F. (2016). Coherence and elicitability. *Mathematical Finance* **26** 901–918.

BELLINI, F. and BIGNOZZI, V. (2015). On elicitable risk measures. *Quantitative Finance* **15** 725–733.

BELLONI, A., CHERNOZHUKOV, V., CHETVERIKOV, D. and KATO, K. (2015). Some new asymptotic theory for least squares series: Pointwise and uniform results. *Journal of Econometrics* **186** 345–366.