Electronic Journal of Statistics
Vol. 17 (2023) 2067–2119
ISSN: 1935-7524
https://doi.org/10.1214/23-EJS2147

High-dimensional composite quantile regression: Optimal statistical guarantees and fast algorithms

Haeseong Moon

Department of Mathematics, University of California, San Diego, La Jolla, CA 92093, USA
e-mail: h5moon@ucsd.edu

Wen-Xin Zhou

Department of Information and Decision Sciences, University of Illinois at Chicago, Chicago, IL 60607, USA
e-mail: wenxinz@uic.edu

Abstract: The composite quantile regression (CQR) was introduced by Zou and Yuan [*Ann. Statist.* **36** (2008) 1108–1126] as a robust regression method for linear models with heavy-tailed errors while achieving high efficiency. Its penalized counterpart for high-dimensional sparse models was recently studied in Gu and Zou [*IEEE Trans. Inf. Theory* **66** (2020) 7132–7154], along with a specialized optimization algorithm based on the alternating direct method of multipliers (ADMM). Compared to the various first-order algorithms for penalized least squares, ADMM-based algorithms are not well-adapted to large-scale problems. To overcome this computational hardness, in this paper we employ a convolution-smoothed technique to CQR, complemented with iteratively reweighted ℓ_1-regularization. The smoothed composite loss function is convex, twice continuously differentiable, and locally strong convex with high probability. We propose a gradient-based algorithm for penalized smoothed CQR via a variant of the majorize-minimization principal, which gains substantial computational efficiency over ADMM. Theoretically, we show that the iteratively reweighted ℓ_1-penalized smoothed CQR estimator achieves near-minimax optimal convergence rate under heavy-tailed errors without any moment constraint, and further achieves near-oracle convergence rate under a weaker minimum signal strength condition than needed in Gu and Zou (2020). Numerical studies demonstrate that the proposed method exhibits significant computational advantages without compromising statistical performance compared to two state-of-the-art methods that achieve robustness and high efficiency simultaneously.

MSC2020 subject classifications: Primary 62J07; secondary 62A01.
Keywords and phrases: Asymptotic efficiency, composite quantile regression, convolution smoothing, high-dimensional data, oracle property, sparsity.

Received April 2022.

Contents

1. Introduction

Consider a sparse linear regression model $y = \beta_0^* + \boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta}^* + \varepsilon$, where $y \in \mathbb{R}$ is the response variable, $\boldsymbol{x} = (x_1, \ldots, x_p)^{\mathrm{T}}$ is the p-vector of explanatory variables (covariates), and $\varepsilon \in \mathbb{R}$ is the observation noise. In high-dimensional settings where the number of covariates considerably exceeds the number of observations, a common practice is to impose a low-dimensional structure on $\boldsymbol{\beta}^*$, the p-vector of regression coefficients. Over the last three decades, various penalized regression methods have been developed for fitting high-dimensional models with low intrinsic dimensions, typified by the L_1-penalized least squares method, also known as the Lasso [34, 9]. We refer to the monographs [8], [18], [37] and [13] for comprehensive expositions of high-dimensional statistical methods and theory.

One of the main challenges in high-dimensional linear regression is that the maximum spurious correlation between the covariates and the realized noise can be large even when the population counterpart is small. Therefore, the penalized least squares methods are sensitive to the tails of the error distribution, or equivalently, the response distribution. The statistical properties are often derived under exponentially light-tailed error distributions [4, 8], including but not limited to Gaussian, sub-Gaussian or sub-exponential distributions. Heavy-tailedness, however, has been frequently observed in empirical data, such as the high-dimensional microarray data as well as financial and economic data. To cope with heavy-tailed error contamination in high dimensions, many robust penalized regression methods have been proposed; see, for example, [38], [40], [27], [26], [11], [1] and [32]. A common thread in these methods is the use of a robust loss function (that replaces the L_2 loss) to achieve either high breakdown point under arbitrary contamination or near-optimal error bounds under heavy-tailed errors. For the latter, [26] considered the case where ε has a symmetric distribution, including the standard Cauchy; [11] and [32] provided a concentration study for penalized Huber regression with a properly tuned cut-off parameter when ε has a bounded variance but can be skewed/asymmetric. To achieve robustness against gross outliers, we refer to [29] the most recent advance and the references therein.

In this work, we focus on heavy-tailed error contamination in a more general scenario. When the error distribution is not only heavy-tailed but also asymmetric, using a classical robust loss function, such as the L_1 loss, the Huber loss and the Tukey loss, may induce non-negligible bias. The impact of this bias can be alleviated by letting the cut-off parameter in the Huber/Tukey loss grow with the sample size, yet we still need ε to have finite variance in order to achieve (near-)optimal convergence rate, and the parameter tuning is quite delicate in practice. Although the least absolute deviation (LAD) regression requires no moment condition on ε, the relative efficiency of the LAD can be arbitrarily small when compared with the least squares [46]. To overcome the efficiency loss while being robust against heavy-tailed errors, [46] introduced the composite quantile regression (CQR), as a robust regression method, by combining quantile information across various quantile levels. The asymptotic efficiency of the CQR relative to the least squares has a universal lower bound 86.4% [21]. Theoretically, CQR requires the existence of an everywhere non-vanishing density function of ε without any moment constraint, thus allowing the infinite variance case. By complementing a composite loss function with sparsity-inducing penalties, [7] and [17] further proposed penalized composite quasi-likelihood and quantile regression estimators, respectively.

While the CQR method inherits the robustness property of quantile regression [22], it also inherits the computational hardness especially in high dimensions. Note that the L_1-penalized quantile regression can be recast as a linear program (LP) [38, 25], solvable by general-purpose optimization toolboxes. These toolboxes are convenient to use yet are only adapted to small-scale problems [2]. [42] and [16] proposed more efficient algorithms based on the alternating direction method of multipliers (ADMM). For penalized CQR, [17] proposed an ADMM-based algorithm which we will revisit in Section 4.1. The computational complexity of each ADMM update is of order $O(pnq + (p+q)^2)$, where $q \geq 1$

is the number of quantile levels used in the CQR. This can be computationally intensive when applied to large-scale datasets; see Section 4 for more detailed discussions.

To extend the capability of CQR with large-scale data, in this paper we propose a convolution-smoothed CQR (SCQR) method, complemented with iteratively reweighted L_1-penalization for fitting sparse models. Convolution smoothing turns the piecewise linear check function into a twice continuously differentiable, convex and locally strongly convex surrogate. Its success has recently been witnessed in the context of quantile regression in both statistical and computational aspects [15, 19].

Under a Lipschitz continuity condition on the density of ε and sub-Gaussian (stochastic) designs, we show that the L_1-penalized SCQR (SCQR-Lasso) estimator achieves the same rate of convergence as the Lasso estimator when ε is sub-Gaussian. We do not require the symmetry of the error distribution nor the existence of any moment, including the mean. Moreover, under a mild minimum signal strength (also known as the beta-min) condition, we show that the iteratively reweighted L_1-penalized SCQR estimator converges at a near-oracle rate $O(\sqrt{(s+\log q)/n})$. This reveals the advantage of folded-concave penalization in terms of its adaptivity to strong signals. Heuristically, the L_1 penalty applies soft-thresholding to all signals ignoring their magnitudes, thus creating a bias that is of order λ for all non-zero signals, where $\lambda > 0$ is the regularization parameter. Furthermore, we employ a variant of the local adaptive majorize-minimization (LAMM) algorithm [14] for solving weighted L_1-penalized SCQR estimator. The main idea is to construct an isotropic quadratic objective function that locally majorizes the smoothed composite quantile loss such that closed-form updates are available at each iteration. The quadratic coefficient is adaptively chosen so that the objective function is non-increasing along the iteration path. Compared to ADMM, LAMM is a simpler gradient-based algorithm that is particularly suited for large-scale problems, where the dominant computational effort is a relatively cheap matrix-vector multiplication at each step. The (local) strong convexity of the convolution smoothed loss facilitates the convergence of such a first order method.

Our work complements [17] in two aspects. The theoretical results in [17] are derived in the case of fixed designs satisfying conditions (C1) and (C2) therein. It is unclear whether these conditions hold with high probability for Gaussian or sub-Gaussian covariates. We provide a random design analysis for sub-Gaussian covariates. To achieve oracle convergence rate, our beta-min condition is weaker than that in [17] by relaxing the $\sqrt{s}$-factor, where s denotes the model sparsity. Computationally, we develop a fast algorithm for penalized CQR without sacrificing statistical efficiency by means of convolution smoothing. We believe that this paper introduces an interesting compromise between robustness, statistical performance and numerical efficiency for sparse linear regression with heavy-tailed errors.

This work is also closely related to [39], in which a new robust regression method is proposed along with a simulation-based procedure for choosing the regularization parameter. In low dimensions, we refer to [39]'s method as pairwise-LAD as it applies LAD regression to the pairwise differences of the observations, namely, $\{(y_i - y_j, \boldsymbol{x}_i - \boldsymbol{x}_j)\}_{1 \leq i \neq j \leq n}$. Although CQR and pairwise-LAD are motivated quite differently, an intriguing connection is that the asymptotic relative efficiency of pairwise-LAD is equivalent to that of CQR (compared to the least squares) when q, the number of quantile levels, goes to infinity. Computationally, [39] reformulates L_1-penalized pairwise-LAD as a linear program with $2n^2 + 2p$ variables and $O(n^2 + p)$ constraints. Due to the high computational complexity and storage cost, generic LP solvers can be extremely slow in practice. To alleviate the computational burden, [39] suggested using the resampling technique [10] that is able to reduce the effective sample size $O(n^2)$ (for pairwise differences) to $O(n)$.

The rest of the paper is organized as follows. Section 2 starts with a brief review of (penalized) composite quantile regression, followed by the proposed convolution smoothed CQR with iteratively reweighted L_1-penalization. The selection of tuning parameters is discussed in Section 2.3. Section 3 provides the statistical guarantees for penalized SCQR, including a bias analysis and rates of convergence of the solution path. In Section 4, we first revisit the ADMM-based algorithm proposed in [17], and then introduce a gradient-based LAMM algorithm for convolution smoothed CQR. Numerical comparisons of the three methods, CQR, SCQR and pairwise-LAD, are conducted in Section 5. All the proofs are placed in the appendix. The Python code for the proposed method and our implementation of the methods in [17] and [39] is available at https://github.com/hsmoonjohn/scqr.

2. Sparse composite quantile regression

2.1. Preliminaries

Suppose we observe n independent samples $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$ of a random variable $(\boldsymbol{x}, y) \in \mathbb{R}^p \times \mathbb{R}$ satisfying the linear model

$$y = \beta_0^* + \boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta}^* + \varepsilon = \beta_0^* + \sum_{j=1}^p x_j \beta_j^* + \varepsilon, \tag{2.1}$$

where β_0^* is the intercept, $\boldsymbol{\beta}^* = (\beta_1^*, \ldots, \beta_p^*)^{\mathrm{T}} \in \mathbb{R}^p$ is the p-vector of slope coefficients, and ε denotes the observation noise. Assume that ε is independent of $\boldsymbol{x}$, and has cumulative distribution function $F(\cdot)$ and probability density function $f(\cdot)$. Without loss of generality, we assume $\beta_0^* = 0$; otherwise we set $\widetilde{\varepsilon} = \beta_0^* + \varepsilon$ so that the model becomes $y = \boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta}^* + \widetilde{\varepsilon}$. Under these assumptions, the conditional τ-quantile $(0 < \tau < 1)$ of $y|\boldsymbol{x}$ is

$$F^{-1}(\tau) + \sum_{j=1}^p x_j \beta_j^*,$$

where $F^{-1}(\tau) := \inf\{u \in \mathbb{R} : F(u) \geq \tau\}$ is the τ-quantile of ε.

To robustly estimate $\boldsymbol{\beta}^*$ in model (2.1), we consider the composite quantile regression (CQR) approach proposed in [46], which delivers consistent estimates even when the error distribution has infinite variance and enjoys high efficiency otherwise [46, 21]. Given a positive integer q, let $\{\tau_k\}_{k=1}^q \subseteq (0,1)$ be an increasing sequence of quantile indexes and write $\alpha_k^* = F^{-1}(\tau_k)$. When $p < n$, the canonical CQR estimator of $\boldsymbol{\beta}^*$, denoted by $\widehat{\boldsymbol{\beta}}^{\mathrm{CQR}}$, is defined as

$$(\widehat{\alpha}_1, \ldots, \widehat{\alpha}_q, \widehat{\boldsymbol{\beta}}^{\mathrm{CQR}}) \in \underset{\substack{\boldsymbol{\alpha}=(\alpha_1,\ldots,\alpha_q)^{\mathrm{T}} \in \mathbb{R}^q, \\ \boldsymbol{\beta} \in \mathbb{R}^p}}{\operatorname{argmin}} \underbrace{\frac{1}{nq} \sum_{i=1}^n \sum_{k=1}^q \rho_{\tau_k}(y_i - \alpha_k - \boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{\beta})}_{=: \widehat{Q}(\boldsymbol{\alpha}, \boldsymbol{\beta})}, \qquad (2.2)$$

where $\rho_\tau(u) = \{\tau - I(u < 0)\}u$ is the check function. In the special case of $q = 1$, this becomes the usual quantile regression [22]. [46] established the asymptotic normality of $\widehat{\boldsymbol{\beta}}^{\mathrm{CQR}}$ when the density function $f(\cdot)$ of ε is non-vanishing at the selected quantile levels. Therefore, the root-n consistency of $\widehat{\boldsymbol{\beta}}^{\mathrm{CQR}}$ requires no moment condition on ε, thus allowing very heavy-tailed errors such as the Cauchy error.

For high-dimensional sparse models in which $\boldsymbol{\beta}^*$ is s-sparse with $s \ll n$, [17] proposed the penalized CQR estimator, defined as the global optimum to the optimization problem

$$\min_{\substack{\alpha_1,\ldots,\alpha_q \in \mathbb{R}, \\ \boldsymbol{\beta} \in \mathbb{R}^p}} \left\{ \frac{1}{nq} \sum_{i=1}^n \sum_{k=1}^q \rho_{\tau_k}(y_i - \alpha_k - \boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{\beta}) + \sum_{j=1}^p P_\lambda(|\beta_j|) \right\}, \qquad (2.3)$$

where $P_\lambda(\cdot) := \lambda^2 P(\cdot/\lambda)$ for some penalty function $P : [0,\infty) \to [0,\infty)$ and regularization parameter $\lambda > 0$. The regularizer $P(\cdot)$ is allowed to be non-convex (concave), which helps reduce the bias and leads to oracle estimators when the signals are sufficiently strong [44]. The most commonly used sparsity-inducing penalty functions are

(i) L_1 function [34]: $P(t) = t$ for $t \geq 0$.
(ii) Smoothly clipped absolute deviation (SCAD) penalty [12]: $P(0) = 0$ and $P'(t) = I(t \leq 1) + \frac{(a-t)_+}{a-1} I(t > 1)$ for $t \geq 0$ and some constant $a > 2$.
(iii) Minimax concave (MC) penalty [43]: $P(0) = 0$ and $P'(t) = (1 - t/a)_+$ for $t \geq 0$ and some constant $a \geq 1$.

Computationally, [17] employed the local linear approximation (LLA) algorithm [45] to obtain an approximate solution to the nonconvex problem (2.3), which enjoys desirable statistical properties. The LLA algorithm for the optimization problem (2.3) is iterative, starting at iteration 0 with an initial estimate $\widehat{\boldsymbol{\beta}}^0 \in \mathbb{R}^p$. At iteration $t = 1, 2, \ldots$, it combines a weighted L_1-penalty with the composite quantile loss to obtain the updated estimates $(\widehat{\boldsymbol{\alpha}}^t, \widehat{\boldsymbol{\beta}}^t)$. The procedure involves two steps.

1) Using the previous estimate $\widehat{\boldsymbol{\beta}}^{t-1} = (\widehat{\beta}_1^{t-1}, \ldots, \widehat{\beta}_p^{t-1})^{\mathrm{T}}$, compute the penalty weights

$$w_j^{t-1} = P_\lambda'(|\widehat{\beta}_j^{t-1}|) = \lambda P'(|\widehat{\beta}_j^{t-1}|/\lambda) \geq 0, \quad j = 1, \ldots, p.$$

2) Solve the convex optimization problem

$$\min_{\alpha_1,\dots,\alpha_q\in\mathbb{R},\,\boldsymbol{\beta}\in\mathbb{R}^p}\left\{\frac{1}{nq}\sum_{i=1}^{n}\sum_{k=1}^{q}\rho_{\tau_k}(y_i-\alpha_k-\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta})+\sum_{j=1}^{p}w_j^{t-1}|\beta_j|\right\} \tag{2.4}$$

to obtain $\widehat{\boldsymbol{\alpha}}^t=(\alpha_1^t,\dots,\alpha_q^t)^{\mathrm{T}}$ and $\widehat{\boldsymbol{\beta}}^t\in\mathbb{R}^p$.

Under the following “beta-min” (minimum signal strength) condition

$$\min_{1\le j\le p:\beta_j^*\neq 0}|\beta_j^*|\gtrsim\sqrt{\frac{s\log p}{n}} \tag{2.5}$$

among other regularity conditions on the non-stochastic design matrix $\mathbf{X}=(\boldsymbol{x}_1,\dots,\boldsymbol{x}_n)^{\mathrm{T}}\in\mathbb{R}^{n\times p}$, [17] showed that initialized with the L_1-penalized CQR (CQR-Lasso) estimator, the LLA algorithm converges to the oracle estimator in two iterations with high probability.

2.2. Convolution smoothed composite quantile regression

Motivated by the smoothed QR approach [15] that has attractive statistical properties and computational benefits [19, 33], we propose a penalized smoothed CQR estimator by complementing the convolution-smoothed composite quantile loss with a folded concave regularizer. We show that the proposed estimator, computed by a combination of the LLA and the iterative local adaptive majorize-minimization (LAMM) [14] algorithms, achieves oracle statistical properties under a relaxed “beta-min” condition compared to (2.5). We refer to Section 4 for a computational comparison between ADMM and LAMM.

For every $\boldsymbol{\beta}\in\mathbb{R}^p$, let $\widehat{F}(\cdot;\boldsymbol{\beta})$ be the empirical cumulative distribution function of the residuals $\{r_i(\boldsymbol{\beta}):=y_i-\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}\}_{i=1}^n$. Then, the empirical composite quantile loss in (2.2) can be written as

$$\widehat{Q}(\boldsymbol{\alpha},\boldsymbol{\beta})=\frac{1}{q}\int_{-\infty}^{\infty}\sum_{k=1}^{q}\rho_{\tau_k}(u-\alpha_k)\mathrm{d}\widehat{F}(u;\boldsymbol{\beta}),$$

where $\boldsymbol{\alpha}=(\alpha_1,\dots,\alpha_q)^{\mathrm{T}}\in\mathbb{R}^q$. Let $K:\mathbb{R}\to[0,\infty)$ be a symmetric, non-negative kernel function (a function that integrates to 1). For a given sequence of bandwidth parameters $h=h_n>0$, we smooth the loss $\widehat{Q}(\cdot,\cdot)$ by

$$\begin{aligned}\widehat{Q}_h(\boldsymbol{\alpha},\boldsymbol{\beta})&:=\frac{1}{q}\sum_{k=1}^{q}\int_{-\infty}^{\infty}\rho_{\tau_k}(u-\alpha_k)\mathrm{d}\widehat{F}_h(u,\boldsymbol{\beta}) \\ &=\frac{1}{nq}\sum_{i=1}^{n}\sum_{k=1}^{q}\int_{-\infty}^{\infty}\rho_{\tau_k}(u)K_h(u+\alpha_k-r_i(\boldsymbol{\beta}))\mathrm{d}u,\end{aligned} \tag{2.6}$$

where

$$\widehat{F}_h(u,\boldsymbol{\beta})=\frac{1}{n}\sum_{i=1}^{n}\int_{-\infty}^{u}K_h(v-r_i(\boldsymbol{\beta}))\mathrm{d}v\quad\text{and}\quad K_h(u)=\frac{1}{h}K(u/h).$$

For each $k = 1, \ldots, m$, define the convolution smoothed counterpart of $\rho_{\tau_k}(\cdot)$ as

$$\ell_{h,k}(u) = (\rho_{\tau_k} * K_h)(u) = \int_{-\infty}^{\infty} \rho_{\tau_k}(v) K_h(u - v) \mathrm{d}v, \tag{2.7}$$

where $*$ denotes the convolution operator. Consequently, the smoothed composite loss $\widehat{Q}_h(\boldsymbol{\alpha}, \boldsymbol{\beta})$ defined in (2.6) can be equivalently written as

$$\widehat{Q}_h(\boldsymbol{\alpha}, \boldsymbol{\beta}) = \frac{1}{nq} \sum_{i=1}^{n} \sum_{k=1}^{q} \ell_{h,k}(y_i - \alpha_k - \boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{\beta}).$$

Starting with an initial estimate $\widehat{\boldsymbol{\beta}}_h^0 \in \mathbb{R}^p$, we define a sequence of iteratively reweighted L_1-penalized smoothed CQR estimators

$$\big\{\widehat{\boldsymbol{\alpha}}_h^t = (\widehat{\alpha}_{h,1}^t, \ldots, \widehat{\alpha}_{h,q}^t)^{\mathrm{T}},\, \widehat{\boldsymbol{\beta}}_h^t = (\widehat{\beta}_{h,1}^t, \ldots, \widehat{\beta}_{h,p}^t)^{\mathrm{T}}\big\}_{t=1,2,\ldots}$$

as follows. At iteration $t = 1, 2, \ldots$, $(\widehat{\boldsymbol{\alpha}}_h^t, \widehat{\boldsymbol{\beta}}_h^t)$ is defined as a solution to the convex optimization problem

$$\min_{\boldsymbol{\alpha} \in \mathbb{R}^q,\, \boldsymbol{\beta} \in \mathbb{R}^p} \bigg\{ \widehat{Q}_h(\boldsymbol{\alpha}, \boldsymbol{\beta}) + \sum_{j=1}^{p} P'_\lambda(|\widehat{\beta}_{h,j}^{t-1}|) \cdot |\beta_j| \bigg\}, \tag{2.8}$$

where $\lambda > 0$ is the regularization parameter. Section 3 establishes the statistical properties of the solution path $\{(\widehat{\alpha}_h^t, \widehat{\boldsymbol{\beta}}_h^t)\}_{t \geq 1}$ in the stochastic design setting. Specifically, we assume that the random covariate vectors $\boldsymbol{x}_i$'s are *sub-Gaussian*; see Condition (A3) below. Recall that the theoretical results in [17] are derived in the case of fixed designs satisfying conditions (C1) and (C2) therein. It is natural to question whether or not these conditions hold with high probability for Gaussian or sub-Gaussian covariates. In Section 3.2, we describe a variant of the LAMM algorithm for solving (2.8), which will be compared to the ADMM algorithm for solving (2.4) in terms of statistical accuracy and computation time; see Section 5.

2.3. Selection of tuning parameters

The penalized smoothed CQR method relies primarily on two key tuning parameters, the regularization parameter λ and the bandwith h. As for the quantile indexes, we follow the suggestion in [46] and take $\tau_k = k/(q+1)$, $k = 1, \ldots, q$ with $q = 19$. The resulting estimator thus combines the strength across multiple QR estimators at levels $5\%, 10\%, \ldots, 90\%, 95\%$.

[19] and [33] demonstrated numerically that the convolution-smoothed QR estimator is rather insensitive to the choice of the bandwidth as long as it is in a reasonable range (neither too small nor too large). Motivated by [33], we set the default value of h as $\max\{0.01, \sqrt{\overline{\tau}(1-\overline{\tau})}\{\log(p)/n\}^{1/4}\}$, where $\overline{\tau} = q^{-1} \sum_{k=1}^{q} \tau_k$. The penalty level λ, on the other hand, has a more visible impact on the performance as it directly controls the sparsity of the solution. One

general approach is to use K-fold cross-validation (e.g. $K = 5$ or 10) when given a set of λ values. If model selection is of more interest than prediction, information criteria typically produce much smaller models and thus are preferable. As a variant of the high-dimensional Bayesian information criterion (BIC) for penalized QR [24], [17] considered the following BIC in the context of composite QR:

$$\mathrm{BIC}(\lambda) := \log\left(\frac{1}{q}\sum_{i=1}^{n}\sum_{k=1}^{q}\rho_{\tau_k}(y_i - \widehat{\alpha}_k(\lambda) - \boldsymbol{x}_i^{\mathrm{T}}\widehat{\boldsymbol{\beta}}(\lambda))\right) + |\widehat{\mathcal{S}}_\lambda|\frac{C_n\log(p)}{n}, \quad (2.9)$$

where $(\widehat{\boldsymbol{\alpha}}(\lambda), \widehat{\boldsymbol{\beta}}(\lambda))$ is a penalized CQR estimator with regularization parameter λ, $\widehat{\mathcal{S}}_\lambda$ is the support of $\widehat{\boldsymbol{\beta}}(\lambda)$, and C_n is a positive number depending on n. Typically C_n is chosen as a slowly growing function of n, e.g., $C_n = \log(\log n)$.

Motivated by the simulation-based method proposed by [3], we further describe a λ-tuning procedure that is computationally much cheaper than the cross-validation and BIC methods. The key is to utilize the the pivotal property of the L_1-loss [3].

As we shall see from the theoretical results in Section 3, the magnitude of λ depends in theory on $\|\boldsymbol{\omega}^*\|_\infty$, where $\boldsymbol{\omega}^* = (nq)^{-1}\sum_{i=1}^{n}\sum_{k=1}^{q}\{\bar{K}((\alpha^*_{h,k} - \varepsilon_i)/h) - \tau_k\}\boldsymbol{x}_i$, and $\bar{K}((\alpha^*_{h,k} - \varepsilon_i)/h)$ serves as a smoothed proxy of $I(\varepsilon_i \le \alpha^*_k)$. Recall that $\mathbb{P}(\varepsilon_i \le \alpha^*_k) = \tau_k$ for each $k = 1, \ldots, q$. Therefore, we consider a pivotal proxy of $\boldsymbol{w}^*$, defined as

$$\widetilde{\boldsymbol{w}}^* = \frac{1}{nq}\sum_{i=1}^{n}\sum_{k=1}^{q}\{I(u_{i,k} \le \tau_k) - \tau_k\}\boldsymbol{x}_i, \quad u_{ik} \overset{\text{i.i.d.}}{\sim} \mathrm{Unif}(0,1).$$

For some constant $c > 1$ and $\alpha \in (0,1)$, we set

$$\lambda^* = \lambda^*(c, \alpha) = c \cdot F^{-1}_{\|\widetilde{\boldsymbol{\omega}}^*\|_\infty|\mathbf{X}}(1-\alpha), \quad (2.10)$$

where $F^{-1}_{\|\widetilde{\boldsymbol{\omega}}^*\|_\infty|\mathbf{X}}(1-\alpha)$ denotes the $(1-\alpha)$-quantile of $\|\widetilde{\boldsymbol{\omega}}^*\|_\infty$ given $\mathbf{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)^{\mathrm{T}}$. We can calculate λ^* numerically with any specified precision by simulation. In particular, we choose $(c, \alpha) = (1.9, 0.05)$ for SCQR-Lasso and $(c, \alpha) = (3.1, 0.05)$ for SCQR-SCAD, and simulate the conditional distribution of $\|\widetilde{\boldsymbol{\omega}}^*\|_\infty$ given $\mathbf{X}$ based on 200 replications.

3. Statistical analysis

In this section, we establish the statistical properties of the penalized smoothed CQR estimators $\{(\widehat{\boldsymbol{\alpha}}^t_h, \widehat{\boldsymbol{\beta}}^t_h)\}_{t\ge1}$ initialized at $\widehat{\boldsymbol{\beta}}^0_h = \mathbf{0}$ and for an ordered sequence of quantile indexes $0 < \tau_1 < \cdots < \tau_q < 1$ with $q \ge 1$. To begin with, Section 3.1 provides non-asymptotic upper bounds on the smoothing bias. Throughout the section, we assume that $\boldsymbol{\beta}^* \in \mathbb{R}^p$ in model (2.1) is s-sparse, that is, its support $\mathcal{S} = \{1 \le j \le p : \beta^*_j \ne 0\}$ has cardinality s.

3.1. Smoothing bias

For any $h > 0$, define the population composite quantile loss $Q_h(\boldsymbol{\alpha}, \boldsymbol{\beta}) = \mathbb{E}\{\widehat{Q}_h(\boldsymbol{\alpha}, \boldsymbol{\beta})\}$ for $\boldsymbol{\alpha} \in \mathbb{R}^q$ and $\boldsymbol{\beta} \in \mathbb{R}^p$, and its minimizer

$$(\boldsymbol{\alpha}_h^*, \boldsymbol{\beta}_h^*) \in \underset{\boldsymbol{\alpha}\in\mathbb{R}^q,\, \boldsymbol{\beta}\in\mathbb{R}^p}{\operatorname{argmin}} Q_h(\boldsymbol{\alpha}, \boldsymbol{\beta}). \tag{3.1}$$

We first show that $Q_h : \mathbb{R}^{q+p} \to \mathbb{R}$ is convex under mild regularity conditions.

Lemma 3.1. Assume that the random covariate vector $\boldsymbol{x} \in \mathbb{R}^p$ is non-degenerate with $\bar{\Sigma} = \mathbb{E}(\bar{\boldsymbol{x}}\bar{\boldsymbol{x}}^{\mathrm{T}}) \succ \mathbf{0}$, where $\bar{\boldsymbol{x}} = (1, \boldsymbol{x}^{\mathrm{T}})^{\mathrm{T}}$. Moreover, let the kernel function $K(\cdot)$ and bandwidth $h > 0$ be such that

$$\min_{k=1,\ldots,q} \int_{-\infty}^{\infty} K(u) f(F^{-1}(\tau_k) + hu) \mathrm{d}u > 0, \tag{3.2}$$

where F and f denote the CDF and density function of ε, respectively. Then, the population smoothed composite quantile loss $Q_h : \mathbb{R}^{q+p} \to \mathbb{R}$ is convex and strictly convex at $(\boldsymbol{\alpha}^*, \boldsymbol{\beta}^*)$, where $\boldsymbol{\alpha}^* = (\alpha_1^*, \ldots, \alpha_q^*)^{\mathrm{T}}$ with $\alpha_k^* = F^{-1}(\tau_k)$.

Condition (3.2) can easily be verified if either the kernel function $K(\cdot)$ or the density function $f(\cdot)$ is positive everywhere. Without loss of much generality, we assume the former throughout this section. Intuitively, convolution smoothing induces bias which allows us to think $(\boldsymbol{\alpha}_h^*, \boldsymbol{\beta}_h^*) \neq (\boldsymbol{\alpha}^*, \boldsymbol{\beta}^*)$ for any given $h > 0$. By exploiting the independence of ε and $\boldsymbol{x}$ and a strictly positive kernel (e.g., Gaussian, Laplacian or logistic), we find that $\boldsymbol{\beta}_h^* = \boldsymbol{\beta}^*$ for any $h > 0$, and therefore the proposed smoothing mechanism only generates bias on the intercepts $\alpha_1^*, \ldots, \alpha_q^*$ that are of less interest.

To obtain explicit upper bounds on the (smoothing) bias, we impose the following regularity conditions of the density function $f(\cdot)$ of ε as well as the kernel function $K(\cdot)$.

(A1) There exist constants $\underline{f}, l_0 > 0$ such that $\min_{k=1,\ldots,q} f(F^{-1}(\tau_k)) \geq \underline{f}$ and $|f(u) - f(v)| \leq l_0|u - v|$ for all $u, v \in \mathbb{R}$.

(A2) The kernel function $K(\cdot)$ is symmetric around zero and positive everywhere. Moreover, $\kappa_k := \int_{-\infty}^{\infty} |u|^k K(u) \mathrm{d}u < \infty$ for $k \leq 2$, and $\underline{\kappa} := \min_{|u|\leq 1} K(u) > 0$.

Define the function

$$m_h(\boldsymbol{b}) = \mathbb{E}\Bigg\{\frac{1}{q}\sum_{k=1}^{q} \ell_{h,k}(\varepsilon - b_k)\Bigg\}, \quad \boldsymbol{b} = (b_1, \ldots, b_q)^{\mathrm{T}} \in \mathbb{R}^q, \tag{3.3}$$

where $\ell_{h,k} = \rho_{\tau_k} * K_h$. Under assumptions (A1) and (A2), we will show that $m_h : \mathbb{R}^q \to \mathbb{R}$ is strictly convex with a unique minimizer $\boldsymbol{b}_h = (b_{h,1}, \ldots, b_{h,q})^{\mathrm{T}}$, satisfying $\max_{1\leq k\leq q} |b_{h,k} - F^{-1}(\tau_k)| \lesssim h^2$. Consequently, the smoothed population composite quantile loss $Q_h : \mathbb{R}^{q+p} \to \mathbb{R}$ also has a unique minimizer, which is $(\boldsymbol{\alpha}_h^*, \boldsymbol{\beta}_h^*) = (\boldsymbol{b}_h, \boldsymbol{\beta}^*)$ and satisfies $\|\boldsymbol{\alpha}_h^* - \boldsymbol{\alpha}^*\|_\infty \lesssim h^2$.

Proposition 3.1. Suppose assumptions (A1) and (A2) hold and that $\bar{\Sigma}$ is positive definite. Then, the smoothed population composite quantile loss $Q_h : \mathbb{R}^{q+p} \to \mathbb{R}$ for any $h > 0$ is strictly convex and has a unique minimizer given by $(\boldsymbol{b}_h, \boldsymbol{\beta}^*)$, where $\boldsymbol{b}_h = (b_{h,1}, \dots, b_{h,q})^{\mathrm{T}} \in \mathbb{R}^q$ is the unique minimizer of the function m_h defined in (3.3). Furthermore, provided $0 < h \le \underline{f}/(2\kappa_2^{1/2} l_0)$, we have

$$|b_{h,k} - f(F^{-1}(\tau_k))| \le \frac{2\kappa_2 l_0}{f(F^{-1}(\tau_k))} h^2 \quad \text{for} \quad k = 1, \dots, q. \tag{3.4}$$

The proposition above suggests that the smoothing bias primarily affects intercept terms, but this statement only holds true under the assumption of a homoskedastic error distribution. It is important to note that the theoretical properties we discuss in the next section rely on this assumption, and therefore do not apply to data with heteroskedastic errors. We acknowledge that future work is needed to explore theoretical properties in such cases.

3.2. Oracle rate of convergence

Initialized at $\widehat{\boldsymbol{\beta}}_h^0 = \mathbf{0}$, let $\{(\widehat{\boldsymbol{\alpha}}_h^t, \widehat{\boldsymbol{\beta}}_h^t)\}_{t\ge1}$ be a sequence of penalized smoothed CQR estimators defined in (2.8). Without loss of generality, we assume $\boldsymbol{\mu} = \mathbb{E}(\boldsymbol{x}) = \mathbf{0}$; otherwise, we can rewrite model (2.1) as $y = \beta_0^\star + (\boldsymbol{x} - \boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\beta}^* + \varepsilon$ with $\beta_0^\star := \beta_0^* + \boldsymbol{\mu}^{\mathrm{T}}\boldsymbol{\beta}^*$. Hence, it suffices to work with the centered data $\{(y_i, \boldsymbol{x}_i - \boldsymbol{\mu})\}_{i=1}^n$. In addition, we assume that the random covariate $\boldsymbol{x}$ is sub-Gaussian; see condition (A3) below. For the regularizer $P_\lambda : [0, \infty) \to [0, \infty)$, we impose the following conditions that encompass the L_1, SCAD and MC penalties.

(A3) (sub-Gaussian covariates) The covariance matrix $\Sigma = (\sigma_{jk})_{1\le j,k\le p} = \mathbb{E}(\boldsymbol{x}\boldsymbol{x}^{\mathrm{T}})$ is positive definite. There exist constants $\nu_0, c_0 \ge 1$ such that $\mathbb{P}(|\bar{\boldsymbol{z}}^{\mathrm{T}}\boldsymbol{u}| \ge \nu_0\|\boldsymbol{u}\|_2 \cdot u) \le c_0 e^{-u^2/2}$ for all $\boldsymbol{u} \in \mathbb{R}^{p+1}$ and $u \ge 0$, where $\bar{\boldsymbol{z}} = \bar{\Sigma}^{-1/2}\bar{\boldsymbol{x}}$ and

$$\bar{\Sigma} = \mathbb{E}(\bar{\boldsymbol{x}}\bar{\boldsymbol{x}}^{\mathrm{T}}) = \begin{bmatrix} 1 & \mathbf{0}_{1\times p} \\ \mathbf{0}_{p\times 1} & \Sigma \end{bmatrix}.$$

For simplicity, we assume $c_0 = 1$ and write $\sigma_{\boldsymbol{x}}^2 = \max_{1\le j\le p} \sigma_{jj}$. Moreover, let $\gamma_1 \ge 1 \ge \gamma_p > 0$ be the largest and smallest eigenvalues of Σ.

(A4) $P_\lambda(u) = \lambda^2 P(u/\lambda)$ for $u \ge 0$, where the function $P : [0, \infty) \to [0, \infty)$ is non-decreasing, differentiable almost everywhere on $(0, \infty)$, and satisfies $P(0) = 0$, $0 \le P'(u) \le 1$, $\lim_{u\downarrow 0} P'(u) = 1$ and $P'(u_1) \le P'(u_2)$ for all $u_1 \ge u_2 \ge 0$.

Under condition (A4), $(\widehat{\boldsymbol{\alpha}}_h^1, \widehat{\boldsymbol{\beta}}_h^1)$ is essentially the L_1-penalized smoothed CQR estimator (SCQR-Lasso), that is,

$$(\widehat{\boldsymbol{\alpha}}_h^1, \widehat{\boldsymbol{\beta}}_h^1) \in \underset{\boldsymbol{\alpha}\in\mathbb{R}^q, \boldsymbol{\beta}\in\mathbb{R}^p}{\operatorname{argmin}} \big\{\widehat{Q}_h(\boldsymbol{\alpha}, \boldsymbol{\beta}) + \lambda\|\boldsymbol{\beta}\|_1\big\}. \tag{3.5}$$

Without smoothing, [17] obtained the convergence rates (under L_2-loss) for the L_1-penalized CQR (CQR-Lasso) estimator $(\widehat{\boldsymbol{\alpha}}, \widehat{\boldsymbol{\beta}}) \in \operatorname{argmin}_{\boldsymbol{\alpha}\in\mathbb{R}^q, \boldsymbol{\beta}\in\mathbb{R}^p}\{\widehat{Q}(\boldsymbol{\alpha}, \boldsymbol{\beta}) +$

$\lambda\|\boldsymbol{\beta}\|_1\}$ under fixed designs. For sub-Gaussian (stochastic) designs, we first establish estimation error bounds for $(\widehat{\boldsymbol{\alpha}}_h^1, \widehat{\boldsymbol{\beta}}_h^1)$, which complement the results in [17].

Recall from Proposition 3.1 that $(\boldsymbol{\alpha}_h^*, \boldsymbol{\beta}^*)$ is the unique minimizer of the population loss Q_h. Write $\boldsymbol{\alpha}_h^* = (\alpha_{h,1}^*, \ldots, \alpha_{h,q}^*)^{\mathrm{T}} \in \mathbb{R}^q$. For the smoothed loss $\widehat{Q}_h(\boldsymbol{\alpha}, \boldsymbol{\beta})$, define its partial gradient vectors at $(\boldsymbol{\alpha}_h^*, \boldsymbol{\beta}^*)$ as

$$\boldsymbol{\zeta}^* := \nabla_{\boldsymbol{\alpha}} \widehat{Q}_h(\boldsymbol{\alpha}_h^*, \boldsymbol{\beta}^*) = \frac{1}{nq} \sum_{i=1}^n \begin{bmatrix} \bar{K}((\alpha_{h,1}^* - \varepsilon_i)/h) - \tau_1 \\ \vdots \\ \bar{K}((\alpha_{h,q}^* - \varepsilon_i)/h) - \tau_q \end{bmatrix} \in \mathbb{R}^q, \tag{3.6}$$

$$\boldsymbol{\omega}^* := \nabla_{\boldsymbol{\beta}} \widehat{Q}_h(\boldsymbol{\alpha}_h^*, \boldsymbol{\beta}^*) = \frac{1}{nq} \sum_{i=1}^n \sum_{k=1}^q \{\bar{K}((\alpha_{h,k}^* - \varepsilon_i)/h) - \tau_k\} \boldsymbol{x}_i \in \mathbb{R}^p, \tag{3.7}$$

where $\bar{K}(u) = \int_{-\infty}^u K(t)\mathrm{d}t$.

The key elements of our analysis are (i) a cone-like property for $\widehat{\boldsymbol{\beta}}_h^1 - \boldsymbol{\beta}^*$, and (ii) a local restricted strong convexity (RSC) property for the empirical loss $\widehat{Q}_h$, which is based on the function

$$D(\boldsymbol{\alpha}, \boldsymbol{\beta}) := \left\langle \nabla \widehat{Q}_h(\boldsymbol{\alpha}, \boldsymbol{\beta}) - \nabla \widehat{Q}_h(\boldsymbol{\alpha}_h^*, \boldsymbol{\beta}^*), \begin{bmatrix} \boldsymbol{\alpha} - \boldsymbol{\alpha}_h^* \\ \boldsymbol{\beta} - \boldsymbol{\beta}^* \end{bmatrix} \right\rangle, \quad \boldsymbol{\alpha} \in \mathbb{R}^q, \boldsymbol{\beta} \in \mathbb{R}^p. \tag{3.8}$$

For any regularization parameter $\lambda > 0$, define the event

$$\mathcal{G}(\lambda) := \{\|\boldsymbol{\zeta}^*\|_\infty \le 3\lambda/(2q), \, \|\boldsymbol{\omega}^*\|_\infty \le \lambda/2\} \tag{3.9}$$

and the restricted (cone-like) set

$$\mathcal{C} = \mathcal{C}(\mathcal{S}) := \left\{ \begin{bmatrix} \boldsymbol{\delta} \\ \boldsymbol{\Delta} \end{bmatrix} \in \mathbb{R}^{q+p} : \|\boldsymbol{\Delta}_{\mathcal{S}^{\mathrm{c}}}\|_1 \le 3\|\boldsymbol{\Delta}_{\mathcal{S}}\|_1 + 3q^{-1/2}\|\boldsymbol{\delta}\|_2 \right\}. \tag{3.10}$$

It can be shown that $\begin{bmatrix} \widehat{\boldsymbol{\alpha}}_h^1 - \boldsymbol{\alpha}^* \\ \widehat{\boldsymbol{\beta}}_h^1 - \boldsymbol{\beta}^* \end{bmatrix} \in \mathcal{C}(\mathcal{S})$ conditioned on $\mathcal{G}(\lambda)$. Proposition 3.2 below validates that for all sufficiently large λ, the event $\mathcal{G}(\lambda)$ holds with high probability.

Proposition 3.2. *Under assumption (A3), the event $\mathcal{G}(\lambda)$ holds with probability at least $1 - 2q\exp(-9n\lambda^2/2) - 2p\exp\{-n\lambda^2/(32\nu_0^2\sigma_x^2)\}$.*

Due to high dimensionality, the empirical loss $\widehat{Q}_h : \mathbb{R}^q \times \mathbb{R}^p \to \mathbb{R}$ does not have a curvature along all directions. In fact, there exists a subspace with dimension at least $p-n$ of directions in which it is completely flat. Instead, it can be shown that the cone-like subset $\mathcal{C} = \mathcal{C}(\mathcal{S})$ is well-aligned with the curved directions of the Hessian of $\widehat{Q}_h$ in a local region with high probability, which we refer to as the local RSC property. For $r, l > 0$, define the (rescaled) ℓ_2-ball and ℓ_1-cone as

$$\mathbb{B}_\Omega(r) := \left\{ \begin{bmatrix} \boldsymbol{\delta} \\ \boldsymbol{\Delta} \end{bmatrix} \in \mathbb{R}^{q+p} : \|(\boldsymbol{\delta}^{\mathrm{T}}, \boldsymbol{\Delta}^{\mathrm{T}})^{\mathrm{T}}\|_\Omega \le r \right\} \tag{3.11}$$

$$\text{and} \quad \mathbb{C}_\Omega(l) := \left\{ \begin{bmatrix} \boldsymbol{\delta} \\ \boldsymbol{\Delta} \end{bmatrix} \in \mathbb{R}^{q+p} : \|(\boldsymbol{\delta}^{\mathrm{T}}, \boldsymbol{\Delta}^{\mathrm{T}})^{\mathrm{T}}\|_1 \le l \|(\boldsymbol{\delta}^{\mathrm{T}}, \boldsymbol{\Delta}^{\mathrm{T}})^{\mathrm{T}}\|_\Omega \right\}, \tag{3.12}$$

where

$$\Omega = \begin{bmatrix} \mathrm{I}_q & \mathbf{0}_{q\times p} \\ \mathbf{0}_{p\times q} & \Sigma \end{bmatrix} \in \mathbb{R}^{(q+p)\times(q+p)}.$$

Here, for a matrix A, $\|\cdot\|_A$ denote the vector norm induced by $A : \|u\|_A = \|A^{1/2}u\|_2$. For any curvature parameter $c > 0$, radius parameter $r > 0$, and cone parameter $l > 0$, define the event

$$\mathcal{R}(c, r, l) := \left\{ D(\boldsymbol{\alpha}, \boldsymbol{\beta}) \ge c\big(\|\boldsymbol{\Delta}\|_\Sigma^2 + q^{-1}\|\boldsymbol{\delta}\|_2^2\big) \text{ for all } \begin{bmatrix} \boldsymbol{\delta} \\ \boldsymbol{\Delta} \end{bmatrix} \in \mathbb{B}_\Omega(r) \cap \mathbb{C}_\Omega(l) \right\}, \tag{3.13}$$

where $\boldsymbol{\delta} = \boldsymbol{\alpha} - \boldsymbol{\alpha}_h^* \in \mathbb{R}^q$, $\boldsymbol{\Delta} = \boldsymbol{\beta} - \boldsymbol{\beta}^* \in \mathbb{R}^p$. The following proposition shows that under certain conditions on (s, p, n) and h, there exists some curvature parameter $c > 0$ such that event $\mathcal{R}(c, r, l)$ occurs with high probability.

Proposition 3.3. Let assumptions (A1), (A2), and (A3) hold. Furthermore, assume $\max_{k=1,\dots,q} f(F^{-1}(\tau_k)) \le \overline{f}$ for some constant $\overline{f} > 0$. Then, the event $\mathcal{R}(c, r, l)$ with $c = 0.5\underline{f} \cdot \underline{\kappa}$ holds with probability at least $1 - q/(2p)$ as long as

$$12\nu_0^2 r \le h \le \underline{f}/\{\max(4\kappa_2^{1/2} l_0, 2l_0)\} \quad \text{and} \quad n \ge C(\nu_0 \sigma_x l/\underline{f} r)^2 \overline{f} h \log(2p),$$

where $C > 0$ is a constant independent of (n, s, p, h).

Based on the above preparations, we are now ready to present the first main result of this subsection regarding the estimation error of the SCQR-Lasso estimator defined in (3.5).

Theorem 3.1. Assume $\max_{k=1,\dots,q} f(F^{-1}(\tau_k)) \le \overline{f}$ for some constant $\overline{f} > 0$. Under assumptions (A1)–(A3), the SCQR-Lasso estimator $(\widehat{\boldsymbol{\alpha}}_h, \widehat{\boldsymbol{\beta}}_h)$ with $\lambda \asymp \nu_0 \sigma_x \sqrt{\log(2p)/n}$ satisfies the following error bound

$$\left\| \begin{matrix} \frac{\widehat{\boldsymbol{\alpha}}_h - \boldsymbol{\alpha}_h^*}{\sqrt{q}} \\ \widehat{\boldsymbol{\beta}}_h - \boldsymbol{\beta}^* \end{matrix} \right\|_\Omega \le C \underline{f}^{-1} s^{1/2} \lambda \tag{3.14}$$

with probability at least $1 - q/p$, provided that the bandwidth h satisfies

$$\max\left\{ \frac{\sigma_x}{\underline{f}} \sqrt{\frac{sq \log(2p)}{n}}, \frac{\sigma_x^2 \overline{f}}{\underline{f}^2} \frac{\max(s, q) \cdot \log(2p)}{n} \right\} \lesssim h \lesssim \underline{f}/l_0,$$

where $C > 0$ is a constant independent of (n, s, p, h).

The above theorem shows that the L_1-penalized SCQR estimator achieves the same rate of convergence as the L_1-penalized quantile regression estimator [3], with a proper choice of the bandwidth parameter h, which is yet flexible. Moreover, Theorem 3.1 indicates that, the regularization parameter $\lambda \asymp$

$\nu_0\sigma_x\sqrt{\log(2p)/n}$ is independent of the error distibution, which alleviates the difficulty of tuning parameter selection of the LS estimator that typically depends on the standard deviation of the error distribution, and of the high-dimensional CQR estimator by [17] that is dependent on the minimum density of the error distribution at each quantile level.

As we can observe from Proposition 3.1 and Theorem 3.1, the bandwidth parameter h adapts to the sample size n, quantile parameter q and dimensionality p to achieve a tradeoff between statistical accuracy and computational stability. When error distribution is independent of covariates, the larger the bandwidth, the bias on the intercept terms gets larger while making the sample size requirement in Theorem 3.1 more lenient. Also, when the dimension p and the quantile parameter q get larger, the bandwidth should get larger to adapt to the requirement in Theorem 3.1, which makes the bias larger. On the other hand, when the bandwith parameter is small, the bias gets smaller. However, smaller bandwidth require more sample size to satisfy the sample size requirement in Theorem 3.1, and makes gradient term less stable computaionally dut to $1/h$ term inside it.

In [33], they have assumed sub-exponential random vector for the design matrix when deriving the error rate, which is arguably the weakest moment condition in high-dimensional regression analysis under random design. This random design plays crucial role for deriving probability bound of restrictive strong convexity. We expect the sub-Gaussian condition can be relaxed to sub-exponential condition by following the proof of Theorem 4.2 of [33]. However, to derive a prediction error bound for our estimator, sub-Gaussianity is crucial, as well as for the strong oracle property which will be discussed in the following section. For the sake of brevity, only sub-Gaussian design is considered for deriving the error bound. As a corollary, we derive a prediction error bound for $\widehat{\boldsymbol{\beta}}_h$, which is a direct consequence of Theorem 3.1.

Corollary 3.1. Under the conditions of Theorem 3.1, it holds

$$\frac{1}{\sqrt{n}}\|\boldsymbol{X}(\widehat{\boldsymbol{\beta}}_h-\boldsymbol{\beta}^*)\|_2 \lesssim \underline{f}^{-1}\nu_0\sigma_x s^{1/2}\left(\frac{\log p}{n}\right)^{1/2} \tag{3.15}$$

with probability at least $1-(q+2)/p$.

Next, we investigate the statistical properties of the iteratively reweighted L_1-penalized estimators $(\widehat{\boldsymbol{\alpha}}_h^t, \widehat{\boldsymbol{\beta}}_h^t)$ when $t \geq 2$. Define the error vectors

$$\boldsymbol{\theta}^t = \begin{bmatrix} \frac{\widehat{\boldsymbol{\alpha}}_h^t-\boldsymbol{\alpha}_h^*}{\sqrt{q}} \\ \widehat{\boldsymbol{\beta}}_h^t-\boldsymbol{\beta}^* \end{bmatrix} \in \mathbb{R}^{q+p}, \quad t=1,2,\ldots.$$

The following result characterizes the dependence of the estimation error $\|\boldsymbol{\theta}^t\|_2$ at t-th step on $\|\boldsymbol{\theta}^{t-1}\|_2$ from the previous step. It reveals how iteratively reweighted L_1-penalization refines the statistical rate when the signals are sufficiently strong. We first derive a deterministic bound of the estimation error, conditioned on some "good" events.

Theorem 3.2. Suppose assumptions (A1)–(A4) hold, and let $a_0, c > 0$ be such that

$$P'(a_0) > 0 \quad \text{and} \quad \sqrt{1 + \{P'(a_0)\}^2/2} < a_0 c \gamma_p. \tag{3.16}$$

Furthermore, let $b > 0$ satisfy

$$\sqrt{\frac{b^2+1}{2}} P'(a_0) + 2 = a_0 c \gamma_p \cdot b, \tag{3.17}$$

and define $r_{\rm opt} = a_0 b (\gamma_p s)^{1/2} \lambda$. Then, conditioned on $\mathcal{G}(P'(a_0)\lambda) \cap \mathcal{R}(c, r, l)$ with $r \geq q^{1/2} r_{\rm opt}$ and $l = 4\gamma_p^{-1/2}\sqrt{2 \cdot \max(s,q)}$, the sequence of solutions $\{(\widehat{\boldsymbol{\alpha}}_h^t, \widehat{\boldsymbol{\beta}}_h^t)\}_{t\geq 1}$ to programs (2.4) satisfies

$$\|\boldsymbol{\theta}^t\|_\Omega \leq \delta \cdot \|\boldsymbol{\theta}^{t-1}\|_\Omega + \underbrace{c^{-1}\Big\{\gamma_p^{-1/2}\|P'_\lambda((|\boldsymbol{\beta}^*_{\mathcal{S}}| - a_0\lambda)_+)\|_2 + \|\boldsymbol{\omega}^*_{\mathcal{S}}\|_2\Big\}}_{=:\, r_{\rm ora}} + c^{-1}\sqrt{q}\|\boldsymbol{\zeta}^*\|_2, \tag{3.18}$$

where $\delta := \sqrt{1 + \{P'(a_0)\}^2/2}/(a_0 c \gamma_p) \in (0,1)$. In addition, it holds for for any $t \geq 2$ that

$$\|\boldsymbol{\theta}^t\|_\Omega \leq \delta^{t-1} r_{\rm opt} + (1-\delta)^{-1}(r_{\rm ora} + c^{-1}\sqrt{q}\|\boldsymbol{\zeta}^*\|_2). \tag{3.19}$$

The above theorem shows that under proper conditions on the curvature parameter c and penalty function, the estimation error, at least its leading term, can be refined iteratively via reweighted L_1-penalization. From (3.18) we see that there are three terms on the right-hand side that cannot be improved, which are

$$\|\boldsymbol{\omega}^*_{\mathcal{S}}\|_2, \quad \|P'_\lambda((|\boldsymbol{\beta}^*_{\mathcal{S}}| - a_0\lambda)_+)\|_2 \quad \text{and} \quad \sqrt{q}\|\boldsymbol{\zeta}^*\|_2.$$

The first term, $\|\boldsymbol{\omega}^*_{\mathcal{S}}\|_2$, determines the oracle convergence rate because it corresponds to the estimation error of the oracle SCQR estimator when only the significant covariates (indexed by $\mathcal{S}$) are used in the fitting. The oracle SCQR estimator is formally defined as

$$(\widehat{\boldsymbol{\alpha}}^o, \widehat{\boldsymbol{\beta}}^o) = \underset{\substack{(\boldsymbol{\alpha},\boldsymbol{\beta}) \in \mathbb{R}^q \times \mathbb{R}^p \\ \boldsymbol{\beta}_{\mathcal{S}^c} = \mathbf{0}}}{\operatorname{argmin}} \widehat{Q}_h(\boldsymbol{\alpha}, \boldsymbol{\beta}), \tag{3.20}$$

where subscript h is ommited for the brevity of notation. The second term $\|P'_\lambda((|\boldsymbol{\beta}^*_{\mathcal{S}}| - a_0\lambda)_+)\|_2$ is the shrinkage bias induced by the penalty function. For the L_1-penalty $P_\lambda(t) = \lambda t$ ($t \geq 0$), it is easy to see that this term is of order $s^{1/2}\lambda$ regardless of how large the non-zero coordinates of $\boldsymbol{\beta}^*$ are (in magnitude). For a concave penalty that has a decreasing P'_λ, there is a chance that this shrinkage bias might be reduced when the signals are sufficiently strong. A concave penalty function P_λ satisfying (A4) is called *folded concave* if it further satisfies the following property.

(A5) $a_* := \inf\{a > 0 : P'(a) = 0\}$ is finite.

Under assumption (A5) and the minimum signal strength condition (also known as the beta-min condition) that $\|\boldsymbol{\beta}^*_{\mathcal{S}}\|_{\min} \geq (a_0 + a_*)\lambda$, the shrinkage bias becomes zero. The third term, $\sqrt{q}\|\boldsymbol{\zeta}^*\|_2$, depends on the partial gradient of the empirical loss with respect to $\boldsymbol{\alpha}$. Therefore, its order is independent of p and only scales with q.

Recall that Theorem 3.2 is a deterministic result conditioned on some "good" event related to the local RSC structure and the magnitude of the gradient of the empirical loss. Combining Theorem 3.2 with Propositions 3.2 and 3.3, we further provide a complete result characterizing the oracle convergence rate of the iteratively reweighted L_1-penalized SCQR estimator under a weaker beta-min condition than needed in [17].

Theorem 3.3. Suppose assumptions (A1)–(A5) hold, and that there exist $a_1 \geq a_* > a_0 > 0$ such that

$$P'(a_0) > 0 \quad \text{and} \quad \sqrt{4 + 2\{P'(a_0)\}^2} < a_0 \gamma_p \underline{\kappa}\, \underline{f}. \tag{3.21}$$

Moreover, let the regularization parameter λ and bandwidth h satisfy $\lambda \asymp \nu_0 \sigma_x \sqrt{\log(2p)/n}$ and

$$\max\left\{ C_1 \sqrt{\frac{sq\log(2p)}{n}}, C_2 \frac{\max(s,q)\cdot\log(2p)}{n} \right\} \lesssim h \lesssim \underline{f}/l_0,$$

where $C_1 = \sigma_x \nu_0^3$ and $C_2 = \sigma_x^2 \nu_0^6 \overline{f}\,\underline{f}^{-2}$. Then, for any $z > 0$, under the beta-min condition $\|\boldsymbol{\beta}^*_{\mathcal{S}}\|_{\min} \geq (a_0 + a_1)\lambda$, the iteratively reweighted L_1-penalized SCQR estimator $(\widehat{\boldsymbol{\alpha}}^t_h, \widehat{\boldsymbol{\beta}}^t_h)$ with $t \gtrsim \log\log(2p)/\log(1/\delta)$ satisfies the bounds

$$\|\widehat{\boldsymbol{\beta}}^t_h - \boldsymbol{\beta}^*\|_2 \lesssim \underline{f}^{-1} \sqrt{\frac{s + \log q + z}{n}},$$
$$\|\widehat{\boldsymbol{\alpha}}^t_h - \boldsymbol{\alpha}^*\|_2 \lesssim \underline{f}^{-1} q^{1/2} \left(\sqrt{\frac{s + \log q + z}{n}} + h^2 \right)$$

with probability at least $1 - q/p - 2e^{-(s+z)}$ as long as $n \gtrsim s\log p + \log q + t$, where $\delta = \sqrt{4 + 2\{P'(a_0)\}^2}/(a_0\gamma_p\underline{\kappa}\,\underline{f}) \in (0,1)$.

Theorem 3.3 shows that under a beta-min condition $\|\boldsymbol{\beta}^*_{\mathcal{S}}\|_{\min} \gtrsim \sqrt{\log(p)/n}$, the proposed estimator (of $\boldsymbol{\beta}^*$) achieves a near-oracle rate $\sqrt{(s + \log q)/n}$ after as many as $\log(\log p)$ steps, where $q \geq 1$ is a predetermined number of quantile levels. This complements the strong oracle property established in [17], which requires the minimum signal strength to be of order $\sqrt{s\log(p)/n}$.

3.3. Strong oracle property

To establish the strong oracle property of our proposed multi-step estimator $(\widehat{\boldsymbol{\alpha}}^t_h, \widehat{\boldsymbol{\beta}}^t_h)$, we need to show that the estimator equals to the oracle estimator defined in (3.20) for sufficiently large t. We define a similar event that resembles the local RSC property. Let

$$D_{rsc}(\boldsymbol{\alpha}_1, \boldsymbol{\beta}_1, \boldsymbol{\alpha}_2, \boldsymbol{\beta}_2) := \frac{\langle \nabla \widehat{Q}_h(\boldsymbol{\alpha}_1, \boldsymbol{\beta}_1) - \nabla \widehat{Q}_h(\boldsymbol{\alpha}_2, \boldsymbol{\beta}_2), (\boldsymbol{\alpha}_1^{\mathrm{T}} - \boldsymbol{\alpha}_2^{\mathrm{T}}, \boldsymbol{\beta}_1^{\mathrm{T}} - \boldsymbol{\beta}_2^{\mathrm{T}})^{\mathrm{T}} \rangle}{\|\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2\|_{\Sigma}^2 + q^{-1}\|\boldsymbol{\alpha}_1 - \boldsymbol{\alpha}_2\|_2^2}. \tag{3.22}$$

Given radius parameters $r, l > 0$ and a curvature parameter $c > 0$, define

$$\mathcal{R}_{rsc}(r, l, c) := \Big\{ D_{rsc}(\boldsymbol{\alpha}_1, \boldsymbol{\beta}_1, \boldsymbol{\alpha}_2, \boldsymbol{\beta}_2) \geq c \text{ for all } (\boldsymbol{\alpha}_1, \boldsymbol{\beta}_1, \boldsymbol{\alpha}_2, \boldsymbol{\beta}_2) \in \Lambda(r, l) \Big\}, \tag{3.23}$$

where

$$\begin{aligned} \Lambda(r, l) &:= \cap_{k=1}^{q} \Lambda_k(r, l) \cap \{(\boldsymbol{\alpha}_1, \boldsymbol{\beta}_2, \boldsymbol{\alpha}_2, \boldsymbol{\beta}_2) : \begin{bmatrix} \boldsymbol{\alpha}_2 - \boldsymbol{\alpha}_h^* \\ \boldsymbol{\beta}_2 - \boldsymbol{\beta}^* \end{bmatrix} \in \mathbb{B}_{\Omega}(r/2), (\boldsymbol{\beta}_2)_{\mathcal{S}^c} = \mathbf{0}\}, \\ \Lambda_k(r, l) &:= \{(\boldsymbol{\alpha}_1, \boldsymbol{\beta}_1, \boldsymbol{\alpha}_2, \boldsymbol{\beta}_2) : \begin{bmatrix} \alpha_{1k} - \alpha_{2k} \\ \boldsymbol{\beta}_1 - \boldsymbol{\beta}_2 \end{bmatrix} \in \mathbb{B}_{\bar{\Sigma}}(r) \cap \mathbb{C}_{\bar{\Sigma}}(l)\}. \end{aligned} \tag{3.24}$$

Theorem 3.4. Suppose assumptions (A1)–(A5) hold, and for some predetermined $\delta \in (0, 1)$ and $c > 0$, there exist constants $a_1 > a_0 > 0$ such that

$$P'(a_1) = 0, P'(a_0) > 0, \text{ and } a_0 \delta c \gamma_p > \sqrt{1 + P'(a_0)^2/2}. \tag{3.25}$$

Moreover, let $r \geq q^{1/2}\gamma_p^{1/2} a_0 b s^{1/2} \lambda$ and $l = \sqrt{2}\{2 + 2/P'(a_0)\} \cdot \big[\max\{q, (1 + b^2)s\}/\gamma_p\big]^{1/2}$, where $b > 0$ is a constant that satisfies

$$\sqrt{\frac{1 + b^2}{2}} P'(a_0) + 1 = a_0 c \gamma_p b. \tag{3.26}$$

Assume $\|\boldsymbol{\beta}_S^*\|_{\min} \geq (a_0 + a_1)\lambda$. Then, conditioned on the event

$$\begin{aligned} &\{\|\nabla_{\boldsymbol{\beta}} \widehat{Q}_h(\widehat{\boldsymbol{\alpha}}^o, \widehat{\boldsymbol{\beta}}^o)\|_\infty \leq \lambda/2\} \cap \{\|\boldsymbol{\theta}^o\|_\Omega \leq r/2\} \cap \mathcal{R}_{rsc}(r, l, c) \\ &\qquad \cap \left\{ \|\widehat{\boldsymbol{\beta}}^o - \boldsymbol{\beta}^*\|_\infty \leq \left[a_0 - \frac{\sqrt{1 + \{P'(a_0)/2\}^2}}{\delta c \gamma_p} \right] \lambda \right\} \end{aligned} \tag{3.27}$$

where

$$\boldsymbol{\theta}^o = \begin{bmatrix} \frac{\widehat{\boldsymbol{\alpha}}^o - \boldsymbol{\alpha}_h^*}{\sqrt{q}} \\ \widehat{\boldsymbol{\beta}}^o - \boldsymbol{\beta}^* \end{bmatrix},$$

the strong oracle property holds: $\widehat{\boldsymbol{\beta}}^\ell = \widehat{\boldsymbol{\beta}}^o$ provided $\ell \geq \log(s^{1/2}/\delta)/\log(1/\delta)$.

Similar to the Theorem 3.2, Theorem 3.4 is a deterministic result that depends on the event described in (3.27). Thus, our next goal is to control the probability of the event (3.27). To control such probability, we require deviation bound and a non-asymptotic Kiefer-Bahadur representation of the oracle estimator that are of independent interest.

(A1') In addition to Condition (A1), assume $\sup_{u\in\mathbb{R}} |f_\varepsilon(u)| \le \overline{f}$ for some constant $\overline{f} > 0$.

(A2') In addition to Condition (A2), assume $\sup_{u\in\mathbb{R}} K(u) \le \kappa_u$ for some $\kappa_u \in (0, 1]$.

Since the oracle estimator is essentially an unpenalized smoothed CQR estimator in the low-dimensional regime where $s \ll n$, we need to derive relevant results for the low-dimensional smoothed CQR estimator. The following proposition summarizes those results about the oracle estimator that is essential to deriving necessary conditions for the strong oracle property. Same result has been derived in low dimensional smoothed CQR paper by [41], we refer their paper for detailed proof of the following result.

Proposition 3.4. Assume Conditions (A1'), (A2'), and (A3) hold. Then, for any $t \ge 0$, the oracle estimator $(\widehat{\boldsymbol{\alpha}}^o, \widehat{\boldsymbol{\beta}}^o)$ defined in (3.20) satisfies

$$\left\| \begin{matrix} \frac{\widehat{\boldsymbol{\alpha}}^o - \boldsymbol{\alpha}_h^*}{\sqrt{q}} \\ \widehat{\boldsymbol{\beta}}^o - \boldsymbol{\beta}^* \end{matrix} \right\|_\Omega \lesssim \underline{f}^{-1} \sqrt{\frac{s+t}{n}} \tag{3.28}$$

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

Moreover, let $\boldsymbol{S} = \mathbb{E}(\boldsymbol{x}_{\mathcal{S}} \boldsymbol{x}_{\mathcal{S}}^{\mathrm{T}})$, and $\boldsymbol{D} := q^{-1} \sum_{k=1}^q f_\varepsilon(F^{-1}(\tau_k))\boldsymbol{S}$. Then,

$$\left\| \boldsymbol{D}(\widehat{\boldsymbol{\beta}}^o - \boldsymbol{\beta}^*) + \frac{1}{nq} \sum_{i=1}^n \sum_{k=1}^q \{\bar{K}((\alpha_k^* - \varepsilon_i)/h) - \tau_k\} \boldsymbol{x}_{i,\mathcal{S}} \right\|_2$$
$$\lesssim \frac{(s+t)}{h^{1/2} n} + h^{3/2} \sqrt{\frac{q(s+t)}{n}} + h^4$$

with probability at least $1 - 3qe^{-t}$.

Before presenting our final theorem that characterizes the strong oracle property, there is one more event that we need to make sure that it holds with high probability, which is $\mathcal{R}_{rsc}(r, l, c)$. The following Proposition characterizes the event and its probability bound.

Proposition 3.5. Let r, l, and h satisfy

$$24\nu_0^2 r \le h \le \underline{f}/\{\max(4\kappa_2^{1/2} l_0, 2l_0)\} \quad \text{and} \quad nh \ge C\overline{f}\underline{f}^{-2} \max\{s, l^2 \log(p)\}, \tag{3.29}$$

for some sufficiently large constant C independent of (n, s, p, h). Then, the event $\mathcal{R}_{rsc}(r, l, c)$ holds with probability at least $1 - q/(2p)$ with $c = 0.5\underline{\kappa} \cdot \underline{f}$.

With the above preparations, we finally establish the strong oracle property of our iterative estimator.

Theorem 3.5. Assume (A1'), (A2'), and (A3)–(A5) hold. Assume also that

$$\max_{j\in\mathcal{S}^c} \|\boldsymbol{J}_{j\mathcal{S}}(\boldsymbol{J}_{\mathcal{S}\mathcal{S}})^{-1}\|_1 \le A_0 \tag{3.30}$$

for some $A_0 \geq 1$, where $\boldsymbol{J} := q^{-1}\sum_{k=1}^{q} f_\varepsilon(F^{-1}(\tau_k))\Sigma$. Moreover, let $\mu_4 := \sup_{\boldsymbol{u}\in\mathbb{S}^p} \mathbb{E}|\bar{\Sigma}^{-1/2}\bar{\boldsymbol{x}}^{\mathrm{T}}\boldsymbol{u}|^4 < \infty$. For a predetermined $\delta \in (0,1)$, suppose there exist $a_1 > a_0$ satisfying (3.25) with $c = 0.5\underline{\kappa}\underline{f}$, and the beta-min condition $\|\boldsymbol{\beta}^*_{\mathcal{S}}\|_{\min} \geq (a_0+a_1)\lambda$. Let the smoothing bandwidth $h \asymp \{\log(p)/n\}^{1/4}$ and $\lambda \asymp \sqrt{\log(p)/n}$. Then, $\widehat{\boldsymbol{\beta}}^t = \widehat{\boldsymbol{\beta}}^o$ for all $t \geq \lceil \log(s^{1/2}/\delta)/\log(1/\delta)\rceil$ with probability at least $1 - 2q/p - (5q+1)/n$, provided that the sample size satisfies $n \gtrsim \max\{s^{8/3}/(\log p)^{5/3}, s^{4/3}\log(p)\}$.

The above strong oracle property theorem, the required beta-min condition is $\|\boldsymbol{\beta}^*_{\mathcal{S}}\|_{\min} \gtrsim \sqrt{\log(p)/n}$ which further complements the strong oracle property established in [17].

4. Algorithms

In this section, we discuss the computational methods for penalized composite quantile regression, with a particular focus on the weighted L_1-penalty. We first revisit the ADMM-based algorithm proposed in [17], and then describe a local adaptive majorize-minimization (LAMM) for convolution-smoothed CQR. Complexities of the two algorithms are also discussed.

4.1. An alternating direction method of multipliers algorithm

The computation of either L_1-penalized or folded concave penalized composite quantile regression boils down to solving a weighted L_1-penalized problem

$$\min_{\boldsymbol{\alpha},\boldsymbol{\beta}} \left\{ \frac{1}{nq}\sum_{i=1}^{n}\sum_{k=1}^{q} \rho_{\tau_k}(y_i - \alpha_k - \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}) + \sum_{j=1}^{p}\lambda_j|\beta_j| \right\}, \tag{4.1}$$

where $\lambda_j \geq 0$ for $j = 1,\ldots,d$. A conventional strategy is to formulate (4.1) as a linear program, solvable by general-purpose optimization toolboxes. These toolboxes are convenient to use yet are only adapted to small-scale problems. To solve (4.1) more efficiently under high-dimensional settings, [17] proposed an algorithm based on the alternating direction method of multipliers (ADMM). The idea is to cast (4.1) as an equivalent program solvable by ADMM. Specifically, they consider the following reformulation

$$\begin{aligned} &\text{minimize} \quad \frac{1}{nq}\sum_{i=1}^{n}\sum_{k=1}^{q}\rho_{\tau_k}(z_{ik}) + \sum_{j=1}^{p}\lambda_j|\gamma_j| \\ &\text{subject to} \quad \boldsymbol{Z} = \boldsymbol{1}_q^{\mathrm{T}} \otimes \boldsymbol{y} - \boldsymbol{1}_n \otimes \boldsymbol{\alpha}^{\mathrm{T}} - \boldsymbol{1}_q \otimes (\boldsymbol{X}\boldsymbol{\beta}), \boldsymbol{\gamma} = \boldsymbol{\beta}, \end{aligned}$$

where $\boldsymbol{Z} = (z_{ik})_{n\times q} \in \mathbb{R}^{n\times q}$ with $z_{ik} = y_i - \alpha_k - \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}$, $\boldsymbol{X} = (\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n)^{\mathrm{T}} \in \mathbb{R}^{n\times p}$ and $\boldsymbol{\gamma} = (\gamma_1,\ldots,\gamma_p)^{\mathrm{T}}$. Here $\otimes$ denotes the Kronecker product. Let $\boldsymbol{\varphi} = (\boldsymbol{\alpha}^{\mathrm{T}}, \boldsymbol{\beta}^{\mathrm{T}})^{\mathrm{T}} \in \mathbb{R}^{p+q}$ be the total vector of parameters (of interest). The augmented Lagrangian of the above problem is

$$L_\sigma(\boldsymbol{\varphi}, \boldsymbol{Z}, \boldsymbol{\gamma}, \boldsymbol{U}, \boldsymbol{v}) := \frac{1}{nq}\sum_{i=1}^{n}\sum_{k=1}^{q}\rho_{\tau_k}(z_{ik}) + \sum_{j=1}^{p}\lambda_j|\gamma_j| + \langle \mathrm{vec}(\boldsymbol{U}), \mathrm{vec}(\boldsymbol{Z}) + \mathbb{X}_1\boldsymbol{\varphi}\rangle$$

$$+\langle \boldsymbol{v}, \boldsymbol{\gamma} - \mathbb{X}_2\boldsymbol{\varphi}\rangle + \frac{\sigma}{2}\|\mathrm{vec}(\boldsymbol{Z}) + \mathbb{X}_1\boldsymbol{\varphi} - \boldsymbol{Y}\|_{\mathrm{F}}^2 + \frac{\sigma}{2}\|\boldsymbol{\gamma} - \mathbb{X}_2\boldsymbol{\varphi}\|_2^2, \tag{4.2}$$

where $\boldsymbol{U} = (u_{ik}) \in \mathbb{R}^{n\times q}$ and $\boldsymbol{v} = (v_1, \ldots, v_p)^{\mathrm{T}}$ are the Lagrangian multipliers, $\boldsymbol{Y} = \boldsymbol{1}_q \otimes \boldsymbol{y}$ is an (nq)-dimensional vector that stacks q copies of $\boldsymbol{y} \in \mathbb{R}^n$ one underneath the other, $\sigma > 0$ is a optimization parameter and

$$\mathbb{X}_1 = \begin{pmatrix} \boldsymbol{1}_n & \ldots & \boldsymbol{0} & \boldsymbol{X} \\ \vdots & \ddots & \vdots & \vdots \\ \boldsymbol{0} & \ldots & \boldsymbol{1}_n & \boldsymbol{X} \end{pmatrix} \in \mathbb{R}^{(nq)\times(p+q)}, \; \mathbb{X}_2 = (\mathbf{O}_{p\times q} \quad \mathbf{I}_p) \in \mathbb{R}^{p\times(p+q)}.$$

Moreover, let us define $\mathrm{Prox}_\tau(v, a) := v - \max\{(\tau - 1)/a, \min(v, \tau/a)\}$, and $\mathrm{Shrink}(v, a) := \mathrm{sign}(v)(|v| - a)_+$. The former is the proximity operator of the check loss ρ_τ with respect to parameter $a > 0$, and the latter is the proximity operator of $|\cdot|$, also known as the soft-thresholding operator. The ADMM-based algorithm [17] for solving problem (4.1) is summarized in Algorithm 1.

Algorithm 1 The ADMM Algorithm for Solving Weighted L_1-penalized CQR

```
Input: Initialize with $(\boldsymbol{\varphi}^0, \boldsymbol{Z}^0, \boldsymbol{\gamma}^0, \boldsymbol{U}^0, \boldsymbol{v}^0)$, where $\boldsymbol{\varphi}^0 = ((\boldsymbol{\alpha}^0)^{\mathrm{T}}, (\boldsymbol{\beta}^0)^{\mathrm{T}})^{\mathrm{T}}$
For $t = 0, 1, \ldots$, repeat the following steps until convergence.
  1. Update
       $\boldsymbol{\varphi}^{t+1} = \frac{1}{\sigma}(\mathbb{X}_1^{\mathrm{T}}\mathbb{X}_1 + \mathbb{X}_2^{\mathrm{T}}\mathbb{X}_2)^{-1} \cdot \left[\mathbb{X}_1^{\mathrm{T}}\{\sigma\boldsymbol{Y} - \sigma\mathrm{vec}(\boldsymbol{Z}^t) - \mathrm{vec}(\boldsymbol{U}^t)\} + \mathbb{X}_2^{\mathrm{T}}(\sigma\boldsymbol{\gamma}^t + \boldsymbol{v}^t)\right]$    (4.3)
  2. Update
       $z_{ik}^{t+1} = \mathrm{Prox}_{\tau_k}\left(y_i - \alpha_k^{t+1} - \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}^{t+1} - \frac{u_{ik}^t}{\sigma}, nq\sigma\right), 1 \le i \le n, 1 \le k \le q,$    (4.4)
       $\gamma_j^{t+1} = \mathrm{Shrink}\left(\beta_j^{t+1} - \frac{v_j^t}{\sigma}, \frac{\lambda d_j}{\sigma}\right), 1 \le j \le p.$    (4.5)
  3. Update
       $\mathrm{vec}(\boldsymbol{U}^{t+1}) = \mathrm{vec}(\boldsymbol{U}^t) + \sigma\{\mathrm{vec}(\boldsymbol{Z}^{t+1}) + \mathbb{X}_1\boldsymbol{\varphi}^{t+1} - \boldsymbol{Y}\},$
       $\boldsymbol{v}^{t+1} = \boldsymbol{v}^t + \sigma(\boldsymbol{\gamma}^{t+1} - \mathbb{X}_2\boldsymbol{\varphi}^{t+1}).$
```

4.2. A local adaptive majorize-minimization algorithm for smoothed CQR

In this section, we focus on solving a smoothed version of (4.1) with each ρ_{τ_k} replaced by $\ell_{h,k} = \rho_{\tau_k} \circ K_h$. To take advantage of the smoothness and the local strong convexity of the smoothed loss, we employ a variant of the local adaptive majorize-minimization algorithm (LAMM) proposed by [14]. The main idea of LAMM is to construct an isotropic quadratic objective function that locally majorizes the smoothed composite quantile loss such that closed-form updates are available at each iteration. To see this, recall the smoothed objective function $\widehat{Q}_h(\boldsymbol{\alpha}, \boldsymbol{\beta}) = (nq)^{-1} \sum_{i=1}^n \sum_{k=1}^q \ell_{h,k}(y_i - \alpha_k - \boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{\beta})$. For $k = 1, 2, \ldots$, let $\boldsymbol{\varphi}^k = ((\boldsymbol{\alpha}^k)^{\mathrm{T}}, (\boldsymbol{\beta}^k)^{\mathrm{T}})^{\mathrm{T}}$ be the iterate after the k iteration. At the k-th iteration, for some sufficiently large quadratic parameter $\phi_k > 0$, we define a locally majorizing isotropic quadratic function

$$F(\boldsymbol{\varphi}; \phi_k, \boldsymbol{\varphi}^{k-1}) := \widehat{Q}_h(\boldsymbol{\varphi}^{k-1}) + \langle \nabla \widehat{Q}_h(\boldsymbol{\varphi}^{k-1}), \boldsymbol{\varphi} - \boldsymbol{\varphi}^{k-1} \rangle + \frac{\phi_k}{2} \| \boldsymbol{\varphi} - \boldsymbol{\varphi}^{k-1} \|_2^2, \tag{4.6}$$

satisfying $F(\boldsymbol{\varphi}^{k-1}; \phi_k, \boldsymbol{\varphi}^{k-1}) = \widehat{Q}_h(\boldsymbol{\varphi}^{k-1})$. For a large enough ϕ_k, say no less than the largest eigenvalue of $\nabla^2 \widehat{Q}_h(\boldsymbol{\varphi}^{k-1})$, we have $F(\boldsymbol{\varphi}; \phi_k, \boldsymbol{\varphi}^{k-1}) \geq \widehat{Q}_h(\boldsymbol{\varphi})$ for all $\boldsymbol{\varphi}$. Then, we define the updated iterate $\boldsymbol{\varphi}^k = ((\boldsymbol{\alpha}^k)^{\mathrm{T}}, (\boldsymbol{\beta}^k)^{\mathrm{T}})^{\mathrm{T}}$ as the solution to

$$\min_{\boldsymbol{\varphi}} \{ F(\boldsymbol{\varphi}; \phi_k, \boldsymbol{\varphi}^{k-1}) + \| \boldsymbol{\lambda} \circ \boldsymbol{\beta} \|_1 \}, \tag{4.7}$$

where $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_p)^{\mathrm{T}}$. It is easy to see that

$$\begin{aligned} &\widehat{Q}_h(\boldsymbol{\varphi}^k) + \| \boldsymbol{\lambda} \circ \boldsymbol{\beta}^k \|_1 \leq F(\boldsymbol{\varphi}^k; \phi_k, \boldsymbol{\varphi}^{k-1}) + \| \boldsymbol{\lambda} \circ \boldsymbol{\beta}^k \|_1 \\ &\quad \leq F(\boldsymbol{\varphi}^{k-1}; \phi_k, \boldsymbol{\varphi}^{k-1}) + \| \boldsymbol{\lambda} \circ \boldsymbol{\beta}^{k-1} \|_1 = \widehat{Q}_h(\boldsymbol{\varphi}^{k-1}) + \| \boldsymbol{\lambda} \circ \boldsymbol{\beta}^{k-1} \|_1. \end{aligned} \tag{4.8}$$

This ensures that the objective function (with penalty) decreases after each iteration. From the first-order optimality condition we obtain the following closed forms for $\boldsymbol{\alpha}^k$ and $\boldsymbol{\beta}^k$:

$$\begin{aligned} &\alpha_j^k = \alpha_j^{k-1} - \phi_k^{-1} \partial_{\alpha_j} \widehat{Q}_h(\boldsymbol{\varphi}^{k-1}), \quad j = 1, 2, \ldots, q, \\ &\beta_j^k = \mathrm{Shrink}(\beta_j^{k-1} - \phi_k^{-1} \partial_{\beta_j} \widehat{Q}_h(\boldsymbol{\varphi}^{k-1}), \phi_k^{-1} \lambda_j), \quad j = 1, \ldots, p. \end{aligned}$$

To choose a sufficiently large quadratic coefficient ϕ_k that ensures majorization property, we start from a relatively small number, say $\phi_0 = 0.01$, and successively inflate it by a factor $\gamma > 1$, denoted by $\phi_{k,l} = \gamma^l \phi_0$ for $l = 1, 2, \ldots$. If the solution $\boldsymbol{\varphi}^{k,l}$ to (4.7) with $\phi_k = \phi_{k,l}$ satisfies (4.8) for some $l \geq 0$, we stop the search and set $\phi_k = \phi_{k,l}$. Therefore, the quadratic coefficient ϕ_k is automatically determined at each step. By default, we set the optimization parameters to be $(\phi_0, \gamma) = (0.01, 1.25)$. We summarize the whole procedure in Algorithm 2.

From Algorithms 1 and 2 we see that the dominant computational effort of each LAMM update is the multiplication of a $p \times nq$ matrix and (nq)-dimensional vectors, which can be implemented in $O(pnq)$ operations. In addition to this, each ADMM update also involves the multiplication of a $(p+q) \times (p+q)$ matrix and $(p+q)$-dimensional vectors with a complexity $O(pnq + (p+q)^2)$. Moreover, the ADMM needs to compute and store the inverse of $\mathbb{X}_1^{\mathrm{T}} \mathbb{X}_1 + \mathbb{X}_2^{\mathrm{T}} \mathbb{X}_2 \in \mathbb{R}^{(p+q) \times (p+q)}$, hence incurring extra computational cost and memory allocation. Via the Sherman-Morrison-Woodbury formula, the real computational effort of this step is to evaluate the inverse of an $n \times n$ matrix (with complexity $O(n^3)$), which is still expensive when n is large.

Algorithm 2 The LAMM Algorithm for Smoothed CQR with Weighted L_1-penalization

```
Input: Initialize with $\boldsymbol{\alpha}^0 = \mathbf{0}$ and $\boldsymbol{\beta}^0 = \mathbf{0}$
For $k = 0, 1, \ldots$, repeat the following steps until convergence.
1. Set $\phi_k = \max\{\phi_0, \phi_{k-1}/\gamma\}$.
2. Update
       $\alpha_j^k = \alpha_j^{k-1} - \phi_k^{-1} \partial_{\alpha_j} \widehat{Q}_h(\boldsymbol{\varphi}^{k-1}), \quad j = 1, 2, \ldots, q,$
       $\beta_j^k = \mathrm{Shrink}(\beta_j^{k-1} - \phi_k^{-1} \partial_{\beta_j} \widehat{Q}_h(\boldsymbol{\varphi}^{k-1}), \phi_k^{-1}\lambda_j), \quad j = 1. \ldots, p.$
3  If $F(\boldsymbol{\varphi}^k; \phi_k, \boldsymbol{\varphi}^{k-1}) < \widehat{Q}_h(\boldsymbol{\varphi}^k)$, set $\phi_k = \gamma\phi_k$ and repeat Step 2 until $F(\boldsymbol{\varphi}^k; \phi_k, \boldsymbol{\varphi}^{k-1}) \geq \widehat{Q}_h(\boldsymbol{\varphi}^k)$.
```

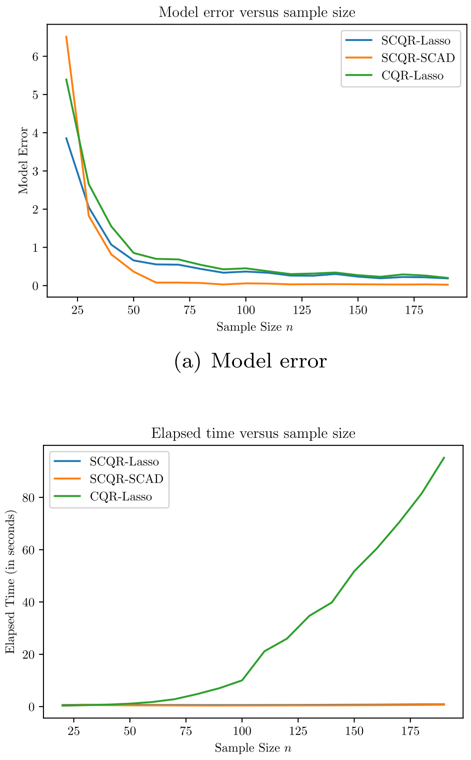
Figure 1 shows a preliminary comparison between the ADMM and LAMM algorithms for computing L_1-penalized CQR estimators on a simulated dataset with increasing n, p subject to $p = 5n$. To make comparisons that are as fair as possible, each algorithm is implemented in `Python`, using the `NumPy` library for basic linear algebra operations. On the statistical aspect, the CQR-Lasso (by ADMM) and SCQR-Lasso (by LAMM) estimators exhibit nearly identical estimation errors (under squared model error); in a speed comparison, the runtime of ADMM grows significantly faster than that of LAMM as the sample size and dimension increase. These preliminary numerical results show evidence that LAMM can also be faster than ADMM by several orders of magnitude. More empirical evidence will be given in the next section.

5. Numerical studies

Recall that composite quantile regression was introduced by [46] as a robust regression method for linear models with heavy-tailed errors that may have infinite variance. The relative efficiency of CQR compared to the least squares is at least 70% regardless the error distribution, could be arbitrarily close to 95.5% in the Gaussian model and arbitrarily large with very heavy-tailed errors. The least absolute deviation (LAD) regression, however, may have an arbitrarily small relative efficiency with respect to the least squares. Recently, [39] introduced a new robust method for high-dimensional regression along with a simulation-based procedure for choosing the regularization parameter. In the Gaussian model, their oracle estimator achieves the same asymptotic relative efficiency (with respect to the least squares) as the CQR.

In the following simulation study, we first compare the penalized SCQR method with its non-smoothed counterpart [17], and then with the robust regression method proposed by [39] when the tuning parameters are automatically chosen for both methods. Data are generated independently from the linear model

$$y = \boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta}^* + \varepsilon, \qquad \boldsymbol{x} \sim N_p(0, \Sigma), \tag{5.1}$$



(a) Model error

(b) Runtime

Fig 1. *A numerical comparison between CQR via ADMM and SCQR via LAMM. Panels (a) and (b) display, respectively, the "model error versus sample size" curve and the "runtime versus sample size" curve. The sample size n increases from 20 to 200, and the dimension p is set as* $5n$.

where $\boldsymbol{\beta}^* = (3, 1.5, 0, 0, 2, 0, \ldots, 0)^{\mathrm{T}} \in \mathbb{R}^p$ and $\Sigma = (0.5^{|j-k|})_{1 \le j,k \le p}$. Independent of $\boldsymbol{x}$, the observation noise ε is generated from one of following four distributions:

(a) The normal distribution with mean 0 and variance 3—$N(0, 3)$.
(b) The mixture normal (MN) distribution—$\sqrt{6} \times \{0.5N(0, 1) + 0.5N(0, 0.5^6)\}$.
(c) The t-distribution with 3 degrees of freedom—t_3.
(d) The standard Cauchy distribution with the density $f(t) = 1/\{\pi(1 + t^2)\}$.

We consider two moderate-scale settings with $(n, p) = (100, 600)$ and $(n, p) = (200, 1200)$.

The statistical performance of each method is measured via the (average) squared model error (with the standard error in the parenthesis), which is $\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_{\Sigma}^2$, the number of false positive results (FP), which is the number of spurious covariates that are selected, and the number of true positive results (TP), which is the number of significant covariates that are selected. For the implementation of CQR, we set $q = 19$ and choose quantile levels $\tau_k = k/20$ for $k = 1, \ldots, 19$.

The theoretical bandwidth requirements suggested in Theorem 3.1 is not scale invariant, which is not a good property in applications. In [15], the authors have evaluated their smoothed quantile regression estimator at the rule-of-thumb bandwidth of [30], which is $h_{\mathrm{ROT}} = 1.06\widehat{s}n^{-1/5}$, where $\widehat{s}$ is the minimum between the sample standard deviation and the interquartile range (divided by 1.38898) of the standard median regression estimator's residuals. In [33], they have shown that the results are insensitive to the choice of the bandwith provided that it is in a reasonable range (neither too small nor too large). In our numerical simulation, we endorsed their choice of the bandwidth parameter $h = \max\left\{0.05, \sqrt{\tau(1-\tau)}\{\log(p)/n\}^{1/4}\right\}$ which is scale invariant, with slight modification that we put $\tau = q^{-1}\sum_{k=1}^{q}\tau_k$.

Table 1 summarizes the simulation results for CQR-Lasso, SCQR-Lasso and SCQR-SCAD that uses the SCAD penalty to compute the weights in (2.8). For a fair comparison between the two methods in terms of statistical and numerical efficiency, we first compute an "oracle" λ value based on the true model error $\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_{\Sigma}$ for each estimator. To be specific, we first compute each estimator along a predetermined sequence of λ values, and choose the λ that minimizes the true model error averaged over 50 replications. Next, we run 100 additional simulations for each method using the optimally chosen λ, and report the results in Table 1. When the L_1 penalty is used, the SCQR has slightly lower model errors than the CQR yet at the cost of more false positives. From the runtime comparison we see a significant computational advantage of the SCQR via LAMM over the CQR via ADMM. As mentioned in Section 3.2, both algorithms are implemented in `Python` using the `NumPy` library for basic linear algebra operations. Moreover, with the optimally chosen λ, the SCQR-SCAD estimator considerably outperforms the Lasso counterparts and achieves oracle-like performance.

We further implement both methods with λ chosen by two data-driven procedures, the (five-fold) cross-validation and a modified BIC method; see Section 2.3 for details. Under the four error distributions, Tables 2 and 3 show the simulation results for CQR-Lasso, SCQR-Lasso and SCQR-SCAD with λ chosen by five-fold cross-validation and BIC, respectively. Statistically, the L_1-penalized CQR and SCQR methods perform similarly in terms of model selection accuracy and estimation accuracy (for $\boldsymbol{\beta}^*$). This empirically validates the theoretical results that smoothing only affects the intercept terms and thus does not compromise the estimation of $\boldsymbol{\beta}^*$. The runtime comparison, on the other hand, shows that the computational cost of ADMM, combined with either cross-validation or BIC, becomes prohibitive as soon as the data has moderately large scales.

We end this section with a numerical comparison of the (smoothed) composite quantile regression method and the robust regression method proposed by [39]. The latter is a combination of the pairwise difference technique and LAD regression. For simplicity, we focus on L_1-penalization. Following the terminol-

Table 1

Statistical performance comparison between the CQR (via ADMM) and the SCQR (via LAMM) estimators under linear model (5.1) with four error distributions. Optimally chosen λ values are used for both methods.

		$n = 100, p = 600$				$n = 200, p = 1200$			
Error	Method	ME	TP	FP	Runtime	ME	TP	FP	Runtime
$N(0,3)$	CQR (Lasso)	0.7680 (0.41)	3	11.62	21.05	0.3581 (0.14)	3	14.66	166.01
	SCQR (Lasso)	0.6617 (0.26)	3	18.37	0.88	0.3317 (0.15)	3	24.31	1.21
	SCQR (SCAD)	0.1096 (0.99)	3	0.06	0.46	0.0474 (0.04)	3	0.04	0.81
MN	CQR (Lasso)	0.3871 (0.18)	3	10.15	18.95	0.1940 (0.08)	3	12.65	147.39
	SCQR (Lasso)	0.3298 (0.16)	3	17.87	0.72	0.1808 (0.06)	3	23.89	1.08
	SCQR (SCAD)	0.0484 (0.04)	3	0.06	0.44	0.0290 (0.02)	3	0.03	0.79
t_3	CQR (Lasso)	0.4131 (0.23)	3	10.3	19.78	0.2196 (0.09)	3	12.78	150.94
	SCQR (Lasso)	0.3661 (0.17)	3	17.57	0.81	0.1839 (0.08)	3	23.58	1.15
	SCQR (SCAD)	0.0561 (0.05)	3	0.04	0.52	0.0246 (0.02)	3	0.03	0.88
Cauchy	CQR (Lasso)	1.3223 (0.99)	3	13.33	24.64	0.6289 (0.38)	3	17.92	190.73
	SCQR (Lasso)	1.0474 (0.68)	3	17.57	0.96	0.4655 (0.26)	3	23.36	1.40
	SCQR (SCAD)	0.2350 (0.39)	3	0.41	0.64	0.1174 (0.16)	3	0.69	1.11

Table 2

Statistical performance comparison between the CQR (via ADMM) and the SCQR (via LAMM) estimators under linear model (5.1) with four error distributions—$N(0,3)$, mixture normal, t_3 and Cauchy. The average of the squared model L_2 error (and standard error), true positives (TP), and false positives (FP), and runtime (in seconds), over 100 replications, are reported. 5-fold CV is used to select λ.

		$n = 100, p = 600$			
Error	Method	ME	TP	FP	Runtime
$N(0,3)$	CQR (Lasso)	0.7924 (0.36)	3	3.52	768.38
	SCQR (Lasso)	0.8066 (0.43)	3	8.77	52.41
	SCQR (SCAD)	0.2074 (0.24)	3	0.31	49.94
MN	CQR (Lasso)	0.4110 (0.20)	3	4.14	626.67
	SCQR (Lasso)	0.5066 (0.29)	3	9.69	46.29
	SCQR (SCAD)	0.1383 (0.13)	3	0.29	44.05
t_3	CQR (Lasso)	0.4125 (0.23)	3	4.27	722.36
	SCQR (Lasso)	0.4966 (0.34)	3	8.60	48.12
	SCQR (SCAD)	0.1395 (0.12)	3	0.35	47.68
Cauchy	CQR (Lasso)	1.2951(1.12)	3	4.74	1158.31
	SCQR (Lasso)	1.2584 (1.06)	3	6.32	63.46
	SCQR (SCAD)	0.3848 (0.48)	2.96	0.22	79.90

Table 3

Statistical performance comparison between the CQR (via ADMM) and the SCQR (via LAMM) estimators under linear model (5.1) with four error distributions—$N(0,3)$, mixture normal, t_3 and Cauchy. The average of the squared model L_2 error (and standard error), true positives (TP), false positives (FP) and runtime (in seconds), over 200 replications, are reported. The BIC (2.9) is used to select λ.

		$n = 100, p = 600$			
Error	Method	ME	TP	FP	Runtime
$N(0,3)$	CQR (Lasso)	1.0595 (0.64)	3	0.64	197.86
	SCQR (Lasso)	0.9438 (0.54)	3	0.62	9.51
	SCQR (SCAD)	0.3394 (0.49)	3	0.95	8.18
MN	CQR (Lasso)	0.5632 (0.27)	3	0.75	157.95
	SCQR (Lasso)	0.5159 (0.24)	3	0.67	8.48
	SCQR (SCAD)	0.1208 (0.17)	3	0.49	7.28
t_3	CQR (Lasso)	0.6291 (0.39)	3	0.54	173.23
	SCQR (Lasso)	0.5659 (0.34)	3	0.53	8.81
	SCQR (SCAD)	0.0876 (0.13)	3	0.22	7.81
Cauchy	CQR (Lasso)	2.7542 (2.64)	2.83	0.30	283.25
	SCQR (Lasso)	2.1228 (1.81)	2.93	0.27	11.51
	SCQR (SCAD)	0.4055 (0.94)	2.88	0.03	12.60

ogy in [39], we refer to their estimator as Rank Lasso, defined as

$$\widehat{\boldsymbol{\beta}}^{\mathrm{RL}}(\lambda) = \underset{\boldsymbol{\beta}\in\mathbb{R}^p}{\operatorname{argmin}} \left\{ \frac{1}{n(n-1)} \sum_{i,j=1, i\neq j}^{n} |(y_i - \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}) - (y_j - \boldsymbol{x}_j^{\mathrm{T}}\boldsymbol{\beta})| + \lambda \sum_{k=1}^{p} |\beta_k| \right\}. \tag{5.2}$$

By utilizing the pivotal property of the L_1-loss [3], they further proposed a simulation-based procedure to choose λ automatically from the data. Computationally, [39] reformulate the optimization problem (5.2) as a linear program (LP), and then use general-purpose optimization toolboxes. We thus follow this route and implement Rank Lasso using the `SciPy` library with method "highs" [20]. In the following simulation study, we use equation (7) in [39] with $c = 1.01$ and $\alpha_0 = 0.1$ to compute the λ in (5.2); for SCQR-Lasso, we simulate λ via (2.10) with $c = 1.9$ and $\alpha = 0.05$.

For data-driven Rank Lasso and SCQR-Lasso estimators, we summarize results on the statistical and computational performance in Table 4 under the four error distributions when $(n, p) = (100, 600)$. The data-driven SCQR typically has much smaller estimation errors but more false positives than the data-driven Rank Lasso. This could just be a consequence of the different tuning procedures. The runtime comparison confirms SCQR as a practical and computational efficient approach to robust regression. The linear program reformulation of (5.2), on the other hand, involves $2n^2 + 2p$ variables and $O(n^2 + p)$ constraints. Even the state-of-the-art LP solvers are not adapted to large-scale problems.

Table 4

Statistical and computational performance comparison of the Rank Lasso and the SCQR methods, among four error distributions: $N(0, 3)$, MN, t_3, and the standard Cauchy, under model (5.1). The mean (and standard error) of the model estimation error, true positives (TP), false positives (FP), and runtime (in seconds) are reported.

		$n = 100, p = 600$			
Error	Method	ME	TP	FP	Runtime
$N(0, 3)$	Rank Lasso	1.5222(0.52)	3	0.30	249.28
	SCQR (Lasso)	0.6970(0.24)	3	17.95	0.62
	SCQR (SCAD)	0.1237(0.13)	3	0.05	0.34
MN	Rank Lasso	0.8145(0.51)	3	0.55	247.97
	SCQR (Lasso)	0.3478(0.22)	3	18.30	0.60
	SCQR (SCAD)	0.0726(0.05)	3	0.05	0.38
t_3	Rank Lasso	0.8324(0.48)	3	0.35	241.71
	SCQR (Lasso)	0.3660(0.21)	3	16.20	0.56
	SCQR (SCAD)	0.0551(0.042)	3	0.05	0.40
Cauchy	Rank Lasso	4.6628(2.89)	3	0.40	241.81
	SCQR (Lasso)	1.3241(0.84)	3	17.55	0.82
	SCQR (SCAD)	0.4395(1.47)	2.95	0.20	0.43

Appendix A: Proof of main results

By a change of variable, we can rewrite the smoothed composite quantile loss $\widehat{Q}_h : \mathbb{R}^{q+p} \to \mathbb{R}$ in (2.6) as

$$\widehat{Q}_h(\boldsymbol{\alpha}, \boldsymbol{\beta}) = \frac{1}{q} \sum_{k=1}^{q} \left\{ (1-\tau_k) \int_{-\infty}^{0} \widehat{F}_h(u+\alpha_k; \boldsymbol{\beta}) \mathrm{d}u + \tau_k \int_{0}^{\infty} (1-\widehat{F}_h(u+\alpha_k; \boldsymbol{\beta})) \mathrm{d}u \right\}.$$

Using this expression, we obtain

$$\partial_{\alpha_k}\widehat{Q}_h(\boldsymbol{\alpha},\boldsymbol{\beta}) = \frac{1}{qn}\sum_{i=1}^{n}\{\bar{K}_h(\alpha_k - r_i(\boldsymbol{\beta})) - \tau_k\}, \quad k=1,\dots,q, \tag{A.1}$$

$$\nabla_{\boldsymbol{\beta}}\widehat{Q}_h(\boldsymbol{\alpha},\boldsymbol{\beta}) = \frac{1}{qn}\sum_{k=1}^{q}\sum_{i=1}^{n}\{\bar{K}_h(\alpha_k - r_i(\boldsymbol{\beta})) - \tau_k\}\boldsymbol{x}_i, \tag{A.2}$$

where

$$\bar{K}(u) = \int_{-\infty}^{u} K(v)\mathrm{d}v \quad \text{and} \quad \bar{K}_h(u) = \bar{K}(u/h).$$

In this notation, we have $\bar{K}_h'(u) = K_h(u) = (1/h)K(u/h)$.

A.1. Proof of Lemma 3.1

Combining (A.1) and (A.2), we see that the full gradient of $\widehat{Q}_h$ with respect to $(\boldsymbol{\alpha}^{\mathrm{T}},\boldsymbol{\beta}^{\mathrm{T}})^{\mathrm{T}} \in \mathbb{R}^{q+p}$ is

$$\nabla\widehat{Q}_h(\boldsymbol{\alpha},\boldsymbol{\beta}) = \frac{1}{qn}\sum_{i=1}^{n}\begin{pmatrix} \bar{K}_h(\alpha_1 - r_i(\boldsymbol{\beta})) - \tau_1 \\ \vdots \\ \bar{K}_h(\alpha_q - r_i(\boldsymbol{\beta})) - \tau_q \\ \sum_{k=1}^{q}\{\bar{K}_h(\alpha_k - r_i(\boldsymbol{\beta})) - \tau_k\}\boldsymbol{x}_i \end{pmatrix} \in \mathbb{R}^{q+p}, \tag{A.3}$$

where $r_i(\boldsymbol{\beta}) = y_i - \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}$. For the Hessian, note that for any $1 \le k,l \le q$ and $1 \le j \le p$,

$$\frac{\partial^2\widehat{Q}_h}{\partial\alpha_k\partial\alpha_l} = \frac{1}{qn}\sum_{i=1}^{n}K_h(\alpha_k - r_i(\boldsymbol{\beta}))\delta_{kl}, \quad \frac{\partial^2\widehat{Q}_h}{\partial\beta_j\partial\alpha_k} = \frac{1}{qn}\sum_{i=1}^{n}K_h(\alpha_k - r_i(\boldsymbol{\beta}))x_{ij}$$

with $\delta_{kl} = I(k=l)$, and

$$\nabla^2_{\boldsymbol{\beta}}\widehat{Q}_h(\boldsymbol{\alpha},\boldsymbol{\beta}) = \frac{1}{nq}\sum_{i=1}^{n}\sum_{k=1}^{q}K_h(\alpha_k - r_i(\boldsymbol{\beta}))\boldsymbol{x}_i\boldsymbol{x}_i^{\mathrm{T}}.$$

For every $\boldsymbol{\beta} \in \mathbb{R}^p$, write

$$\begin{aligned} \boldsymbol{v}_i = \boldsymbol{v}_i(\boldsymbol{\beta}) &= (K_h(\alpha_1 - r_i(\boldsymbol{\beta})),\cdots,K_h(\alpha_q - r_i(\boldsymbol{\beta})))^{\mathrm{T}}, \quad i=1,\dots,n, \\ \boldsymbol{v} = (v_1,\dots,v_q)^{\mathrm{T}} = \boldsymbol{v}(\boldsymbol{\beta}) &= (K_h(\alpha_1 - r(\boldsymbol{\beta})),\cdots,K_h(\alpha_q - r(\boldsymbol{\beta})))^{\mathrm{T}}, \end{aligned}$$

where $r(\boldsymbol{\beta}) = y - \boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta}$. It follows that

$$\nabla^2\widehat{Q}_h(\boldsymbol{\alpha},\boldsymbol{\beta}) = \frac{1}{nq}\sum_{i=1}^{n}\begin{pmatrix} \mathrm{diag}(\boldsymbol{v}_i) & \boldsymbol{v}_i\boldsymbol{x}_i^{\mathrm{T}} \\ \boldsymbol{x}_i\boldsymbol{v}_i^{\mathrm{T}} & 1_q^{\mathrm{T}}\boldsymbol{v}_i \cdot \boldsymbol{x}_i\boldsymbol{x}_i^{\mathrm{T}} \end{pmatrix} \tag{A.4}$$

and

$$\nabla^2 Q_h(\boldsymbol{\alpha},\boldsymbol{\beta}) = \frac{1}{q}\begin{pmatrix} \mathbb{E}(\mathrm{diag}(\boldsymbol{v})) & \mathbb{E}(\boldsymbol{v}\boldsymbol{x}^{\mathrm{T}}) \\ \mathbb{E}(\boldsymbol{x}\boldsymbol{v}^{\mathrm{T}}) & \mathbb{E}(1_q^{\mathrm{T}}\boldsymbol{v}\cdot\boldsymbol{x}\boldsymbol{x}^{\mathrm{T}}) \end{pmatrix}, \tag{A.5}$$

where $\mathbf{1}_q = (1,\ldots,1)^{\mathrm{T}} \in \mathbb{R}^q$.

For any $\boldsymbol{a} = (a_1,\ldots,a_q)^{\mathrm{T}} \in \mathbb{R}^q$ and $\boldsymbol{b} \in \mathbb{R}^p$, note that

$$(\boldsymbol{a}^{\mathrm{T}},\boldsymbol{b}^{\mathrm{T}})\nabla^2 Q_h(\boldsymbol{\alpha},\boldsymbol{\beta})(\boldsymbol{a}^{\mathrm{T}},\boldsymbol{b}^{\mathrm{T}})^{\mathrm{T}} = \frac{1}{q}\sum_{k=1}^q \mathbb{E}\{v_k(a_k+\boldsymbol{x}^{\mathrm{T}}\boldsymbol{b})^2\} \geq 0,$$

where $v_k = K_h(\alpha_k - r(\boldsymbol{\beta}))$.

This verifies the positive semidefiniteness of $\nabla^2 Q_h(\boldsymbol{\alpha},\boldsymbol{\beta})$ and so is the convexity of Q_h. At $(\boldsymbol{\alpha},\boldsymbol{\beta}) = (\boldsymbol{\alpha}^*,\boldsymbol{\beta}^*)$, $v_k = K_h(\alpha_k^* - \beta_0^* - \varepsilon) = K_h(F^{-1}(\tau_k) - \varepsilon)$. Using the independence of ε and $\boldsymbol{x}$, and condition (3.2), we obtain that

$$\begin{aligned}
&(\boldsymbol{a}^{\mathrm{T}},\boldsymbol{b}^{\mathrm{T}})\nabla^2 Q_h(\boldsymbol{\alpha},\boldsymbol{\beta})(\boldsymbol{a}^{\mathrm{T}},\boldsymbol{b}^{\mathrm{T}})^{\mathrm{T}}\big|_{(\boldsymbol{\alpha},\boldsymbol{\beta})=(\boldsymbol{\alpha}^*,\boldsymbol{\beta}^*)} \\
&= \frac{1}{q}\sum_{k=1}^q \mathbb{E}(a_k+\boldsymbol{x}^{\mathrm{T}}\boldsymbol{b})^2 \cdot \mathbb{E}K_h(F^{-1}(\tau_k)-\varepsilon) \\
&= \frac{1}{q}\sum_{k=1}^q \mathbb{E}(a_k+\boldsymbol{x}^{\mathrm{T}}\boldsymbol{b})^2 \cdot \int_{-\infty}^{\infty} K(v) f(F^{-1}(\tau_k)-hv)\mathrm{d}v > 0
\end{aligned}$$

for all $\boldsymbol{a} \in \mathbb{R}^q$ and $\boldsymbol{b} \in \mathbb{R}^p$, where the second equality follows from integration by parts and a change of variable. This proves the strict convexity of Q_h at $(\boldsymbol{\alpha}^*,\boldsymbol{\beta}^*)$.

Turning to the sample Hessian, for any $\boldsymbol{a} = (a_1,\ldots,a_q)^{\mathrm{T}} \in \mathbb{R}^q, \boldsymbol{b} \in \mathbb{R}^p$ we have

$$(\boldsymbol{a}^{\mathrm{T}},\boldsymbol{b}^{\mathrm{T}})\nabla^2 \widehat{Q}_h(\boldsymbol{\alpha},\boldsymbol{\beta})(\boldsymbol{a}^{\mathrm{T}},\boldsymbol{b}^{\mathrm{T}})^{\mathrm{T}} = \frac{1}{nq}\sum_{i=1}^n\sum_{k=1}^q K_h(\alpha_k - r_i(\boldsymbol{\beta}))(a_k+\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{b})^2 \geq 0.$$

Hence, the empirical composite quantile loss $\widehat{Q}_h$ is twice-differentiable and convex. □

A.2. Proof of Proposition 3.1

We first show that the function $m_h : \mathbb{R}^q \to \mathbb{R}$ has a unique minimizer, denoted by $\boldsymbol{b}_h$. For each $1 \leq k \leq q$, define the univariate function $m_{h,k}(b) = \mathbb{E}\ell_{h,k}(\varepsilon - b)$ whose first and second-order derivatives are

$$m'_{h,k}(b) = \int_{-\infty}^{\infty} K(v)F(b-hv)\mathrm{d}v - \tau_k, \quad m''_{h,k}(b) = \int_{-\infty}^{\infty} K(v)F(b-hv)\mathrm{d}v.$$

Since K is positive everywhere, we have $m''_{h,k}(b) > 0$ for all b. Therefore, $m_{h,k}(\cdot)$ is strictly convex and has a unique minimizer, denoted by $b_{h,k}$. Noting further that $\nabla^2 m_h(\boldsymbol{b}) = q^{-1}\mathrm{diag}(\{m''_{h,1}(b_1), \ldots, m''_{h,q}(b_q)\})$, the function $m_h : \mathbb{R}^q \to \mathbb{R}$ is also strictly convex with a unique minimizer $\boldsymbol{b}_h = (b_{h,1}, \ldots, b_{h,q})^{\mathrm{T}}$.

For any $\boldsymbol{\alpha} \in \mathbb{R}^q, \boldsymbol{\beta} \in \mathbb{R}^p$, we write $\Delta_{\boldsymbol{x}} = \boldsymbol{x}^{\mathrm{T}}(\boldsymbol{\beta} - \boldsymbol{\beta}^*)$ and obtain that

$$\begin{aligned}
Q_h(\boldsymbol{\alpha}, \boldsymbol{\beta}) &= \frac{1}{q}\sum_{k=1}^{q} \mathbb{E}\ell_{h,k}(\varepsilon + \beta_0^* - \alpha_k - \Delta_{\boldsymbol{x}}) \\
&= \mathbb{E}\bigg[\frac{1}{q}\sum_{k=1}^{q} \mathbb{E}\{\ell_{h,k}(\varepsilon + \beta_0^* - \alpha_k - \Delta_{\boldsymbol{x}})|\boldsymbol{x}\}\bigg] \\
&= \mathbb{E}\{m_h(\alpha_1 + \Delta_{\boldsymbol{x}} - \beta_0^*, \ldots, \alpha_q + \Delta_{\boldsymbol{x}} - \beta_0^*)\} \\
&\geq m_h(\boldsymbol{b}_h) = Q_h(\beta_0^* + \boldsymbol{b}_h, \boldsymbol{\beta}^*).
\end{aligned}$$

In other words, the function $Q_h : \mathbb{R}^{q+p} \to \mathbb{R}$ is minimized at $(\beta_0^* + \boldsymbol{b}_h, \boldsymbol{\beta}^*)$. With a everywhere positive kernel, Q_h is strictly convex so that $(\beta_0^* + \boldsymbol{b}_h, \boldsymbol{\beta}^*)$ is the unique minimizer, implying $\boldsymbol{\alpha}_h^* = \beta_0^* + \boldsymbol{b}_h$ and $\boldsymbol{\beta}_h^* = \boldsymbol{\beta}^*$ as claimed.

Finally, it remains to bound $|b_{h,k} - F^{-1}(\tau_k)|$. For each $k = 1, \ldots, q$, define $\widetilde{b}_k = F^{-1}(\tau_k) + t_k\{b_{h,k} - F^{-1}(\tau_k)\}$ with

$$t_k = \sup\{t \in [0,1] : t|b_{h,k} - F^{-1}(\tau_k)| \leq \kappa_2^{1/2}h\}.$$

When $|b_{h,k} - F^{-1}(\tau_k)| \leq \kappa_2^{1/2}h$, we have $t_k = 1$, otherwise $t_k \in (0,1)$. Set $\delta_k = \widetilde{b}_k - F^{-1}(\tau_k)$, satisfying $|\delta_k| \leq \kappa_2^{1/2}h$ and in particular, $|\delta_k| = \kappa_2^{1/2}h$ if $|b_{h,k} - F^{-1}(\tau_k)| > \kappa_2^{1/2}h$. By Lemma 3.1 and the fact that $m'_{h,k}(b_{h,k}) = 0$, we have

$$\begin{aligned}
\{m'_{h,k}(\widetilde{b}_k) - m'_{h,k}(F^{-1}(\tau_k))\}\delta_k &\leq \{m'_{h,k}(b_{h,k}) - m'_{h,k}(F^{-1}(\tau_k))\}\delta_k \\
&\leq |m'_{h,k}(F^{-1}(\tau_k))| \cdot |\delta_k|.
\end{aligned}$$

For the left-hand side,

$$\begin{aligned}
&m'_{h,k}(\widetilde{b}_k) - m'_{h,k}(F^{-1}(\tau_k)) = \int_{F^{-1}(\tau_k)}^{\widetilde{b}_k} m''_{h,k}(t)\,\mathrm{d}t \\
&= \int_{F^{-1}(\tau_k)}^{\widetilde{b}_k} \int_{-\infty}^{\infty} K(u) f(t - hu)\,\mathrm{d}u\,\mathrm{d}t \\
&= f(F^{-1}(\tau_k))\delta_k + \int_{F^{-1}(\tau_k)}^{\widetilde{b}_k} \int_{-\infty}^{\infty} K(u)\{f(t - hu) - f(F^{-1}(\tau_k))\}\,\mathrm{d}u\,\mathrm{d}t.
\end{aligned}$$

This, together with the Lipschitz continuity of f, implies

$$\{m'_{h,k}(\widetilde{b}_k) - m'_{h,k}(F^{-1}(\tau_k))\}\delta_k \geq f(F^{-1}(\tau_k))\delta_k^2 - \frac{l_0}{2}|\delta_k|^3 - l_0\kappa_1 h \cdot \delta_k^2.$$

On the other hand, we have

$$|m'_{h,k}(F^{-1}(\tau_k))| = |\int_{-\infty}^{\infty} K(u)\{F(F^{-1}(\tau_k) - hu) - F(F^{-1}(\tau_k))\}\,\mathrm{d}u| \leq l_0\kappa_2 h^2/2.$$

Putting together the pieces, we conclude that

$$f(F^{-1}(\tau_k))\delta_k^2 \leq \frac{l_0}{2}\kappa_2 h^2|\delta_k| + \frac{l_0}{2}|\delta_k|^3 + l_0\kappa_1 h\delta_k^2 < \kappa_2 l_0 h^2|\delta_k| + \frac{1}{2}f(F^{-1}(\tau_k))\delta_k^2$$

provided $0 < h \leq f(F^{-1}(\tau_k))/(2\kappa_2^{1/2}l_0)$, where the second inequality follows from the fact that $\kappa_1 < \kappa_2^{1/2}$. Canceling out $|\delta_k|$ from both sides yields

$$|\delta_k| < \frac{2\kappa_2^{1/2}l_0 h}{f(F^{-1}(\tau_k))} \cdot \kappa_2^{1/2}h \leq \kappa_2^{1/2}h.$$

By the definition of $\widetilde{b}_k$, we must have $|b_{h,k} - F^{-1}(\tau_k)| \leq \kappa_2^{1/2}h$; otherwise $|\delta_k| = \kappa_2^{1/2}h$ which contradicts the above inequality. Consequently, $t_k = 1$ and $\widetilde{b}_k = b_{h,k}$, thus implying the claimed bound (3.4). □

A.3. Proof of Proposition 3.2

We first consider $\|\boldsymbol{\zeta}^*\|_\infty = \max_{1\leq k\leq q}|\zeta_k^*|$, where

$$\zeta_k^* = \frac{1}{nq}\sum_{i=1}^n\{\bar{K}((\alpha_{h,k}^* - \varepsilon_i)/h) - \tau_k\}$$

is the k-th coordinate of $\boldsymbol{\zeta}^*$. Note that $\bar{K}((\alpha_{h,k}^* - \varepsilon_i)/h) - \tau_k \in [-\tau_k, 1-\tau_k]$. Hence, applying Hoeffding's inequality yields

$$\mathbb{P}\{|\zeta_k^*| > 3\lambda/(2q)\} \leq 2e^{-2n(3\lambda/2)^2} = 2e^{-9n\lambda^2/2}.$$

Taking the union bound over $k = 1, \ldots, q$, it follows that $\mathbb{P}\{\|\boldsymbol{\zeta}^*\|_\infty > 3\lambda/(2q)\} \leq 2qe^{-9n\lambda^2/2}$.

For $\boldsymbol{w}^* = (\omega_1^*, \ldots, \omega_p^*)^{\mathrm{T}} \in \mathbb{R}^p$, let

$$z_{ij} = \frac{x_{ij}}{q}\sum_{k=1}^q\{\bar{K}((\alpha_{h,k}^* - \varepsilon_i)/h) - \tau_k\},$$

so that $\omega_j^* = (1/n)\sum_{i=1}^n z_{ij}$. Note that $\mathbb{E}(z_{ij}) = 0$, $|z_{ij}| \leq |x_{ij}|$, and by assumption (A3), $\mathbb{P}(|x_{ij}| \geq \nu_0\sigma_{jj}^{1/2}t) \leq e^{-t^2/2}$ for all $t \geq 0$. It thus follows from Proposition 2.5 of [37] that

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^n z_{ij}\right| > t\right) \leq 2\exp\{-nt^2/(8\nu_0^2\sigma_{jj})\}. \tag{A.6}$$

Finally, taking $t = \lambda/2$ and applying the union bound over $j = 1, \ldots, p$ prove the claimed bound. □

A.4. Proof of Proposition 3.3

We have

$$D(\boldsymbol{\alpha},\boldsymbol{\beta})=\frac{1}{nq}\sum_{i=1}^{n}\sum_{k=1}^{q}\left\{\bar{K}\left(\frac{\alpha_k-r_i(\boldsymbol{\beta})}{h}\right)-\bar{K}\left(\frac{\alpha^*_{h,k}-r_i(\boldsymbol{\beta}^*)}{h}\right)\right\}\langle\bar{\boldsymbol{x}}_i,(\delta_k,\boldsymbol{\Delta})\rangle, \tag{A.7}$$

by letting $\bar{\boldsymbol{x}}_i=(1,\boldsymbol{x}_i^{\mathrm{T}})^{\mathrm{T}}$, where $\boldsymbol{\delta}=\boldsymbol{\alpha}-\boldsymbol{\alpha}^*_h$, $\boldsymbol{\Delta}=\boldsymbol{\beta}-\boldsymbol{\beta}^*$.

We restrict our focus on the symmetrized Bregmann divergence with each quantile index $k=1,\dots,q$, by letting

$$D_k(\boldsymbol{\alpha},\boldsymbol{\beta}):=\frac{1}{n}\sum_{i=1}^{n}\left\{\bar{K}\left(\frac{\alpha_k-r_i(\boldsymbol{\beta})}{h}\right)-\bar{K}\left(\frac{\alpha^*_{h,k}-r_i(\boldsymbol{\beta}^*)}{h}\right)\right\}\langle\bar{\boldsymbol{x}}_i,\boldsymbol{\Delta}_k\rangle,$$

where $\boldsymbol{\Delta}_k:=(\delta_k,\boldsymbol{\Delta})$.

We first claim that, when $\|\boldsymbol{\Delta}_k\|_{\bar{\Sigma}}\le r$, we have the following lower bound

$$D_k(\boldsymbol{\alpha},\boldsymbol{\beta})\ge c\|\boldsymbol{\Delta}_k\|^2_{\bar{\Sigma}}$$

with high probability for some positive $c>0$, so that we can combine all the bounds for each k to derive the desired result of the Proposition 3.3. Let us define an event in the neighborhood of the true parameter, $\mathcal{E}_{i,k}:=\{|\varepsilon_i-\alpha^*_{h,k}|\le h/2\}\cap\{|\langle\bar{\boldsymbol{x}}_i,\boldsymbol{\Delta}_k\rangle|/\|\boldsymbol{\Delta}_k\|_{\bar{\Sigma}}\le h/(2r)\}$, with $r>0$. We can lower bound $D_k(\boldsymbol{\alpha},\boldsymbol{\beta})$ by

$$\frac{\underline{\kappa}}{nh}\sum_{i=1}^{n}\langle\bar{\boldsymbol{x}}_i,\boldsymbol{\Delta}_k\rangle^2 I_{\mathcal{E}_{i,k}}, \tag{A.8}$$

where $\underline{\kappa}:=\min_{|x|\le1}K(x)$. Also, by using similar smoothing technique from [26], we define a Lipshitz continuous function for $R>0$, as

$$\varphi_R(u):=u^2I(|u|\le R/2)+(|u|-R)^2I(R/2<|u|\le R). \tag{A.9}$$

Then, we can further lower bound (A.8) by

$$\frac{\underline{\kappa}}{nh}\sum_{i=1}^{n}\langle\bar{\boldsymbol{x}}_i,\boldsymbol{\Delta}_k\rangle^2 I_{\mathcal{E}_{i,k}}\ge\underline{\kappa}\cdot\underbrace{\frac{1}{nh}\sum_{i=1}^{n}\varphi_{h/2}(\langle\boldsymbol{\Delta}_k,\bar{\boldsymbol{x}}_i\rangle)\cdot\chi_{i,k}}_{=:D_{0,k}(\boldsymbol{\alpha},\boldsymbol{\beta})}, \tag{A.10}$$

where $\chi_{i,k}=I(|\varepsilon_i-\alpha^*_{h,k}|\le h/2)$.

To prove our claim, now it suffices to show that when $\|\boldsymbol{\Delta}_k\|_{\bar{\Sigma}}=\delta r$, for each $(\delta\in(0,1])$, we have

$$\frac{\underline{\kappa}}{nh}\sum_{i=1}^{n}\varphi_{\delta(h/2)}(\langle\boldsymbol{\Delta}_k,\bar{\boldsymbol{x}}_i\rangle)\cdot\chi_{i,k}\ge c(\delta r)^2. \tag{A.11}$$

If it holds for $\delta = 1$, then

$$\frac{\underline{\kappa}}{nh}\sum_{i=1}^{n} \varphi_{h/2}(\langle \boldsymbol{\Delta}_k/\delta, \bar{\boldsymbol{x}}_i\rangle) \cdot \chi_{i,k} \geq cr^2,$$

which gives

$$\frac{\underline{\kappa}}{nh}\sum_{i=1}^{n} \varphi_{\delta(h/2)}(\langle \boldsymbol{\Delta}_k, \bar{\boldsymbol{x}}_i\rangle) \cdot \chi_{i,k} \geq c(\delta r)^2.$$

Hence, we only need to prove when $\|\boldsymbol{\Delta}_k\|_{\bar{\Sigma}} = r$. Suppose $\|\boldsymbol{\Delta}_k\|_{\bar{\Sigma}} = r$, we have

$$|\mathbb{E}\chi_{i,k} - hf(\alpha^*_{h,k})| \leq \int_{\alpha^*_{h,k}-h/2}^{\alpha^*_{h,k}+h/2} |f_\varepsilon(t) - f_\varepsilon(\alpha^*_{h,k})|\mathrm{d}t \leq \frac{l_0}{4}h^2$$

with Proposition 3.1. Moreover, we have the following lower bound

$$\mathbb{E}\chi_{i,k} \geq hf_\varepsilon(\alpha^*_{h,k}) - \frac{l_0}{4}h^2 \geq h(\underline{f} - l_0 C_b h^2 - \frac{l_0}{4}h) \geq \frac{3}{4}\underline{f}h \tag{A.12}$$

when $h \leq \min\{\underline{f}/(4\kappa_2^{1/2} l_0), \underline{f}/(2l_0)\}$, where $C_b := 2\kappa_2 l_0/\underline{f}$ is an upper bound of the bias $|b_{h,k} - f(F^{-1}(\tau_k))|$ that derived in the Proposition 3.1. Then, using above results, we get

$$\begin{aligned}\mathbb{E}\{h^{-1}\varphi_{h/2}(\langle \boldsymbol{\Delta}_k, \bar{\boldsymbol{x}}_i\rangle)\chi_{i,k}\} &\geq \frac{3}{4}\underline{f}\mathbb{E}\varphi_{h/2}(\langle \boldsymbol{\Delta}_k, \bar{\boldsymbol{x}}_i\rangle) \\ &\geq \frac{3}{4}\underline{f}[r^2 - \mathbb{E}\{\langle \boldsymbol{\Delta}_k, \bar{\boldsymbol{x}}_i\rangle^2 I\{|\langle \boldsymbol{\Delta}_k, \bar{\boldsymbol{x}}_i\rangle| \geq h/4)\}].\end{aligned}$$

The last term of the above inequality is equal to

$$\frac{3}{4}\underline{f}r^2\big(1 - \mathbb{E}\{\boldsymbol{\xi}^2_{\boldsymbol{\Delta}_k} I_{|\boldsymbol{\xi}_{\boldsymbol{\Delta}_k}| \geq h/(4r)}\}\big), \tag{A.13}$$

where $\boldsymbol{\xi}_{\boldsymbol{\Delta}_k} = \langle \boldsymbol{\Delta}_k, \bar{\boldsymbol{x}}_i\rangle / \|\boldsymbol{\Delta}_k\|_{\bar{\Sigma}}$.

Since we have sub-Gaussian covariates, for any $u > 0$, we get

$$\begin{aligned}\mathbb{E}\{\boldsymbol{\xi}^2_{\boldsymbol{\Delta}_k} I(|\boldsymbol{\xi}_{\boldsymbol{\Delta}_k}| > u)\} &= 2\mathbb{E}\left\{\int_0^\infty t \cdot I(|\boldsymbol{\xi}_{\boldsymbol{\Delta}_k}| > t) I(|\boldsymbol{\xi}_{\boldsymbol{\Delta}_k}| > u)\mathrm{d}t\right\} \\ &= 2\mathbb{E}\int_0^u t \cdot I(|\boldsymbol{\xi}_{\boldsymbol{\Delta}_k}| > t) I(|\boldsymbol{\xi}_{\boldsymbol{\Delta}_k}| > u)\mathrm{d}t \\ &+ 2\mathbb{E}\int_u^\infty t \cdot I(|\boldsymbol{\xi}_{\boldsymbol{\Delta}_k}| > t)\mathrm{d}t \\ &= u^2\mathbb{P}(|\boldsymbol{\xi}_{\boldsymbol{\Delta}_k}| > u) + 2\int_u^\infty t \cdot \mathbb{P}(|\boldsymbol{\xi}_{\boldsymbol{\Delta}_k}| > t)\mathrm{d}t \\ &\leq u^2 e^{-(u/\sqrt{2}\nu_0)^2} + 2\nu_0^2\int_{u/\nu_0}^\infty t \cdot \mathbb{P}(|\boldsymbol{\xi}_{\boldsymbol{\Delta}_k}| > \nu_0 t)\mathrm{d}t \\ &\leq (u^2 + 2\nu_0^2)e^{-(u/\sqrt{2}\nu_0)^2},\end{aligned}$$

where we use the condition (A3) to establish the last two inequalities.

To simplify, define $L_\delta := \min\big\{L : \mathbb{E}(\boldsymbol{v}^{\mathrm{T}}\bar{\boldsymbol{x}})^2 \cdot I(|\boldsymbol{v}^{\mathrm{T}}\bar{\boldsymbol{x}}| > L) \le \delta$ for all $\boldsymbol{v} \in \mathbb{R}^{p+1}, \|\boldsymbol{v}\|_{\bar{\Sigma}} = 1\big\}$. Then, we get $L_{1/8} \le h/(4r)$ when $h/(4r) \ge 3\nu_0^2$., which leads to

$$\inf_{\|\boldsymbol{\Delta}_k\|_{\bar{\Sigma}}=r} \mathbb{E}\{D_{0,k}(\boldsymbol{\alpha},\boldsymbol{\beta})\} > \frac{21}{32}\underline{f}r^2. \tag{A.14}$$

Now, we need to bound $|D_{0,k}(\boldsymbol{\alpha},\boldsymbol{\beta}) - \mathbb{E}\{D_{0,k}(\boldsymbol{\alpha},\boldsymbol{\beta})\}|$. The domain of interest is $\mathbb{B}_\Omega(r) \cap \mathbb{C}_\Omega(l)$, in particular $\|(\boldsymbol{\delta}^{\mathrm{T}}, \boldsymbol{\Delta}^{\mathrm{T}})^{\mathrm{T}}\|_\Omega \le r$. Thus, when $\|\boldsymbol{\Delta}_k\|_{\bar{\Sigma}} = r$, it also satisfies that $\|\boldsymbol{\Delta}_k\|_1 \le l\|\boldsymbol{\Delta}_k\|_{\bar{\Sigma}}$. Let us define

$$Z_n(l) := \sup_{\|\boldsymbol{\Delta}_k\|_1 \le lr} |D_{0,k}(\boldsymbol{\alpha},\boldsymbol{\beta}) - \mathbb{E}\{D_{0,k}(\boldsymbol{\alpha},\boldsymbol{\beta})\}|.$$

Then, we have $0 \le h^{-1}\varphi_{h/(2r)}(\boldsymbol{\xi}_{\boldsymbol{\Delta}_k}) \le (4r)^{-2}h$, and

$$\begin{aligned}\mathbb{E}\{h^{-2}\varphi^2_{h/(2r)}(\boldsymbol{\xi}_{\boldsymbol{\Delta}_k})\chi_{i,k}\} &\le (h/4r)^2h^{-2}\mathbb{E}\big(\boldsymbol{\xi}^2_{\boldsymbol{\Delta}_k} \cdot \chi_{i,k}\big)\\ &\le (4r)^{-2} \cdot \{hf_\varepsilon(\alpha^*_{h,k}) + l_0h^2/4\}\\ &\le (4r)^{-2}(5\bar{f}h/4),\end{aligned} \tag{A.15}$$

where the last inequality follows from Proposition 3.1.

To control $Z_n(l)$ defined above, let us divide it by r^2 for convenience, which gives

$$\frac{1}{r^2} \cdot Z_n(l) = \sup_{\|\boldsymbol{\Delta}_k\|_1 \le l\|\boldsymbol{\Delta}_k\|_{\bar{\Sigma}}} \left|\frac{1}{nh}\sum_{i=1}^n \{\varphi_{h/(2r)}(\boldsymbol{\xi}_{\boldsymbol{\Delta}_k}) - \mathbb{E}\varphi_{h/(2r)}(\boldsymbol{\xi}_{\boldsymbol{\Delta}_k})\}\right|. \tag{A.16}$$

With (A.15) and above preparations on $Z' := Z_n(l)/r^2$, we can apply Theorem 7.3 of [6], a refined Talagrand's inequality, which gives

$$\mathbb{P}\Big\{Z' \le \mathbb{E}(Z') + h/(16r^2n)\sqrt{40n\bar{f}r^2x/h + 2\mathbb{E}(Z')x} + hx/(48r^2n)\Big\} \le e^{-x}. \tag{A.17}$$

We further simplify the inequality above as

$$\begin{aligned}Z' &\le \mathbb{E}(Z') + \{\mathbb{E}(Z')\}^{1/2}\sqrt{\frac{hx}{4r^2n}} + (4r)^{-1} \cdot 2\bar{f}^{1/2}\sqrt{\frac{hx}{n}} + \frac{hx}{48r^2n}\\ &\le \frac{5}{4}\mathbb{E}(Z') + \sqrt{\frac{\bar{f}hx}{4r^2n}} + \frac{hx}{3r^2n}\end{aligned} \tag{A.18}$$

with probability at least $1 - e^{-x}$, where the last inequality follows from $ab \le a^2/4 + b^2$.

By taking expectation on the right-hand side of (A.16), we apply Talagrand's contraction principle in Theorem 4.12 of [23], which leads to

$$\mathbb{E}(Z') \le \frac{l}{r} \cdot \mathbb{E}\left\|\frac{1}{n}\sum_{i=1}^n e_i\chi_{i,k}\bar{\boldsymbol{x}}_i\right\|_\infty, \tag{A.19}$$

where e_i's are independent Rademacher random variables.

We have $\mathbb{E}(e_i\chi_{i,k}x_{ij}) = 0$ and $\mathbb{E}(e_i\chi_{i,k}x_{ij})^2 \le \sigma_{jj}c_h$, where $c_h = (9/8)\overline{f}h + l_0h^2/4$. Also, for $k = 3, 4, \ldots$,

$$\begin{aligned}\mathbb{E}|e_i\chi_{i,k}x_{ij}|^k &\le c_h \cdot k\int_0^\infty u^{k-1}\mathbb{P}(|x_{ij}| \ge u)\mathrm{d}u\\ &\le c_h\nu_0^k\sigma_{jj}^{k/2} \cdot k\int_0^\infty \mathbb{P}(|x_{ij}| \ge \nu_0\sigma_{jj}^{1/2}t)\mathrm{d}t\\ &\le c_h\nu_0^k\sigma_{jj}^{k/2} \cdot k\int_0^\infty e^{-t^2/2}t^{k-1}\mathrm{d}t\\ &\le \frac{k!}{2}\cdot\nu_0^2\sigma_{jj}c_h\cdot(2\nu_0\sigma_{jj}^{1/2})^{k-2}.\end{aligned} \tag{A.20}$$

Then, following the proof of Theorem 2.10 in [5], letting $v = \nu_0^2\sigma_x^2c_hn, c = 2\nu_0\sigma_x$ and using Bernstein's inequality, we get

$$\mathbb{E}(Z') \le 2\nu_0\sigma_x\frac{l}{r}\left(\sqrt{\frac{\overline{f}h\log(2p)}{n}} + \frac{\log(2p)}{n}\right). \tag{A.21}$$

Hence, combining bounds (A.18) and (A.21) with $x = \log(2p)$, we get

$$Z' \le \frac{1}{2}(1 + 5\nu_0\sigma_xl)\sqrt{\frac{\overline{f}h\log(2p)}{r^2n}} + 2.5\nu_0\sigma_xl\frac{\log(2p)}{rn} + \frac{h\log(2p)}{3r^2n} \tag{A.22}$$

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

Then, provided that $n \ge C(\nu_0\sigma_xl/\underline{f}r)^2\overline{f}h\log(2p)$ for some large constant C, we get

$$D_k(\boldsymbol{\alpha}, \boldsymbol{\beta}) \ge c\|\boldsymbol{\Delta}_k\|_\Sigma^2 \tag{A.23}$$

with probability at least $1 - (2p)^{-1}$ where $c = 0.5\underline{f}\cdot\underline{\kappa}$. Summing up these results for $k = 1\ldots, q$, we get the desired RSC property with probability at least $1 - q/(2p)$. □

A.5. Proof of Theorem 3.1

Throughout the proof, we write $\widehat{\boldsymbol{\alpha}} = \widehat{\boldsymbol{\alpha}}_h$ and $\widehat{\boldsymbol{\beta}} = \widehat{\boldsymbol{\beta}}_h$ for simplicity. To prove Theorem 3.1, we first derive an upper bound on the symmetrized Bregman divergence given in (3.8), along with a cone property for the estimator. Next, we prove a local RSC property based on Proposition 3.3, which in turns implies a lower bound on the Bregman divergence. Combining these upper and lower bounds yields the claimed estimation error bound.

Set $\widehat{\boldsymbol{\delta}} = \widehat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}_h^* \in \mathbb{R}^q$ and $\widehat{\boldsymbol{\Delta}} = \widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^* \in \mathbb{R}^p$. Conditioned on the event $\mathcal{G}(\lambda)$ defined in (3.9), we have

$$D(\widehat{\boldsymbol{\alpha}}, \widehat{\boldsymbol{\beta}}) \le \lambda(\|\widehat{\boldsymbol{\Delta}}_S\|_1 - \|\widehat{\boldsymbol{\Delta}}_{S^c}\|_1) + \frac{3\lambda}{2q}\|\widehat{\boldsymbol{\delta}}\|_1 + \frac{\lambda}{2}\|\widehat{\boldsymbol{\Delta}}\|_1 \le \frac{3\lambda}{2}\big(\|\widehat{\boldsymbol{\Delta}}_S\|_1 + q^{-1}\|\widehat{\boldsymbol{\delta}}\|_1\big). \tag{A.24}$$

Recall from Proposition 3.2 that a lower bound for $D(\boldsymbol{\alpha}, \boldsymbol{\beta})$ holds when $(\boldsymbol{\delta}^{\mathrm{T}}, \boldsymbol{\Delta}^{\mathrm{T}})^{\mathrm{T}}$ is in a cone-like set, where $\boldsymbol{\delta} = \boldsymbol{\alpha} - \boldsymbol{\alpha}_h^*$ and $\boldsymbol{\Delta} = \boldsymbol{\beta} - \boldsymbol{\beta}^*$. We thus need to show that the estimator satisfies a cone-like property (with high probability). Using the optimality of $(\widehat{\boldsymbol{\alpha}}, \widehat{\boldsymbol{\beta}})$ and the convexity of $\widehat{Q}_h$, we have

$$\begin{aligned}
0 &\geq \widehat{Q}_h(\widehat{\boldsymbol{\alpha}}, \widehat{\boldsymbol{\beta}}) - \widehat{Q}_h(\boldsymbol{\alpha}_h^*, \boldsymbol{\beta}^*) + \lambda(\|\widehat{\boldsymbol{\beta}}\|_1 - \|\boldsymbol{\beta}^*\|_1) \qquad \text{(A.25)}\\
&\geq \boldsymbol{\zeta}^{*\mathrm{T}}(\widehat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}_h^*) + \boldsymbol{\omega}^{*\mathrm{T}}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*) + \lambda(\|\widehat{\boldsymbol{\beta}}\|_1 - \|\boldsymbol{\beta}^*\|_1)\\
&\geq -\|\boldsymbol{\zeta}^*\|_\infty \|\widehat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}_h^*\|_1 - \|\boldsymbol{\omega}^*\|_\infty \|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_1 + \lambda(\|\widehat{\boldsymbol{\beta}}_{S^c} - \boldsymbol{\beta}^*_{S^c}\|_1 - \|\widehat{\boldsymbol{\beta}}_S - \boldsymbol{\beta}^*_S\|_1).
\end{aligned}$$

It follows that

$$(\lambda - \|\boldsymbol{\omega}^*\|_\infty)\|\widehat{\boldsymbol{\beta}}_{S^c} - \boldsymbol{\beta}^*_{S^c}\|_1 \leq (\lambda + \|\boldsymbol{\omega}^*\|_\infty)\|\widehat{\boldsymbol{\beta}}_S - \boldsymbol{\beta}^*_S\|_1 + \|\boldsymbol{\zeta}^*\|_\infty \|\widehat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}_h^*\|_1, \qquad \text{(A.26)}$$

which further implies

$$\|\widehat{\boldsymbol{\beta}}_{S^c} - \boldsymbol{\beta}^*_{S^c}\|_1 \leq 3\|\widehat{\boldsymbol{\beta}}_S - \boldsymbol{\beta}^*_S\|_1 + 3q^{-1/2}\|\widehat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}_h^*\|_2 \qquad \text{(A.27)}$$

conditioned on $\mathcal{G}(\lambda)$. Using the above bound and Cauchy-Schwarz inequality, we get

$$\begin{aligned}
\|(\widehat{\boldsymbol{\delta}}^{\mathrm{T}}, \widehat{\boldsymbol{\Delta}}^{\mathrm{T}})^{\mathrm{T}}\|_1 &\leq 4s^{1/2}\|\widehat{\boldsymbol{\Delta}}_{\mathcal{S}}\|_2 + \bigl(q^{1/2} + 3q^{-1/2}\bigr)\|\widehat{\boldsymbol{\delta}}\|_2\\
&\leq 4\max(s,q)^{1/2}(\|\widehat{\boldsymbol{\Delta}}_{\mathcal{S}}\|_2 + \|\widehat{\boldsymbol{\delta}}\|_2)\\
&\leq 4\sqrt{2}\cdot \max(s,q)^{1/2}\gamma_p^{-1/2}\|(\widehat{\boldsymbol{\delta}}^{\mathrm{T}}, \widehat{\boldsymbol{\Delta}}^{\mathrm{T}})^{\mathrm{T}}\|_\Omega,
\end{aligned}$$

so that $(\widehat{\boldsymbol{\delta}}^{\mathrm{T}}, \widehat{\boldsymbol{\Delta}}^{\mathrm{T}})^{\mathrm{T}} \in \mathbb{C}_\Omega(l)$ with $l = 4\gamma_p^{-1/2}\sqrt{2\cdot\max(s,q)}$, where $\mathbb{C}_\Omega(l)$ is defined in (3.12).

Note further that the RSC property only holds in a local neighborhood of $\boldsymbol{\alpha}_h^*$ and $\boldsymbol{\beta}^*$, for which $(\widehat{\boldsymbol{\alpha}}, \widehat{\boldsymbol{\beta}})$ does not necessarily satisfy. We thus employ a localized argument complemented with proof by contradiction. For some $r > 0$ to be determined, define $\eta := \sup\bigl\{u \in [0,1] : u(\widehat{\boldsymbol{\delta}}, \widehat{\boldsymbol{\Delta}}) \in \mathbb{B}_\Omega(r)\bigr\}$, where $\mathbb{B}_\Omega(r)$ is defined in (3.11). By definition, $\eta = 1$ when $(\widehat{\boldsymbol{\delta}}, \widehat{\boldsymbol{\Delta}}) \in \mathbb{B}_\Omega(r)$, and $\eta \in (0,1)$ otherwise. Then define an intermediate "estimate" $(\widetilde{\boldsymbol{\alpha}}, \widetilde{\boldsymbol{\beta}}) = (\boldsymbol{\alpha}_h^*, \boldsymbol{\beta}^*) + \eta(\widehat{\boldsymbol{\delta}}, \widehat{\boldsymbol{\Delta}})$, which satisfies (i) $(\widetilde{\boldsymbol{\alpha}}, \widetilde{\boldsymbol{\beta}}) = (\widehat{\boldsymbol{\alpha}}, \widehat{\boldsymbol{\beta}})$ if $(\widehat{\boldsymbol{\delta}}, \widehat{\boldsymbol{\Delta}}) \in \mathbb{B}_\Omega(r)$, and (ii) $(\widetilde{\boldsymbol{\alpha}}, \widetilde{\boldsymbol{\beta}})$ lies on the boundary of $(\boldsymbol{\alpha}_h^*, \boldsymbol{\beta}^*) + \mathbb{B}_\Omega(r)$ if $(\widehat{\boldsymbol{\delta}}, \widehat{\boldsymbol{\Delta}}) \notin \mathbb{B}_\Omega(r)$. Moreover, $(\widetilde{\boldsymbol{\alpha}}, \widetilde{\boldsymbol{\beta}})$ inherits the cone property of $(\widehat{\boldsymbol{\alpha}}, \widehat{\boldsymbol{\beta}})$ conditioned on $\mathcal{G}(\lambda)$. Applying Proposition 3.3, we obtain that

$$D(\widetilde{\boldsymbol{\alpha}}, \widetilde{\boldsymbol{\beta}}) \geq c\left(\|\widetilde{\boldsymbol{\Delta}}\|_\Sigma^2 + \frac{1}{q}\|\widetilde{\boldsymbol{\delta}}\|_2^2\right) \qquad \text{(A.28)}$$

with probability at least $1 - q/(2p)$ conditioned on $\mathcal{G}(\lambda)$, where $c = 0.5\underline{f}\cdot\underline{\kappa}$. On the other hand, Lemma F.2 in the supplementary material of [14] states that $D(\widetilde{\boldsymbol{\alpha}}, \widetilde{\boldsymbol{\beta}}) \leq \eta D(\widehat{\boldsymbol{\alpha}}, \widehat{\boldsymbol{\beta}})$. Combining the upper and lower bounds in (A.24) and (A.28), we obtain

$$c\left(\|\widetilde{\boldsymbol{\Delta}}\|_\Sigma^2 + \frac{1}{q}\|\widetilde{\boldsymbol{\delta}}\|_2^2\right) \leq \frac{3\lambda}{2}\left(\|\widetilde{\boldsymbol{\Delta}}_S\|_1 + \frac{1}{\sqrt{q}}\|\widetilde{\boldsymbol{\delta}}\|_2\right)$$

$$
\begin{aligned}
&\le \frac{3\lambda}{2}\left(s^{1/2}\|\widetilde{\boldsymbol{\Delta}}\|_2 + \frac{1}{\sqrt{q}}\|\widetilde{\boldsymbol{\delta}}\|_2\right) \\
&\le \frac{3}{\sqrt{2}} s^{1/2}\lambda \cdot \|(\widetilde{\boldsymbol{\delta}}^{\mathrm{T}}/\sqrt{q}, \widetilde{\boldsymbol{\Delta}}^{\mathrm{T}})\|_2 \\
&\le \frac{3}{\sqrt{2}} \gamma_p^{-1/2} s^{1/2}\lambda \cdot \|(\widetilde{\boldsymbol{\delta}}^{\mathrm{T}}/\sqrt{q}, \widetilde{\boldsymbol{\Delta}}^{\mathrm{T}})\|_\Omega .
\end{aligned}
$$

Canceling out the common factor $\|(\widetilde{\boldsymbol{\delta}}^{\mathrm{T}}/\sqrt{q}, \widetilde{\boldsymbol{\Delta}}^{\mathrm{T}})\|_\Omega$ from both sides yields

$$
\|(\widetilde{\boldsymbol{\delta}}^{\mathrm{T}}/\sqrt{q}, \widetilde{\boldsymbol{\Delta}}^{\mathrm{T}})\|_\Omega \le \frac{3}{\sqrt{2}c}\gamma_p^{-1/2} s^{1/2} = \frac{3\sqrt{2}}{\underline{f}\underline{\kappa}\gamma_p^{1/2}} s^{1/2}\lambda. \tag{A.29}
$$

In view of Proposition 3.3, we choose $r = h/(12\nu_0^2)$ so that $\|(\widetilde{\boldsymbol{\delta}}^{\mathrm{T}}, \widetilde{\boldsymbol{\Delta}}^{\mathrm{T}})\|_\Omega < r$ provided

$$
\frac{3\sqrt{2}}{\underline{f}\underline{\kappa}}\gamma_p^{-1/2}(sq)^{1/2}\lambda < \frac{h}{12\nu_0^2}.
$$

In this case, $(\widetilde{\boldsymbol{\alpha}}, \widetilde{\boldsymbol{\beta}})$ falls into the interior of $\mathbb{B}_\Omega(r)$ and we claim that $(\widehat{\boldsymbol{\delta}}, \widehat{\boldsymbol{\Delta}}) \in \mathbb{B}_\Omega(r)$ and thus $\eta = 1$. Otherwise if $(\widehat{\boldsymbol{\delta}}, \widehat{\boldsymbol{\Delta}}) \notin \mathbb{B}_\Omega(r)$, $(\widetilde{\boldsymbol{\alpha}}, \widetilde{\boldsymbol{\beta}})$ is constructed to be on the boundary of $(\boldsymbol{\alpha}_h^*, \boldsymbol{\beta}^*) + \mathbb{B}_\Omega(r)$ so that $\|(\widetilde{\boldsymbol{\delta}}^{\mathrm{T}}, \widetilde{\boldsymbol{\Delta}}^{\mathrm{T}})\|_\Omega = r$. This contradicts the above, and therefore proves the claim. The desired estimation error bound then holds on the event $\mathcal{R}(c, r, l) \cap \mathcal{G}(\lambda)$. Finally, from Propositions 3.2 and 3.3 we see that event $\mathcal{R}(c, r, l) \cap \mathcal{G}(\lambda)$ with $\lambda \asymp \nu_0\sigma_x\sqrt{\log(2p)/n}$ occurs with probability at least $1 - q/p$ as long as $n \gtrsim sq\log(p)$. □

A.6. Proof of Corollary 3.1

From (A.26), we know that $\|\widehat{\boldsymbol{\beta}}_{\mathcal{S}^c} - \boldsymbol{\beta}^*_{\mathcal{S}^c}\|_1 \le 3\|\widehat{\boldsymbol{\beta}}_{\mathcal{S}} - \boldsymbol{\beta}^*_{\mathcal{S}}\|_1 + 3q^{-1}\|\widehat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}_h^*\|_1$. Also, we have $3q^{-1}\|\widehat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}_h^*\|_1 \le 3|\widehat{\alpha}_j - \alpha^*_{h,1}|$ for some $j \in \{1, \ldots, q\}$, and assume $j = 1$ satisfy the condition without loss of generality. Let $\bar{\boldsymbol{x}}_i^{\mathrm{T}} := (q^{-1/2}, \boldsymbol{x}_i^{\mathrm{T}})$, $\bar{\boldsymbol{X}} = (\bar{\boldsymbol{x}}_1, \ldots, \bar{\boldsymbol{x}}_n)^{\mathrm{T}} \in \mathbb{R}^{n\times(p+1)}$, $\bar{\boldsymbol{\Sigma}} := \mathbb{E}\bar{\boldsymbol{x}}_i\bar{\boldsymbol{x}}_i^{\mathrm{T}}$, which notation will be only used in this proof. Moreover, let $\boldsymbol{\Psi} = \bar{\boldsymbol{X}}\bar{\boldsymbol{\Sigma}}^{-1/2}$, $\boldsymbol{A} = \bar{\boldsymbol{\Sigma}}^{1/2}$, and $\widehat{\boldsymbol{\Delta}}_1^{\mathrm{T}} := (\widehat{\alpha}_1 - \alpha^*_{h,1}, \widehat{\boldsymbol{\beta}}_h^{\mathrm{T}} - \boldsymbol{\beta}^{*\mathrm{T}})$. Then, Definition 1 in [28] holds with $s_0 = s, k_0 = 3, A = \boldsymbol{A}$, and $K(s_0, k_0, A) = \{\min(\gamma_p, 1/q)\}^{-1/2}$. Then, by using Theorem 16 of [28], we obtain that, with probability at least $1 - 2p^{-1}$,

$$
\frac{1}{\sqrt{n}}\|\boldsymbol{X}(\widehat{\boldsymbol{\beta}}_h - \boldsymbol{\beta}^*)\|_2 \le \frac{1}{\sqrt{n}}\|\bar{\boldsymbol{X}}\widehat{\boldsymbol{\Delta}}_1\|_2 \le 2\|\boldsymbol{A}\widehat{\boldsymbol{\Delta}}_1\| \le 2\left\|\begin{matrix}\frac{\widehat{\boldsymbol{\alpha}}_h - \boldsymbol{\alpha}_h^*}{\sqrt{q}} \\ \widehat{\boldsymbol{\beta}}_h - \boldsymbol{\beta}^*\end{matrix}\right\|_\Omega .
$$

Then, the result follows from Theorem 3.1.

A.7. Proof of Theorem 3.2

The idea behind this proof is that we need to control the magnitude of false discoveries at each step to refine the estimation error. The larger value of λ_j

with $j \in \mathcal{S}^{\mathrm{c}}$ tends to penalize the false discoveries harder. Throughout the proof, we write $\widehat{\boldsymbol{\alpha}}^t = \widehat{\boldsymbol{\alpha}}_h^t$ and $\widehat{\boldsymbol{\beta}}^t = \widehat{\boldsymbol{\beta}}_h^t$ for simplicity. Let us define a sequence of sets for $t \geq 1$ as follows

$$\mathcal{S}_t := \mathcal{S} \cup \{1 \leq j \leq p : \lambda_j^{t-1} = P'_\lambda(|\widehat{\beta}_j^{t-1}|) < P'(a_0)\lambda\}. \tag{A.30}$$

Each set depends on the estimator of the previous iterative step. Using above definition of the index set $\mathcal{S}_t$, we claim that

$$|\mathcal{S}_t| < (b^2+1)s, \quad \text{and} \quad \|\boldsymbol{\lambda}^{t-1}_{\mathcal{S}_t^{\mathrm{c}}}\|_{\min} \geq P'(a_0)\lambda. \tag{A.31}$$

We first assume that the above claime holds. On $\mathcal{G}(P'(a_0)\lambda)$, using (A.26), we can derive that $(\widehat{\boldsymbol{\delta}}^t, \widehat{\boldsymbol{\Delta}}^t) \in \mathbb{C}_\Omega(l)$ with $l = 4\gamma_p^{-1/2}\sqrt{2\cdot\max(s,q)}$, where $(\widehat{\boldsymbol{\delta}}^t, \widehat{\boldsymbol{\Delta}}^t) = (\widehat{\boldsymbol{\alpha}}^t - \boldsymbol{\alpha}_h^*, \widehat{\boldsymbol{\beta}}^t - \boldsymbol{\beta}^*)$, and $\mathbb{C}_\Omega(l)$ is defined in (3.12). Consider the symmetrized Bregmann divergence

$$D(\widehat{\boldsymbol{\alpha}}^t, \widehat{\boldsymbol{\beta}}^t) = \langle -\boldsymbol{\lambda} \circ \widehat{\boldsymbol{g}}, \widehat{\boldsymbol{\Delta}}^t\rangle + \langle -\boldsymbol{\zeta}^*, \widehat{\boldsymbol{\delta}}^t\rangle + \langle -\boldsymbol{\omega}^*, \widehat{\boldsymbol{\Delta}}^t\rangle,$$

where $\widehat{\boldsymbol{g}} \in \partial\|\widehat{\boldsymbol{\beta}}^t\|_1$. For the first term of the right-hand side of the equality above, we split it into three parts, which leads to

$$\begin{aligned}\langle \boldsymbol{\lambda} \circ \boldsymbol{g}, \widehat{\boldsymbol{\Delta}}^t\rangle &= \langle (\boldsymbol{\lambda} \circ \boldsymbol{g})_{\mathcal{S}}, \widehat{\boldsymbol{\Delta}}^t_{\mathcal{S}}\rangle + \langle (\boldsymbol{\lambda} \circ \boldsymbol{g})_{\mathcal{S}_t \setminus \mathcal{S}}, \widehat{\boldsymbol{\Delta}}^t_{\mathcal{S}_t \setminus \mathcal{S}}\rangle + \langle (\boldsymbol{\lambda} \circ \boldsymbol{g})_{\mathcal{S}_t^{\mathrm{c}}}, \widehat{\boldsymbol{\Delta}}^t_{\mathcal{S}_t^{\mathrm{c}}}\rangle \\ &\geq -\|\boldsymbol{\lambda}_{\mathcal{S}}\|_2\|\widehat{\boldsymbol{\Delta}}_{\mathcal{S}}\|_2 + 0 + \|\boldsymbol{\lambda}_{\mathcal{S}_t^{\mathrm{c}}}\|_{\min}\|\widehat{\boldsymbol{\Delta}}^t_{\mathcal{S}^{\mathrm{c}}}\|_1.\end{aligned}$$

The inequality above is derived using $\boldsymbol{\beta}^*_{\mathcal{S}^{\mathrm{c}}} = 0$ and the property of subdifferential. Then, combining above result with Hölder's inequality, we get the following upper bound of the symmetrized Bregmann divergence

$$\begin{aligned}D(\widehat{\boldsymbol{\alpha}}^t, \widehat{\boldsymbol{\beta}}^t) &\leq \|\boldsymbol{\zeta}^*\|_2\|\widehat{\boldsymbol{\delta}}^t\|_2 + \|\boldsymbol{\omega}^*_{\mathcal{S}_t}\|_2\|\widehat{\boldsymbol{\Delta}}^t_{\mathcal{S}_t}\|_2 + \|\boldsymbol{\lambda}_{\mathcal{S}}\|_2\|\widehat{\boldsymbol{\Delta}}_{\mathcal{S}}\|_2 \\ &\quad + (\|\boldsymbol{\omega}^*_{\mathcal{S}_t^{\mathrm{c}}}\|_\infty - \|\boldsymbol{\lambda}_{\mathcal{S}_t^{\mathrm{c}}}\|_{\min})\|\widehat{\boldsymbol{\Delta}}^t_{\mathcal{S}^{\mathrm{c}}}\|_1 \\ &\leq \|\boldsymbol{\zeta}^*\|_2\|\widehat{\boldsymbol{\delta}}^t\|_2 + \|\boldsymbol{\omega}^*_{\mathcal{S}_t}\|_2\|\widehat{\boldsymbol{\Delta}}^t_{\mathcal{S}_t}\|_2 + \|\boldsymbol{\lambda}_{\mathcal{S}}\|_2\|\widehat{\boldsymbol{\Delta}}_{\mathcal{S}}\|_2 \qquad \text{(A.32)} \\ &\leq q\|\boldsymbol{\zeta}^*\|_\infty\|\widehat{\boldsymbol{\delta}}^t/\sqrt{q}\|_2 + (0.5P'(a_0)s^{1/2}\sqrt{b^2+1} + s^{1/2})\lambda\|\widehat{\boldsymbol{\Delta}}^t\|_2 \\ &< \gamma_p^{-1/2}s^{1/2}\lambda\big(P'(a_0)\sqrt{(b^2+1)/2} + 2\big)\|\boldsymbol{\theta}^t\|_\Omega = c \cdot r_{\mathrm{opt}}\|\boldsymbol{\theta}^t\|_\Omega. \qquad \text{(A.33)}\end{aligned}$$

As in the proof of Theorem 3.1, define an intermediate vector $(\widetilde{\boldsymbol{\alpha}}^t, \widetilde{\boldsymbol{\beta}}^t) := (\boldsymbol{\alpha}_h^*, \boldsymbol{\beta}^*) + \eta(\widehat{\boldsymbol{\delta}}^t, \widehat{\boldsymbol{\Delta}}^t)$ with $\eta := \sup\big\{u \in [0,1] : u(\widehat{\boldsymbol{\delta}}^t, \widehat{\boldsymbol{\Delta}}^t) \in \mathbb{B}_\Omega(r)\big\}$, where $\mathbb{B}_\Omega(r)$ is defined in (3.11). On event $\mathcal{R}(c,r,l) \cap \mathcal{G}(P'(a_0)\lambda)$, we can ensure that $\eta = 1$, since $D(\widetilde{\boldsymbol{\alpha}}^t, \widetilde{\boldsymbol{\beta}}^t) \leq \eta D(\widehat{\boldsymbol{\alpha}}^t, \widehat{\boldsymbol{\beta}}^t)$ from Lemma F.2 in the supplementary material of [14] combined with the RSC property gives $\|\widetilde{\boldsymbol{\theta}}^t\|_\Omega < r_{\mathrm{opt}}$, which implies $(\widetilde{\boldsymbol{\delta}}^t, \widetilde{\boldsymbol{\Delta}}^t) \in \mathbb{B}_\Omega(r)$, thus ensuring $\eta = 1$ via proof by contradiction.

Now, we need to verify the claim (A.31). For the second part of the claim, it holds trivially for $t = 1$. Assume that it holds for $1, \ldots, t$. Using the definition of the index set, for $j \in \mathcal{S}^{\mathrm{c}}_{t+1}$, we have $\lambda_j^t \geq P'(a_0)\lambda$, which verifies the second part of (A.31). For the first part of (A.31), since $\mathcal{S}_1 = \mathcal{S}$, it holds for $t = 1$ trivially. Suppose it holds for some $t \geq 1$. Then, we get $P'_\lambda(|\widehat{\beta}^t_j|) < P'(a_0)\lambda = P'_\lambda(a_0\lambda)$ for $j \in \mathcal{S}_{t+1} \setminus \mathcal{S}$, which implies that $|\widehat{\beta}^t_j| > a_0\lambda$ due to the monotonicity of $P'(\cdot)$. Thus, we get an upper bound for the size of the set as follows.

$$|\mathcal{S}_{t+1} \setminus \mathcal{S}|^{1/2} < (a_0\lambda)^{-1}\|(\widehat{\boldsymbol{\beta}}^t - \boldsymbol{\beta}^*)_{\mathcal{S}_{t+1}\setminus\mathcal{S}}\|_2 \leq (a_0\lambda)^{-1}\gamma_p^{-1/2} r_{\mathrm{opt}} = bs^{1/2}. \tag{A.34}$$

Hence, we get $|\mathcal{S}_{t+1}| = |\mathcal{S}| + |\mathcal{S}_{t+1} \setminus \mathcal{S}| < s + b^2 s$, which verifies the first part of the claim (A.31).

To refine the rate at each step, we need to control the terms in (A.32). For each j, we consider two cases, where the first case is when $|\widehat{\beta}^{t-1}_j - \beta^*_j| \geq a_0\lambda$ which gives $a_0^{-1}|\widehat{\beta}^{t-1}_j - \beta^*_j| \geq \lambda \geq \lambda^{t-1}_j$, and the other case is when $|\widehat{\beta}^{t-1}_j - \beta^*_j| < a_0\lambda$ which gives $\lambda^{t-1}_j \leq P'_\lambda((|\beta^*_j| - a_0\lambda)_+)$ due to the monotonicity of $P'(\cdot)$ and the fact that $|\beta^*_j| - a_0\lambda < |\widehat{\beta}^{t-1}_j|$ using the triangle inequality. Then, we get following bounds

$$\|\boldsymbol{\lambda}_{\mathcal{S}}\|_2 \leq \|P'_\lambda((|\beta^*_{\mathcal{S}}| - a_0\lambda)_+)\|_2 + a_0^{-1}\|\widehat{\boldsymbol{\Delta}}^{t-1}_{\mathcal{S}}\|_2, \tag{A.35}$$

and

$$\begin{aligned}\|\boldsymbol{\omega}^*_{\mathcal{S}_t}\|_2 &\leq \|\boldsymbol{\omega}^*_{\mathcal{S}}\|_2 + |\mathcal{S}_t \setminus \mathcal{S}|^{1/2}\|\boldsymbol{\omega}^*_{\mathcal{S}^c}\|_\infty \\ &\leq \|\boldsymbol{\omega}^*_{\mathcal{S}}\|_2 + (a_0\lambda)^{-1}\|\boldsymbol{\omega}^*_{\mathcal{S}^c}\|_\infty\|\widehat{\boldsymbol{\Delta}}^{t-1}_{\mathcal{S}_t\setminus\mathcal{S}}\|_2 \\ &\leq \|\boldsymbol{\omega}^*_{\mathcal{S}}\|_2 + \frac{P'(a_0)}{2a_0}\|\widehat{\boldsymbol{\Delta}}^{t-1}_{\mathcal{S}_t\setminus\mathcal{S}}\|_2. \end{aligned} \tag{A.36}$$

Substituting above results into (A.32), we get

$$D(\widehat{\boldsymbol{\alpha}}^t_h, \widehat{\boldsymbol{\beta}}^t_h) \leq \left\{\sqrt{q}\|\boldsymbol{\zeta}^*\|_2 + c \cdot r_{\mathrm{ora}} + \gamma_p^{-1/2} a_0^{-1}\sqrt{1 + \{P'(a_0)/2\}^2}\|\widehat{\boldsymbol{\Delta}}^{t-1}_{\mathcal{S}_t}\|_2\right\}\|\boldsymbol{\theta}^t\|_\Omega.$$

Then, combining with the RSC property, we get

$$\|\boldsymbol{\theta}^t\|_\Omega \leq \delta \cdot \|\boldsymbol{\theta}^{t-1}\|_\Omega + r_{\mathrm{ora}} + c^{-1}\sqrt{q}\|\boldsymbol{\zeta}^*\|_2, \tag{A.37}$$

where $\delta = \sqrt{1 + \{P'(a_0)/2\}^2}/(ca_0\gamma_p) \in (0, 1)$. □

A.8. Proof of Theorem 3.3

Let $\lambda = 8P'(a_0)^{-1}\nu_0\sigma_x\sqrt{\log(2p)/n}$, and $\lambda_1 = P'(a_0)^{-1}\sqrt{\{s + \log(q) + z\}/n} \leq \lambda$. If the inequality does not hold, then let $\lambda = P'(a_0)^{-1}\sqrt{\{s + \log(q) + z\}/n}$.

Then, with slight modification of the proof of Proposition 3.2, the event $\{\|\boldsymbol{\zeta}^*\|_\infty \leq 3P'(a_0)\lambda_1/(2q), \|\boldsymbol{\omega}^*\|_\infty \leq P'(a_0)\lambda/2\} \subset \mathcal{G}(P'(a_0)\lambda)$ holds with probability at least $1 - q/(2p) - e^{-2(s+z)}$.

On $\mathcal{G}(P'(a_0)\lambda)$, we can use the results from Proposition 3.3 and Theorem 3.2. Consider $c = 0.5\underline{f}\cdot\underline{\kappa}$, $l = 4\gamma_p^{-1/2}\sqrt{2\cdot\max(s,q)}$ and b such that satisfies

$$\sqrt{2}P'(a_0)(b^2+1)^{1/2} + 4 = a_0\underline{\kappa}\underline{f}\gamma_p b.$$

Also, let $r \geq q^{1/2}r_{\mathrm{opt}} = a_0 b(\gamma_p sq)^{1/2}\lambda$. Then, using Proposition 3.3 with proper conditions of sample size and the smoothing parameter decribed therein, the event $\mathcal{R}(c,r,l)$ holds with probability at least $1 - q/(2p)$. With above preparations, we get the following estimation error bound using Theorem 3.2 on event $\{\|\boldsymbol{\zeta}^*\|_\infty \leq 3P'(a_0)\lambda_1/(2q), \|\boldsymbol{\omega}^*\|_\infty \leq P'(a_0)\lambda/2\} \cap \mathcal{R}(c,r,l) \subset \mathcal{G}(P'(a_0)) \cap \mathcal{R}(c,r,l)$,

$$\|\boldsymbol{\theta}^t\|_\Omega \leq \delta^{t-1}r_{\mathrm{opt}} + (1-\delta)^{-1}(r_{\mathrm{ora}} + c^{-1}\sqrt{q}\|\boldsymbol{\zeta}^*\|_2) \tag{A.38}$$

where $r_{\mathrm{ora}} = c^{-1}\big\{\gamma_p^{-1/2}\|P'_\lambda((|\boldsymbol{\beta}^*_{\mathcal{S}}| - a_0\lambda)_+)\|_2 + \|\boldsymbol{\omega}^*_{\mathcal{S}}\|_2\big\}$ is defined in Theorem 3.2. Since we have $t \gtrsim \log\{\log(2p)\}/\log(1/\delta)$, we get $\delta^{t-1}r_{\mathrm{opt}} \lesssim \sqrt{s/n}$.

Remaining quantity to bound is $r_{\mathrm{ora}} + c^{-1}\sqrt{q}\|\boldsymbol{\zeta}^*\|_2$ from (A.38). The second term $c^{-1}\sqrt{q}\|\boldsymbol{\zeta}^*\|_2$ is bounded by $(3/2)P'(a_0)^{-1}c^{-1}\sqrt{\{s+\log(q)+z\}/n}$ on the event $\{\|\boldsymbol{\zeta}^*\|_\infty \leq 3P'(a_0)\lambda_1/(2q), \|\boldsymbol{\omega}^*\|_\infty \leq P'(a_0)\lambda/2\} \subset \mathcal{G}(P'(a_0)\lambda)$. By using the beta-min condition, the shrinkage bias term $\|P'_\lambda((|\boldsymbol{\beta}^*_{\mathcal{S}}| - a_0\lambda)_+)\|_2$ vanishes. The only term remaining to bound is $\|\boldsymbol{\omega}^*_{\mathcal{S}}\|_2$.

Consider

$$\|\boldsymbol{S}^{-1/2}\boldsymbol{\omega}^*_{\mathcal{S}}\|_2 = \left\|\frac{1}{nq}\sum_{i=1}^n\sum_{k=1}^q\big\{\bar{K}_h(\alpha^*_{h,k} - \varepsilon_i) - \tau_k\big\}\boldsymbol{S}^{-1}\boldsymbol{x}_{i,\mathcal{S}}\right\|_2, \tag{A.39}$$

where $\boldsymbol{S} := \mathbb{E}(\boldsymbol{x}_{\mathcal{S}}\boldsymbol{x}_{\mathcal{S}}^{\mathrm{T}})$. Since we have $|(\boldsymbol{S}^{-1/2}\boldsymbol{\omega}^*_{\mathcal{S}})_j| = |q^{-1}\sum_{k=1}^q\big\{\bar{K}_h(\alpha^*_{h,k} - \varepsilon_i) - \tau_k\big\}(\boldsymbol{S}^{-1/2}\boldsymbol{x}_{i,\mathcal{S}})_j| \leq |(\boldsymbol{S}^{-1/2}\boldsymbol{x}_{i,\mathcal{S}})_j|$, we get $\|\boldsymbol{S}^{-1/2}\boldsymbol{\omega}^*_{\mathcal{S}}\|_2 \leq \|\frac{1}{n}\sum_{i=1}^n\boldsymbol{S}^{-1/2}\boldsymbol{x}_{i,\mathcal{S}}\|_2$. Then, since each $\boldsymbol{S}^{-1/2}\boldsymbol{x}_{i,\mathcal{S}}$ is an s-dimensional sub-Gaussian random vector with parameter $2\sqrt{2}\nu_0\sigma_x$, we get

$$\|\boldsymbol{S}^{-1/2}\boldsymbol{\omega}^*_{\mathcal{S}}\|_2 \leq 8\nu_0\sigma_x\sqrt{\frac{2s}{n}} + 4\nu_0\sigma_x\sqrt{\frac{2\log(1/\epsilon)}{n}} \tag{A.40}$$

with probability at least $1-\epsilon$. By letting $\epsilon = e^{-(s+z)}$ and combining all the bounds we have for (3.19), we get the desired bound with probability at least $1 - q/p - 2e^{-(s+z)}$. Seperate bounds comes from $\|\widehat{\boldsymbol{\beta}}^t_h - \boldsymbol{\beta}^*\|_\Sigma$, $\|(\widehat{\boldsymbol{\alpha}}^t_h - \boldsymbol{\alpha}^*_h)/\sqrt{q}\|_2 \leq \|\boldsymbol{\theta}^t\|_\Omega$. Finally, the h^2 bias term comes from $\|\boldsymbol{\alpha}^*_h - \boldsymbol{\alpha}^*\|_2$, which was derived in the Proposition 3.1. □

A.9. Proof of Theorem 3.4

For $t = 1, 2, \ldots$, let $\mathcal{T}_t = \mathcal{S} \cup \{1 \leq j \leq p : \lambda_j^{t-1} = P'_\lambda(|\widehat{\beta}_j^{t-1}|) < P'(a_0)\lambda\}$, and $k = |\mathcal{T}_t|$. By using the optimality we get

$$0 = \left\langle \nabla\widehat{Q}_h(\widehat{\boldsymbol{\alpha}}^t, \widehat{\boldsymbol{\beta}}^t) + \lambda \circ \widehat{\boldsymbol{g}}, \begin{bmatrix}\widehat{\boldsymbol{\delta}}^o \\ \widehat{\boldsymbol{\Delta}}^o\end{bmatrix} \right\rangle$$

$$= \left\langle \nabla\widehat{Q}_h(\widehat{\boldsymbol{\alpha}}^t, \widehat{\boldsymbol{\beta}}^t) - \nabla\widehat{Q}_h(\widehat{\boldsymbol{\alpha}}^o, \widehat{\boldsymbol{\beta}}^o), \begin{bmatrix} \widehat{\boldsymbol{\delta}}^o \\ \widehat{\boldsymbol{\Delta}}^o \end{bmatrix} \right\rangle + \left\langle \lambda \circ \widehat{\boldsymbol{g}}, \begin{bmatrix} \widehat{\boldsymbol{\delta}}^o \\ \widehat{\boldsymbol{\Delta}}^o \end{bmatrix} \right\rangle$$
$$+ \left\langle \nabla\widehat{Q}_h(\widehat{\boldsymbol{\alpha}}^o, \widehat{\boldsymbol{\beta}}^o), \begin{bmatrix} \widehat{\boldsymbol{\delta}}^o \\ \widehat{\boldsymbol{\Delta}}^o \end{bmatrix} \right\rangle$$
$$\geq -\|\boldsymbol{\zeta}^o\|_\infty \|\widehat{\boldsymbol{\delta}}^o\|_1 - \|\boldsymbol{\omega}^o\|_\infty \|\widehat{\boldsymbol{\Delta}}^o\|_1 + \|\lambda_{\mathcal{T}_t^c}\|_{\min} \|\widehat{\boldsymbol{\Delta}}^o_{\mathcal{T}^c}\|_1 - \|\lambda_{\mathcal{T}_t}\|_\infty \|\widehat{\boldsymbol{\Delta}}^o_{\mathcal{T}}\|_1,$$

where $\widehat{\boldsymbol{\delta}}^o = \widehat{\boldsymbol{\alpha}}^t - \widehat{\boldsymbol{\alpha}}^o$, $\widehat{\boldsymbol{\Delta}}^o = \widehat{\boldsymbol{\beta}}^t - \widehat{\boldsymbol{\beta}}^o$, $\boldsymbol{\zeta}^o = \nabla_{\boldsymbol{\alpha}} \widehat{Q}_h(\widehat{\boldsymbol{\alpha}}^o, \widehat{\boldsymbol{\beta}}^o)$, and $\boldsymbol{\omega}^o = \nabla_{\boldsymbol{\beta}} \widehat{Q}_h(\widehat{\boldsymbol{\alpha}}^o, \widehat{\boldsymbol{\beta}}^o)$. By rearranging terms and using the optimality ($\|\boldsymbol{\zeta}^o\|_\infty = 0$), we get

$$(\|\lambda_{\mathcal{T}_t^c}\|_{\min} - \|\boldsymbol{\omega}^o\|_\infty) \|\widehat{\boldsymbol{\Delta}}^o_{\mathcal{T}_t^c}\|_1 \leq (\|\boldsymbol{\omega}^o\|_\infty + \|\lambda_{\mathcal{T}_t}\|_\infty) \|\widehat{\boldsymbol{\Delta}}^o_{\mathcal{T}_t}\|_1, \tag{A.41}$$

which leads to

$$\|\widehat{\boldsymbol{\Delta}}^o_{\mathcal{T}_t^c}\|_1 \leq \{1 + 2/P'(a_0)\} \|\widehat{\boldsymbol{\Delta}}^o_{\mathcal{T}_t}\|_1,$$

thus

$$\begin{aligned} \left\| \begin{matrix} \widehat{\boldsymbol{\delta}}^o \\ \widehat{\boldsymbol{\Delta}}^o \end{matrix} \right\|_1 &\leq \{2 + 2/P'(a_0)\} (\|\widehat{\boldsymbol{\delta}}^o\|_1 + \|\widehat{\boldsymbol{\Delta}}^o_{\mathcal{T}_t}\|_1) \\ &\leq \{2 + 2/P'(a_0)\} \left\{ \frac{2\max(q,k)}{\gamma_p} \right\}^{1/2} \left\| \begin{matrix} \widehat{\boldsymbol{\delta}}^o \\ \widehat{\boldsymbol{\Delta}}^o \end{matrix} \right\|_\Omega, \end{aligned} \tag{A.42}$$

which explains the choice of l in the theorem.

Now, using the optimality and properties of subdifferential, we get

$$\left\langle \nabla\widehat{Q}_h(\widehat{\boldsymbol{\alpha}}^t, \widehat{\boldsymbol{\beta}}^t) - \nabla\widehat{Q}_h(\widehat{\boldsymbol{\alpha}}^o, \widehat{\boldsymbol{\beta}}^o), \begin{bmatrix} \widehat{\boldsymbol{\delta}}^o \\ \widehat{\boldsymbol{\Delta}}^o \end{bmatrix} \right\rangle = \left\langle -\lambda \circ \widehat{\boldsymbol{g}} - \nabla\widehat{Q}_h(\widehat{\boldsymbol{\alpha}}^o, \widehat{\boldsymbol{\beta}}^o), \begin{bmatrix} \widehat{\boldsymbol{\delta}}^o \\ \widehat{\boldsymbol{\Delta}}^o \end{bmatrix} \right\rangle$$
$$\leq \|\boldsymbol{\omega}^o_{\mathcal{T}_t}\|_2 \|\widehat{\boldsymbol{\Delta}}^o_{\mathcal{T}_t}\|_2 - (\|\boldsymbol{\lambda}_{\mathcal{T}_t^c}\|_{\min} - \|\boldsymbol{\omega}_{\mathcal{T}_t^c}\|_\infty) \|\widehat{\boldsymbol{\Delta}}^o_{\mathcal{T}_t^c}\|_1 + \|\boldsymbol{\lambda}_{\mathcal{S}}\|_2 \|\widehat{\boldsymbol{\Delta}}^o_{\mathcal{S}}\|_2 \tag{A.43}$$
$$\leq (s^{1/2} + 0.5P'(a_0)k^{1/2})\lambda \|\widehat{\boldsymbol{\Delta}}^o\|_2 \leq \sqrt{2}(s^{1/2} + 0.5P'(a_0)k^{1/2})\lambda\gamma_p^{-1/2} \|\widehat{\boldsymbol{\theta}}^{ot}\|_\Omega \tag{A.44}$$

Using the similar argument given in the proof of Theorem 3.2, we can get $|\mathcal{T}_t| \leq (1+b^2)s$ and $q^{1/2}\|\widehat{\boldsymbol{\theta}}^{ot}\|_\Omega < r$, which makes

$$\left\| \begin{matrix} \widehat{\boldsymbol{\alpha}}^t - \widehat{\boldsymbol{\alpha}}^o \\ \widehat{\boldsymbol{\beta}}^t - \widehat{\boldsymbol{\beta}}^o \end{matrix} \right\|_\Omega \leq r.$$

Now, define $\mathcal{S}_t := \{1 \leq j \leq p : |\widehat{\beta}^t_j - \beta^*_j| > a_0\lambda\}$, which makes $\mathcal{S}_0 = \mathcal{S}$. We have

$$\lambda_j^{t-1} = P'_\lambda\big(|\widehat{\beta}_j^{t-1}|\big) \leq P'_\lambda(|\beta_j^*| - a_0\lambda)$$

if $j \in \mathcal{S} \cap \mathcal{S}^c_{t-1}$., and $\lambda_j^{t-1} \leq \lambda$ for remaining j. Thus, we get

$$\|\boldsymbol{\lambda}^{t-1}_{\mathcal{S}}\|_2 \leq \|P'_\lambda(|\boldsymbol{\beta}^*_{\mathcal{S}}| - a_0\lambda)\|_2 + \lambda|\mathcal{S} \cap \mathcal{S}_{t-1}|^{1/2} = \lambda|\mathcal{S} \cap \mathcal{S}_{t-1}|^{1/2}. \tag{A.45}$$

For each $j \in \mathcal{T}_t \setminus \mathcal{S}$, $\beta_j^* = 0$ and $\lambda_j^{t-1} = P'_\lambda\big(|\widehat{\beta}_j^{t-1}|\big) < P'_\lambda(a_0\lambda)$, which leads to $|\widehat{\beta}_j^{t-1} - \beta_j^*| > a_0\lambda$, thus we get $\mathcal{T}_t \setminus \mathcal{S} \subseteq \mathcal{S}_{t-1} \setminus \mathcal{S}$. Then, we get $\|\boldsymbol{\omega}^o_{\mathcal{T}_t}\|_2 \leq \|\boldsymbol{\omega}^o\|_\infty |\mathcal{T}_t \setminus \mathcal{S}|^{1/2} \leq \|\boldsymbol{\omega}^o\|_\infty |\mathcal{S}_{t-1} \setminus \mathcal{S}|^{1/2}$ since $\boldsymbol{\omega}^o_{\mathcal{S}} = \mathbf{0}$. Then, by combining above results with (A.43), we get

$$c\|\boldsymbol{\theta}^{ot}\|_\Omega^2 \leq \{\|\boldsymbol{\omega}^o\|_\infty |\mathcal{S}_{t-1} \setminus \mathcal{S}|^{1/2} + \lambda|\mathcal{S} \cap \mathcal{S}_{t-1}|^{1/2}\}\gamma_p^{-1/2}\|\boldsymbol{\theta}^{ot}\|_\Omega, \tag{A.46}$$

which leads to

$$\|\widehat{\boldsymbol{\beta}}^t - \widehat{\boldsymbol{\beta}}^o\|_2 \leq \gamma_p^{-1/2}\|\boldsymbol{\theta}^{ot}\|_\Omega \leq \frac{\sqrt{1 + \{P'(a_0)/2\}^2}}{c\gamma_p}|\mathcal{S}_{t-1}|^{1/2}\lambda. \tag{A.47}$$

By the definition of $\mathcal{S}_t$, we have $\min_{j \in \mathcal{S}_t} |\widehat{\beta}_j^t - \widehat{\beta}_j^o| > a_0\lambda - \|\widehat{\boldsymbol{\beta}}^o - \boldsymbol{\beta}^*\|_\infty$. Thus, provided that

$$\left\{\|\widehat{\boldsymbol{\beta}}^o - \boldsymbol{\beta}^*\|_\infty \leq \left[a_0 - \frac{\sqrt{1 + \{P'(a_0)/2\}^2}}{\delta c\gamma_p}\right]\lambda\right\},$$

we have

$$|\mathcal{S}_t|^{1/2} < \frac{\|(\widehat{\boldsymbol{\beta}}^t - \widehat{\boldsymbol{\beta}}^o)_{\mathcal{S}_t}\|_2}{a_0\lambda - \|\widehat{\boldsymbol{\beta}}^o - \boldsymbol{\beta}^*\|_\infty} \leq \delta|\mathcal{S}_{t-1}|^{1/2},$$

which completes the proof.

A.10. Proof of Proposition 3.4

We need to first establish the estimation error bound of the oracle estimator, which is essentially the estimation error bound of low dimensional smoothed CQR estimator. In this proof, with abuse of notations, use same notaion we used for the high-dmensional estiamtion error bound, except for the dimension which is now $s \ll n$. Proof strategy is similar to the proof of Theorem 3.1, but now it is low dimensional, so that some steps can be omitted. We establish upper and lower bounds of $D(\boldsymbol{\alpha}, \boldsymbol{\beta})$ in low dimension. Here, let

$$\mathcal{R}(c, r) := \left\{D(\boldsymbol{\alpha}, \boldsymbol{\beta}) \geq c\big(\|\boldsymbol{\Delta}\|_\Sigma^2 + q^{-1}\|\boldsymbol{\delta}\|_2^2\big) \text{ for all } \begin{bmatrix}\boldsymbol{\delta} \\ \boldsymbol{\Delta}\end{bmatrix} \in \mathbb{B}_\Omega(r)\right\}.$$

We first prove that $\mathcal{R}(c, r)$ holds with high probability. We follow the proof of Proposition 3.3 until (A.18), with slight modification that we no longer require $\mathbb{C}_\Omega(l)$, thus taking supremum only on $\|\boldsymbol{\Delta}_k\|_{\bar{\Sigma}} = r$. Thus, just denote $Z_n(l)$ as Z_n in this proof. Then, we need to bound $\mathbb{E}(Z')$ to establish the RSC property. Using Rademacher symmetrization and Talagrand's contraction principle on (A.16), we get

$$\mathbb{E}(Z') \leq \frac{1}{r} \cdot \mathbb{E}\left\|\frac{1}{n}\sum_{i=1}^n e_i \chi_{i,k} \bar{\boldsymbol{x}}_i\right\|_2 \leq \bar{f}^{1/2}\sqrt{\frac{hs}{r^2 n}}. \tag{A.48}$$

Thus, as long as $n \geq C\overline{f}hs/r^2$ for sufficiently large $C > 0$, we get the desired RSC property with probability at least $1 - qe^{-(s+t)}$. Next, we need to get an upper bound of $D(\boldsymbol{\alpha}, \boldsymbol{\beta})$. Using the optimality, we get $D(\widehat{\boldsymbol{\alpha}}^o, \widehat{\boldsymbol{\beta}}^o) \leq \|\boldsymbol{\zeta}^*\|_\infty \|\widehat{\boldsymbol{\alpha}}^o - \boldsymbol{\alpha}_h^*\|_1 + \|\Sigma^{-1/2}\boldsymbol{\omega}^*\|_2 \|\widehat{\boldsymbol{\beta}}^o - \boldsymbol{\beta}^*\|_\Sigma$. Then, using the result of the Proposition 3.2, we have $\|\boldsymbol{\zeta}^*\|_\infty \|\widehat{\boldsymbol{\alpha}}^o - \boldsymbol{\alpha}_h^*\|_1 \leq 2\lambda \|(\widehat{\boldsymbol{\alpha}}^o - \boldsymbol{\alpha}_h^*)/q^{1/2}\|_2$ with probability at least $1 - 2qe^{-8n\lambda^2}$. Setting $\lambda = \sqrt{(s+t)/n}$ gives the desired probability bound. Now it remains to bound $\|\Sigma^{-1/2}\boldsymbol{\omega}^*\|_2$. Let $\xi_i := q^{-1}\sum_{k=1}^q \{\bar{K}((\alpha_{h,k}^* - \varepsilon_i)/h) - \tau_k\}$. We have $|\xi_i| \leq 1$. Then, $\|\Sigma^{-1/2}\boldsymbol{\omega}^*\|_2 = \|(1/n)\sum_{i=1}^n \xi_i \boldsymbol{w}_i\|_2$, where $\boldsymbol{w}_i = \Sigma^{-1/2}\boldsymbol{x}_i$. Using a covering argument, for any $\epsilon \in (0,1)$, there exist an ϵ-net $\mathcal{N}_\epsilon$ of the unit sphere with cardinality $|\mathcal{N}_\epsilon| \leq (1 + 2/\epsilon)^s$ such that

$$\|\Sigma^{-1/2}\boldsymbol{\omega}^*\|_2 \leq (1-\epsilon)^{-1} \max_{\boldsymbol{u} \in \mathcal{N}_\epsilon} \left\langle \boldsymbol{u}, \frac{1}{n}\sum_{i=1}^n \xi_i \boldsymbol{w}_i \right\rangle. \tag{A.49}$$

Then, we have, for $k = 2, 3, \ldots$

$$\begin{aligned} \mathbb{E}\big(|\langle \boldsymbol{u}, \xi_i \boldsymbol{w}_i \rangle|^k\big) &\leq \mathbb{E}|\langle \boldsymbol{u}, \boldsymbol{w}_i \rangle|^k \\ &\leq \nu_0^k \int_0^\infty \mathbb{P}(|\langle \boldsymbol{u}, \boldsymbol{w}_i \rangle| \geq \nu_0 t) k t^{k-1} \mathrm{d}t \\ &\leq \nu_0 k \int_0^\infty t^{k-1} e^{-t} dt \\ &\leq \frac{k!}{2} \nu_0^2 (2\nu_0)^{k-2}. \end{aligned}$$

Now, using Bernstein's inequality and applying union bound over $\mathcal{N}_\epsilon$, we get

$$\|\Sigma^{-1/2}\boldsymbol{\omega}^*\|_2 \leq \frac{\nu_0}{1-\epsilon}\left(\sqrt{\frac{2u}{n}} + \frac{2u}{n}\right) \tag{A.50}$$

with probability at least $1 - e^{\log(1+2/\epsilon)s - u}$. Choosing $\epsilon = 2/(e^2 - 1)$ ahaend $u = 2s + t$ gives the desired upper bound with probability at least $1 - e^{-t}$. Thus, following the similar argument used in the Theorem 3.1 to combine the lower and upper bounds of the Bregmann divergence, we get the desired estimation error bound for the oracle estimator. For the Bahadur representation part, we refer to the Theorem 2 of [41].

A.11. Proof of Proposition 3.5

We restrict our focus on the symmetrized Bregmann divergence with each quantile index $k = 1, \ldots, q$, by letting

$$\begin{aligned} &D_{rsc}^k(\boldsymbol{\alpha}_1, \boldsymbol{\beta}_1, \boldsymbol{\alpha}_2, \boldsymbol{\beta}_2) \\ &:= \frac{1}{n}\sum_{i=1}^n \left\{ \bar{K}\left(\frac{\alpha_{1k} - r_i(\boldsymbol{\beta}_1)}{h}\right) - \bar{K}\left(\frac{\alpha_{2k} - r_i(\boldsymbol{\beta}_2)}{h}\right) \right\} \langle \bar{\boldsymbol{x}}_i, \boldsymbol{\Delta}_k \rangle, \end{aligned}$$

where $\boldsymbol{\Delta}_k := (\alpha_{1k} - \alpha_{2k}, \boldsymbol{\beta}_1^{\mathrm{T}} - \boldsymbol{\beta}_2^{\mathrm{T}})^{\mathrm{T}}$. Then, we have

$$D_{rsc}^k(\boldsymbol{\alpha}_1, \boldsymbol{\beta}_1, \boldsymbol{\alpha}_2, \boldsymbol{\beta}_2) \geq \frac{\underline{\kappa}}{nh} \sum_{i=1}^n \langle \bar{\boldsymbol{x}}_i, \boldsymbol{\Delta}_k \rangle^2 I_{E_{i,k}} \tag{A.51}$$

where

$$E_{i,k} = \{|\varepsilon_i - \alpha_{h,k}^*| \leq h/2\} \cap \{|\bar{\boldsymbol{x}}_i{}^{\mathrm{T}} \boldsymbol{\Delta}_k^*| \leq h/4\} \cap \{|\langle \boldsymbol{x}_i, \boldsymbol{\Delta}_k \rangle| / \|\boldsymbol{\Delta}_k\|_{\bar{\Sigma}} \leq h/(4r)\},$$

$$\boldsymbol{\Delta}_k^* := \begin{pmatrix} \alpha_{2k} - \alpha_{h,k}^* \\ \boldsymbol{\beta}_2 - \boldsymbol{\beta}^* \end{pmatrix}. \tag{A.52}$$

In addition to φ_R, let

$$\phi_R(u) := I(|u| < R/2) + 2\{1 - |u|/R\} I(R/2 \leq |u| \leq R).$$

Then, we can further lower bound (A.51) by

$$D_{rsc}^k(\boldsymbol{\alpha}_1, \boldsymbol{\beta}_1, \boldsymbol{\alpha}_2, \boldsymbol{\beta}_2) \geq \underline{\kappa} \|\boldsymbol{\Delta}_k\|_{\bar{\Sigma}}^2 \cdot \underbrace{\frac{1}{nh} \sum_{i=1}^n \chi_{i,k} \cdot \varphi_{h/(4r)}\big(\boldsymbol{\xi}_{\boldsymbol{\Delta}_k}\big) \phi_{h/4}\big(\bar{\boldsymbol{x}}_i{}^{\mathrm{T}} \boldsymbol{\Delta}_k^*\big)}_{=: D_{rsc}^{0,k}(\boldsymbol{\alpha}_1, \boldsymbol{\beta}_1, \boldsymbol{\alpha}_2, \boldsymbol{\beta}_2)} \tag{A.53}$$

We have $3\underline{f}h/4 \leq \mathbb{E}\chi_{i,k} \leq 5\overline{f}h/4$ almost surely, which is similar to the proof of the Proposition 3.3. Using the sub-Gaussianity and the similar analyses following (A.13), we have

$$\begin{aligned} &\mathbb{E}\big\{\chi_{i,k} \varphi_{h/(4r)}\big(\bar{\boldsymbol{x}}_i{}^{\mathrm{T}} \boldsymbol{\Delta}_k / \|\boldsymbol{\Delta}_k\|_{\bar{\Sigma}}\big) \phi_{h/4}\big(\bar{\boldsymbol{x}}_i{}^{\mathrm{T}} \boldsymbol{\Delta}_k^*\big)\big\} \\ &\geq \frac{3}{4} \underline{f} h \mathbb{E}\big\{\varphi_{h/(4r)}\big(\bar{\boldsymbol{x}}_i{}^{\mathrm{T}} \boldsymbol{\Delta}_k / \|\boldsymbol{\Delta}_k\|_{\bar{\Sigma}}\big) \phi_{h/4}\big(\bar{\boldsymbol{x}}_i{}^{\mathrm{T}} \boldsymbol{\Delta}_k^*\big)\big\} \\ &\geq \frac{3}{4} \underline{f} h \big(1 - \mathbb{E}\{\boldsymbol{\xi}_{\boldsymbol{\Delta}_k}^2 I_{|\boldsymbol{\xi}_{\boldsymbol{\Delta}_k}| \geq h/(8r)}\} - \mathbb{E}\{\boldsymbol{\xi}_{\boldsymbol{\Delta}_k}^2 I_{|\boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{\Delta}_k^*| \geq h/8}\}\big) > \frac{9}{16} \underline{f} h \end{aligned} \tag{A.54}$$

when $h/(8r) > 3\nu_0^2$.

Now, we need to bound $| - D_{rsc}^{0,k}(\boldsymbol{\alpha}_1, \boldsymbol{\beta}_1, \boldsymbol{\alpha}_2, \boldsymbol{\beta}_2) + \mathbb{E}\{D_{rsc}^{0,k}(\boldsymbol{\alpha}_1, \boldsymbol{\beta}_1, \boldsymbol{\alpha}_2, \boldsymbol{\beta}_2)\}|$ uniformly over $\Lambda(r, l)$. Let

$$Z_k(r, l) := \sup_{\Lambda(r,l)} | - D_{rsc}^{0,k}(\boldsymbol{\alpha}_1, \boldsymbol{\beta}_1, \boldsymbol{\alpha}_2, \boldsymbol{\beta}_2) + \mathbb{E}\{D_{rsc}^{0,k}(\boldsymbol{\alpha}_1, \boldsymbol{\beta}_1, \boldsymbol{\alpha}_2, \boldsymbol{\beta}_2)\}|.$$

If we denote $D_{rsc}^{0,k}(\boldsymbol{\alpha}_1, \boldsymbol{\beta}_1, \boldsymbol{\alpha}_2, \boldsymbol{\beta}_2) = (1/n) \sum_{i=1}^n w_k(\boldsymbol{x}_i, \varepsilon_i)$, where

$$w_k(\boldsymbol{x}_i, \varepsilon_i) := (\chi_{i,k}/h) \cdot \varphi_{h/(4r)}\big(\bar{\boldsymbol{x}}_i{}^{\mathrm{T}} \boldsymbol{\Delta}_k / \|\boldsymbol{\Delta}_k\|_{\bar{\Sigma}}\big) \phi_{h/4}\big(\bar{\boldsymbol{x}}_i{}^{\mathrm{T}} \boldsymbol{\Delta}_k^*\big),$$

we have

$$0 \leq w_k(\boldsymbol{x}_i, \varepsilon_i) \leq h/(8r)^2, \quad \text{and} \quad \mathbb{E} w_k^2(\boldsymbol{x}_i \varepsilon_i) \leq (8r)^{-2} \cdot 5\overline{f}h/4.$$

Again using Talagrand's inequality as in the proof of Proposition 3.3, for any $t > 0$, we get

$$Z_k(r,l) \le \frac{5}{4}\mathbb{E}Z_k(r,l) + \sqrt{\frac{\overline{f}ht}{16r^2n}} + \frac{ht}{12r^2n} \tag{A.55}$$

with probability at least $1-e^{-t}$. From here, we closely follow the proof of Lemma E.2 of [33] to bound $\mathbb{E}Z_k(r,l)$. With Rademacher symmetrization and using the connection between Gaussian and Rademacher complexities that in Lemma 4.5 of [23], we get

$$\mathbb{E}Z_k(r,l) \le 2\sqrt{\frac{\pi}{2}} \cdot \mathbb{E}\left\{ \sup_{\Lambda(r,l)} \mathbb{G}_k(\boldsymbol{\alpha}_1, \boldsymbol{\beta}_1, \boldsymbol{\alpha}_2, \boldsymbol{\beta}_2) \right\} \tag{A.56}$$

where $\mathbb{G}_k(\boldsymbol{\alpha}_1, \boldsymbol{\beta}_1, \boldsymbol{\alpha}_2, \boldsymbol{\beta}_2) := (nh)^{-1}\sum_{i=1}^n g_{i,k}\chi_{i,k} \cdot \varphi_{h/(4r)}\big(\boldsymbol{\xi}_{\boldsymbol{\Delta}_k}\big)\phi_{h/4}\big(\bar{\boldsymbol{x}}_i{}^{\mathrm{T}}\boldsymbol{\Delta}_k^*\big)$, and $g_{i,k}$ are independent standard normal random variables. Denote $\mathbb{E}^*$ be the conditional expectation given data $\{(y_i, \boldsymbol{x}_i)\}_{i=1}^n$. Then, $\{\mathbb{G}_k(\boldsymbol{\alpha}_1, \boldsymbol{\beta}_1, \boldsymbol{\alpha}_2, \boldsymbol{\beta}_2)\}_{\Lambda(r.l)}$ is a Gaussian process, zero at the true parameter. Now, apply the Gaussian comparison theorem to bound $\mathbb{E}^*\{\sup_{\Lambda(r,l)} \mathbb{G}_k(\boldsymbol{\alpha}_1, \boldsymbol{\beta}_1, \boldsymbol{\alpha}_2, \boldsymbol{\beta}_2)\}$.

Let $\boldsymbol{\gamma}_1^{\mathrm{T}} = (\boldsymbol{\alpha}_1^{\mathrm{T}}, \boldsymbol{\beta}_1^{\mathrm{T}}), \boldsymbol{\gamma}_2^{\mathrm{T}} = (\boldsymbol{\alpha}_2^{\mathrm{T}}, \boldsymbol{\beta}_2^{\mathrm{T}})$, and denote $\boldsymbol{\gamma}_{1,k}^{\mathrm{T}} = (\alpha_{1k}, \boldsymbol{\beta}_1^{\mathrm{T}}), \boldsymbol{\gamma}_{2,k}^{\mathrm{T}} = (\alpha_{2k}, \boldsymbol{\beta}_2^{\mathrm{T}}), \boldsymbol{\gamma}_{1,k}'{}^{\mathrm{T}} = (\alpha_{1k}', \boldsymbol{\beta}_1'{}^{\mathrm{T}}), \boldsymbol{\gamma}_{2,k}'{}^{\mathrm{T}} = (\alpha_{2k}', \boldsymbol{\beta}_2'{}^{\mathrm{T}})$, and abbreviate the notation to $\mathbb{G}_k(\boldsymbol{\gamma}_{1,k}, \boldsymbol{\gamma}_{2,k})$.

Let $\boldsymbol{\Delta}_k'{}^{\mathrm{T}} = (\alpha_{1k}' - \alpha_{2k}', \boldsymbol{\beta}_1'{}^{\mathrm{T}} - \boldsymbol{\beta}_2'{}^{\mathrm{T}}), \boldsymbol{\Delta}_k'{}^{*\mathrm{T}} = (\alpha_{2k}' - \alpha_{h,k}^*, \boldsymbol{\beta}_2'{}^{\mathrm{T}} - \boldsymbol{\beta}^{*\mathrm{T}})$. Then, we have

$$\begin{aligned}
&\mathbb{G}_k(\boldsymbol{\gamma}_{1,k}, \boldsymbol{\gamma}_{2,k}) - \mathbb{G}_k(\boldsymbol{\gamma}_{1,k}', \boldsymbol{\gamma}_{2,k}') \\
&= \mathbb{G}_k(\boldsymbol{\gamma}_{1,k}, \boldsymbol{\gamma}_{2,k}) - \mathbb{G}_k(\boldsymbol{\gamma}_{1,k}' + \boldsymbol{\Delta}_k', \boldsymbol{\gamma}_{2,k}') + \mathbb{G}_k(\boldsymbol{\gamma}_{1,k}' + \boldsymbol{\Delta}_k', \boldsymbol{\gamma}_{2,k}') - \mathbb{G}_k(\boldsymbol{\gamma}_{1,k}', \boldsymbol{\gamma}_{2,k}') \\
&= \frac{1}{nh}\sum_{i=1}^n g_{i,k}\chi_{i,k} \cdot \varphi_{h/(4r)}(\boldsymbol{\xi}_{\boldsymbol{\Delta}_k})\{\phi_{h/4}(\bar{\boldsymbol{x}}_i{}^{\mathrm{T}}\boldsymbol{\Delta}_k^*) - \phi_{h/4}(\bar{\boldsymbol{x}}_i{}^{\mathrm{T}}\boldsymbol{\Delta}_k'{}^*)\} \\
&\quad + \frac{1}{nh}\sum_{i=1}^n g_{i,k}\chi_{i,k} \cdot \phi_{h/4}(\bar{\boldsymbol{x}}_i{}^{\mathrm{T}}\boldsymbol{\Delta}_k'{}^*)\{\varphi_{h/(4r)}(\boldsymbol{\xi}_{\boldsymbol{\Delta}_k}) - \varphi_{h/(4r)}(\boldsymbol{\xi}_{\boldsymbol{\Delta}_k'})\}.
\end{aligned}$$

Now, using the Lipshitz continuity of ϕ_R, φ_R and $\varphi_R \le (R/2)^2$, we get

$$\begin{aligned}
&\mathbb{E}^*\big\{\mathbb{G}_k(\boldsymbol{\gamma}_{1,k}, \boldsymbol{\gamma}_{2,k}) - \mathbb{G}_k(\boldsymbol{\gamma}_{1,k}' + \boldsymbol{\Delta}_k', \boldsymbol{\gamma}_{2,k}')\big\}^2 \\
&\le \frac{1}{n^2}\sum_{i=1}^n \frac{h^2}{(8r)^4}\Big(\frac{8}{h}\Big)^2 \langle \bar{\boldsymbol{x}}_i, \boldsymbol{\gamma}_{2,k} - \boldsymbol{\gamma}_{2,k}'\rangle^2\chi_{i,k} = \Big(\frac{1}{8r^2n}\Big)^2 \sum_{i=1}^n \langle \bar{\boldsymbol{x}}_i, \boldsymbol{\gamma}_{2,k} - \boldsymbol{\gamma}_{2,k}'\rangle^2\chi_{i,k}
\end{aligned} \tag{A.57}$$

and

$$\begin{aligned}
&\mathbb{E}^*\big\{\mathbb{G}_k(\boldsymbol{\gamma}_{1,k}' + \boldsymbol{\Delta}_k', \boldsymbol{\gamma}_{2,k}') - \mathbb{G}_k(\boldsymbol{\gamma}_{1,k}', \boldsymbol{\gamma}_{2,k}')\big\}^2 \\
&\le \frac{1}{(nh)^2}\sum_{i=1}^n \big\{\varphi_{h/(4r)}(\bar{\boldsymbol{x}}_i{}^{\mathrm{T}}\boldsymbol{\Delta}_k/\|\boldsymbol{\Delta}_k\|_{\bar{\Sigma}}) - \varphi_{h/(4r)}(\bar{\boldsymbol{x}}_i{}^{\mathrm{T}}\boldsymbol{\Delta}_k'/\|\boldsymbol{\Delta}_k'\|_{\bar{\Sigma}})\big\}^2\chi_{i,k}
\end{aligned}$$

$$\leq \left(\frac{1}{4rn}\right)^2 \sum_{i=1}^{n} \big(\bar{\boldsymbol{x}}_i{}^{\mathrm{T}} \boldsymbol{\Delta}_k / \|\boldsymbol{\Delta}_k\|_{\bar{\Sigma}} - \bar{\boldsymbol{x}}_i{}^{\mathrm{T}} \boldsymbol{\Delta}'_k \|\boldsymbol{\Delta}'_k\|_{\bar{\Sigma}}\big)\chi_{i,k} \tag{A.58}$$

Now, we have an inequality

$$\begin{aligned}
&\mathbb{E}^*\{\mathbb{G}_k(\boldsymbol{\gamma}_{1,k}, \boldsymbol{\gamma}_{2,k}) - \mathbb{G}_k(\boldsymbol{\gamma}'_{1,k}, \boldsymbol{\gamma}'_{2,k})\}^2 \\
&\leq 2\mathbb{E}^*\big\{\mathbb{G}_k(\boldsymbol{\gamma}_{1,k}, \boldsymbol{\gamma}_{2,k}) - \mathbb{G}_k(\boldsymbol{\gamma}'_{1,k} + \boldsymbol{\Delta}'_k, \boldsymbol{\gamma}'_{2,k})\big\}^2 \\
&\quad + 2\mathbb{E}^*\big\{\mathbb{G}_k(\boldsymbol{\gamma}'_{1,k} + \boldsymbol{\Delta}'_k, \boldsymbol{\gamma}'_{2,k}) - \mathbb{G}_k(\boldsymbol{\gamma}'_{1,k}, \boldsymbol{\gamma}'_{2,k})\big\}^2
\end{aligned}$$

which can be bounded by using (A.57) and (A.58). Define another Gaussian process $\{\mathbb{Z}_k(\boldsymbol{\gamma}_1, \boldsymbol{\gamma}_2)\}_{\Lambda(r,l)}$ as

$$\begin{aligned}
\mathbb{Z}_k(\boldsymbol{\gamma}_1, \boldsymbol{\gamma}_2) &= \frac{\sqrt{2}}{8r^2 n} \sum_{i=1}^{n} g'_{i,k} \langle \bar{\boldsymbol{x}}_i, \boldsymbol{\Delta}^*_k \rangle \chi_{i,k} + \frac{\sqrt{2}}{4rn} \sum_{i=1}^{n} g''_{i,k} \frac{\langle \bar{\boldsymbol{x}}_i, \boldsymbol{\Delta}_k \rangle}{\|\boldsymbol{\Delta}_k\|_{\bar{\Sigma}}} \chi_{i,k} \\
&= \frac{\sqrt{2}}{8r^2 n} \sum_{i=1}^{n} \langle g'_{i,k} \bar{\boldsymbol{x}}_{i,\mathcal{S}}, \boldsymbol{\Delta}^*_{k,\mathcal{S}} \rangle \chi_{i,k} + \frac{\sqrt{2}}{4rn} \sum_{i=1}^{n} g''_{i,k} \frac{\langle \bar{\boldsymbol{x}}_i, \boldsymbol{\Delta}_k \rangle}{\|\boldsymbol{\Delta}_k\|_{\bar{\Sigma}}} \chi_{i,k},
\end{aligned}$$

where $\bar{\boldsymbol{x}}_{i,\mathcal{S}} = (1, \boldsymbol{x}_{i,\mathcal{S}}^{\mathrm{T}})^{\mathrm{T}}$, $\boldsymbol{\Delta}^*_{k,\mathcal{S}} = (\alpha_{2k} - \alpha^*_{h,k}, \boldsymbol{\beta}_{2,\mathcal{S}}^{\mathrm{T}} - \boldsymbol{\beta}^{*\mathrm{T}})^{\mathrm{T}}$, and $g'_{1,k}, \ldots, g''_{n,k}$ are i.i.d. standard normal random variable that are independent of other variables.. We can also abbreviate the notation as $\mathbb{Z}_k(\boldsymbol{\gamma}_{1,k}, \boldsymbol{\gamma}_{2,k})$. Then, we have an inequality $\mathbb{E}^*\{\mathbb{G}_k(\boldsymbol{\gamma}_{1,k}, \boldsymbol{\gamma}_{2,k}) - \mathbb{G}_k(\boldsymbol{\gamma}'_{1,k}, \boldsymbol{\gamma}'_{2,k})\}^2 \leq \mathbb{E}^*\{\mathbb{Z}_k(\boldsymbol{\gamma}_{1,k}, \boldsymbol{\gamma}_{2,k}) - \mathbb{Z}_k(\boldsymbol{\gamma}'_{1,k}, \boldsymbol{\gamma}'_{2,k})\}^2$. Applying Sudakov-Fernique's Gaussian comparison inequality (see, e.g. Theorem 7.2.11 in [36]), we get

$$\mathbb{E}^*\bigg\{\sup_{\Lambda(r,l)} \mathbb{G}_k(\boldsymbol{\gamma}_1, \boldsymbol{\gamma}_2)\bigg\} \leq \mathbb{E}^*\bigg\{\sup_{\Lambda(r,l)} \mathbb{Z}_k(\boldsymbol{\gamma}_1, \boldsymbol{\gamma}_2)\bigg\}. \tag{A.59}$$

The above remains valid if we replace $\mathbb{E}^*$ by $\mathbb{E}$. We use the cone-like constraint $\|\boldsymbol{\Delta}_k\|_1 \leq l\|\boldsymbol{\Delta}_k\|_{\bar{\Sigma}}$, and $\|\boldsymbol{\Delta}^*_k\|_{\bar{\Sigma}} \leq r/2$, which leads to

$$\begin{aligned}
\mathbb{E}\bigg\{\sup_{\Lambda(r,l)} \mathbb{Z}_k(\boldsymbol{\gamma}_1, \boldsymbol{\gamma}_2)\bigg\} &\leq \frac{\sqrt{2}}{16r} \mathbb{E}\bigg\| \frac{1}{n} \sum_{i=1}^{n} g'_{i,k} \chi_{i,k} \bar{\boldsymbol{S}}^{-1/2} \bar{\boldsymbol{x}}_{i,\mathcal{S}} \bigg\|_2 \\
&\quad + \frac{\sqrt{2}l}{4r} \mathbb{E}\bigg\| \frac{1}{n} \sum_{i=1}^{n} g''_{i,k} \chi_{i,k} \bar{\boldsymbol{x}}_i \bigg\|_\infty \\
&\leq \frac{\sqrt{2}}{16r} \sqrt{\frac{5\bar{f}h}{4} \frac{s}{n}} + \frac{\sqrt{2}l}{4r} \mathbb{E}\bigg\| \frac{1}{n} \sum_{i=1}^{n} g''_{i,k} \chi_{i,k} \bar{\boldsymbol{x}}_i \bigg\|_\infty.
\end{aligned} \tag{A.60}$$

where $\bar{\boldsymbol{S}} := \mathbb{E}\bar{\boldsymbol{x}}_{\mathcal{S}} \bar{\boldsymbol{x}}_{\mathcal{S}}^{\mathrm{T}}$. Then, from (A.56), (A.59), and (A.60), we get

$$\mathbb{E}Z_k(r,l) \leq \sqrt{\pi}\bigg\{ \frac{\sqrt{5}}{16} \sqrt{\frac{hs}{r^2 n}} + \frac{l}{2r} \mathbb{E}\bigg\| \frac{1}{n} \sum_{i=1}^{n} g''_{i,k} \chi_{i,k} \bar{\boldsymbol{x}}_i \bigg\|_\infty \bigg\}. \tag{A.61}$$

To bound the second term of the right-hand side of (A.61), define $G_j = \sum_{i=1}^n g_{i,k}\chi_{i,k}\bar{x}_{ij}$ for $j = 1,\ldots,p+1$. Using the sub-Gaussianity (A3), for $k \geq 3$, we have

$$\mathbb{E}|\bar{x}_j|^k \leq 2\nu_0^k \sigma_{jj}^{k/2} k \int_0^\infty t^{k-1} e^{-t^2/2}\mathrm{d}t = 2^{k/2}\nu_0^k\sigma_{jj}^{k/2} k\Gamma(k/2).$$

By using the identity $\Gamma(k)\Gamma(k+1/2) = 2^{1-2k}\sqrt{\pi}\Gamma(2k)$, we get

$$\mathbb{E}|g_{i,k}\bar{x}_{ij}| \leq 2^{k/2}\frac{\Gamma(\frac{k+1}{2})}{\sqrt{\pi}} \cdot 2^{k/2}\nu_0^k\sigma_{jj}^{k/2}k\Gamma(k/2) = 2\nu_0^k\sigma_{jj}^{k/2}k!.$$

Thus, for any $0 \leq \lambda < (2\nu_0\sigma_{jj}^{1/2})^{-1}$,

$$\begin{aligned}
\mathbb{E}e^{\lambda g_{i,k}\chi_{i,k}\bar{x}_{ij}} &\leq 1 + \frac{1}{2}\cdot\frac{5\overline{f}h}{4}\sigma_{jj}\lambda^2 + 2\cdot\frac{5\overline{f}h}{4}\sum_{k=3}^\infty \frac{\lambda^{2k}}{(2k)!}\nu_0^{2k}\sigma_{jj}^{2k}(2k)! \\
&\leq 1 + \frac{1}{2}\cdot\frac{5\overline{f}h}{4}\sigma_{jj}\nu_0^2\sum_{k=2}^\infty \lambda^k(2\nu_0\sigma_{jj}^{1/2})^{k-2} \\
&\leq 1 + \frac{1}{2}\cdot\frac{5\overline{f}h}{4}\cdot\frac{\nu_0^2\sigma_{jj}\lambda^2}{1-2\nu_0\sigma_{jj}^{1/2}\lambda}
\end{aligned}$$

which leads to $\log\mathbb{E}e^{\lambda G_j} \leq \frac{1}{2}\cdot\frac{5\overline{f}h}{4}\cdot\frac{\nu_0^2\sigma_{jj}\lambda^2 n}{1-2\nu_0\sigma_{jj}^{1/2}\lambda}$. We can apply same to $-G_j$ using symmetry. By Corollary 2.6 in [5], we have

$$\mathbb{E}\left\|\frac{1}{n}\sum_{i=1}^n g''_{i,k}\chi_{i,k}\bar{\boldsymbol{x}}_i\right\|_\infty \leq \nu_0\sigma_{\boldsymbol{x}}\left\{\frac{5}{2}\sqrt{\frac{\overline{f}h\log(2p)}{n}} + \frac{2\log(2p)}{n}\right\} \tag{A.62}$$

. Then, taking $r = h/(24\nu_0^2)$ and combining above result, we get $Z_k(r,l) \leq \underline{f}/16$ with probability at least $1 - q/(2p)$, which leads to the conclusion by combining those for all $k = 1,\ldots,q$.

A.12. Proof of Theorem 3.5

To prove Theorem 3.5, we need to verify the event in Theorem 3.4 holds with high probability. We closely follow the proof of Lemma E.3 in [33]. First, we bound $\|\nabla_{\boldsymbol{\beta}}\widehat{Q}_h(\widehat{\boldsymbol{\alpha}}^o,\widehat{\boldsymbol{\beta}}^o)\|_\infty$. Let $\boldsymbol{\omega}_h(\boldsymbol{\alpha},\boldsymbol{\beta}) = \nabla_{\boldsymbol{\beta}}\widehat{Q}_h(\boldsymbol{\alpha},\boldsymbol{\beta}) - \nabla_{\boldsymbol{\beta}}Q_h(\boldsymbol{\alpha},\boldsymbol{\beta})$ We have

$$\|\nabla_{\boldsymbol{\beta}}\widehat{Q}_h(\widehat{\boldsymbol{\alpha}}^o,\widehat{\boldsymbol{\beta}}^o)\|_\infty \leq \|\boldsymbol{\omega}_h(\widehat{\boldsymbol{\alpha}}^o,\widehat{\boldsymbol{\beta}}^o) - \boldsymbol{\omega}_h(\boldsymbol{\alpha}_h^*,\boldsymbol{\beta}^*)\|_\infty + \|\nabla_{\boldsymbol{\beta}}Q_h(\widehat{\boldsymbol{\alpha}}^o,\widehat{\boldsymbol{\beta}}^o)\|_\infty + \|\boldsymbol{\omega}^*\|_\infty.$$

Let $\boldsymbol{\gamma}^{\mathrm{T}} = (\boldsymbol{\alpha}^{\mathrm{T}},\boldsymbol{\beta}^{\mathrm{T}}), \boldsymbol{\gamma}_h^{*\mathrm{T}} = (\boldsymbol{\alpha}_h^*,\boldsymbol{\beta}^{*\mathrm{T}})$. Define the oracle neighborhood $\boldsymbol{\Theta}_{\mathcal{S}}^*(r) = \{\boldsymbol{\gamma} \in \boldsymbol{\gamma}_h^* + \mathbb{B}_\Omega(r), \boldsymbol{\beta}_{\mathcal{S}^c} = \boldsymbol{0}\}$. Conditioned on the event $\{\widehat{\boldsymbol{\gamma}}^o \in \boldsymbol{\gamma}_h^* + \mathbb{B}_\Omega(r)\}$, we have

$$\|\boldsymbol{\omega}_h(\widehat{\boldsymbol{\alpha}}^o,\widehat{\boldsymbol{\beta}}^o) - \boldsymbol{\omega}_h(\boldsymbol{\alpha}_h^*,\boldsymbol{\beta}^*)\|_\infty \leq \sup_{\boldsymbol{\Theta}_{\mathcal{S}}^*(r)} \|\boldsymbol{\omega}_h(\boldsymbol{\alpha},\boldsymbol{\beta}) - \boldsymbol{\omega}_h(\boldsymbol{\alpha}_h^*,\boldsymbol{\beta}^*)\|_\infty \tag{A.63}$$

Then, we can bound the above term by bounding it for each quantile index $k = 1. \dots, q$. Let $\boldsymbol{\Delta}^*_{k,\mathcal{S}} = \begin{pmatrix} \alpha_k - \alpha^*_{h,k} \\ \boldsymbol{\beta}_{\mathcal{S}} - \boldsymbol{\beta}^*_{\mathcal{S}} \end{pmatrix}$. Also, let $W_{kj}(\boldsymbol{\Delta}^*_{k,\mathcal{S}}) = (1/n)\sum_{i=1}^n (w_{ikj} - \mathbb{E}w_{ikj})$ where

$$w_{ikj} := \left\{ \bar{K}\left(\frac{\alpha^*_{h,k} - \varepsilon_i + \langle \bar{\boldsymbol{x}}_{i,\mathcal{S}}, \boldsymbol{\Delta}^*_{k,\mathcal{S}} \rangle}{h} \right) - \bar{K}\left(\frac{\alpha^*_{h,k} - \varepsilon_i}{h} \right) \right\} \frac{x_{ij}}{\sigma_{jj}^{1/2}}. \tag{A.64}$$

Then, we have

$$\sup_{\boldsymbol{\Theta}^*_{\mathcal{S}}(r)} \|\boldsymbol{\omega}_h(\boldsymbol{\alpha}, \boldsymbol{\beta}) - \boldsymbol{\omega}_h(\boldsymbol{\alpha}^*_h, \boldsymbol{\beta}^*)\|_\infty \le \sigma_{\boldsymbol{x}} \frac{1}{q} \sum_{k=1}^q \max_{1 \le j \le p} \sup_{\|\boldsymbol{\Delta}^*_{k,\mathcal{S}}\| \le r} |W_{kj}(\boldsymbol{\Delta}^*_{k,\mathcal{S}})| \tag{A.65}$$

Using (A1') and (A2'), we get $|w_{ikj}| \le h^{-1}|x_{ij}\langle \bar{\boldsymbol{x}}_{i,\mathcal{S}}, \boldsymbol{\Delta}^*_{k,\mathcal{S}} \rangle / \sigma_{jj}^{1/2}|, |\mathbb{E}(w_{ikj})| \le \overline{f}\|\boldsymbol{\Delta}^*_{k,\mathcal{S}}\|_{\bar{\boldsymbol{S}}}$. It can be also shown that

$$\mathbb{E}(w_{ijk}^2 | \boldsymbol{x}_i) \le \overline{f} h^{-1} (x_{ij}^2/\sigma_{jj}) \langle \bar{\boldsymbol{x}}_{i,\mathcal{S}}, \boldsymbol{\Delta}^*_{k,\mathcal{S}} \rangle^2$$

using Minkowski's integral inequality. Above inequalities lead to

$$\mathbb{E}\{(w_{ijk} - \mathbb{E}w_{ijk})^2 | \boldsymbol{x}_i\} \le 2\overline{f}^2 \|\boldsymbol{\Delta}^*_{k,\mathcal{S}}\|^2_{\bar{\boldsymbol{S}}} + 2\overline{f} h^{-1} (x_{ij}^2/\sigma_{jj}) \langle \bar{\boldsymbol{x}}_{i,\mathcal{S}}, \boldsymbol{\Delta}^*_{k,\mathcal{S}} \rangle^2.$$

For $\lambda \in \mathbb{R}$,let $\lambda_* = \lambda / \|\boldsymbol{\Delta}^*_{k,\mathcal{S}}\|_{\bar{\boldsymbol{S}}}$, and let $\boldsymbol{\Delta}^{**}_{k,\mathcal{S}} = \boldsymbol{\Delta}^*_{k,\mathcal{S}} / \|\boldsymbol{\Delta}^*_{k,\mathcal{S}}\|_{\bar{\boldsymbol{S}}}$. Then, using $|e^u - 1 - u| \le (u^2/2)e^{u \vee 0}$ we obtain

$$\begin{aligned}
\mathbb{E}e^{\lambda_* W_{kj}(\boldsymbol{\Delta}^*_{k,\mathcal{S}})} &= \prod_{i=1}^n \mathbb{E}e^{\frac{\lambda_*}{n}(w_{ikj} - \mathbb{E}w_{ikj})} \\
&\le \prod_{i=1}^n \mathbb{E}\left\{ 1 + \frac{\lambda_*^2}{2n^2}(w_{ikj} - \mathbb{E}w_{ikj})^2 e^{\frac{|\lambda_*|}{n}|w_{ikj} - \mathbb{E}w_{ikj}|} \right\} \\
&\le \prod_{i=1}^n \left\{ 1 + \frac{\lambda^2 \overline{f}^2}{n^2} e^{\frac{|\lambda| \overline{f}}{n}} \mathbb{E}e^{\frac{|\lambda|}{nh}|\widehat{x}_{ij}\langle \bar{\boldsymbol{x}}_{i,\mathcal{S}}, \boldsymbol{\Delta}^{**}_{k,\mathcal{S}} \rangle|} \right. \\
&\qquad \left. + \frac{\lambda^2 \overline{f}}{n^2 h} e^{\frac{|\lambda| \overline{f}}{n}} \mathbb{E}\widehat{x}_{ij}^2 \langle \bar{\boldsymbol{x}}_{i,\mathcal{S}}, \boldsymbol{\Delta}^{**}_{k,\mathcal{S}} \rangle^2 e^{\frac{|\lambda|}{nh}|\widehat{x}_{ij}\langle \bar{\boldsymbol{x}}_{i,\mathcal{S}}, \boldsymbol{\Delta}^{**}_{k,\mathcal{S}} \rangle|} \right\}.
\end{aligned}$$

where $\widehat{x}_{ij} = x_{ij}/\sigma_{jj}^{1/2}$. Applying Hölder's inequality, we get, for any $t > 0$,

$$\begin{aligned}
&\mathbb{E}\widehat{x}_{ij}^2 \langle \bar{\boldsymbol{x}}_{i,\mathcal{S}}, \boldsymbol{\Delta}^{**}_{k,\mathcal{S}} \rangle^2 e^{t|\widehat{x}_{ij}\langle \bar{\boldsymbol{x}}_{i,\mathcal{S}}, \boldsymbol{\Delta}^{**}_{k,\mathcal{S}} \rangle|} \\
&\le \left\{ \mathbb{E}\widehat{x}_{ij}^2 e^{t\widehat{x}_{ij}^2} \right\}^{1/2} \cdot \left(\mathbb{E}\langle \bar{\boldsymbol{x}}_{i,\mathcal{S}}, \boldsymbol{\Delta}^{**}_{k,\mathcal{S}} \rangle^4 e^{t\langle \bar{\boldsymbol{x}}_{i,\mathcal{S}}, \boldsymbol{\Delta}^{**}_{k,\mathcal{S}} \rangle^2} \right)^{1/2}
\end{aligned}$$

and

$$\mathbb{E}e^{t|\widehat{x}_{ij}\langle \bar{\boldsymbol{x}}_{i,\mathcal{S}}, \boldsymbol{\Delta}^{**}_{k,\mathcal{S}} \rangle|} \le \left(\mathbb{E}e^{t\widehat{x}_{ij}^2} \right)^{1/2} \cdot \left(\mathbb{E}e^{t\langle \bar{\boldsymbol{x}}_{i,\mathcal{S}}, \boldsymbol{\Delta}^{**}_{k,\mathcal{S}} \rangle^2} \right)^{1/2}.$$

For a unit vector $\boldsymbol{u} \in \mathbb{S}^p$,let $Z_{\boldsymbol{u}} = (\boldsymbol{z}^{\mathrm{T}}\boldsymbol{u})^2/(4\nu_0^2)$, where $\boldsymbol{z} = \bar{\Sigma}^{-1/2}\bar{\boldsymbol{x}}$.

Then, using sub-Gaussianity, we can show that

$$\mathbb{E}e^{Z_{\boldsymbol{u}}} = 1 + \int_0^\infty e^u \mathbb{P}(Z_{\boldsymbol{u}} \geq u)\mathrm{d}u \leq 3,$$
$$\mathbb{E}Z_{\boldsymbol{u}}^2 e^{Z_{\boldsymbol{u}}} = \int_0^\infty (u^2+2u)e^u \mathbb{P}(Z_{\boldsymbol{u}} \geq u)\mathrm{d}u \leq 8.$$

Then, we obtain

$$\mathbb{E}e^{\lambda_* W_{kj}(\boldsymbol{\Delta}^*_{k,\mathcal{S}})} \leq \prod_{i=1}^n \{1 + C\nu_0^4 \overline{f}/(n^2h)\} \leq e^{C\nu_0^4\overline{f}/(nh)},$$

for $|\lambda| \leq \min\{nh/(4\nu_0^2), n/\overline{f}\}$, where $C > 0$ is an absolute constant.

Similarly, for each pair $(\boldsymbol{\Delta}^*_{k,\mathcal{S}}, \boldsymbol{\Delta}^{*'}_{k,\mathcal{S}})$, we have a bound

$$\mathbb{E}e^{\lambda\{W_{kj}(\boldsymbol{\Delta}^*_{k,\mathcal{S}}) - W_{kj}(\boldsymbol{\Delta}^{*'}_{k,\mathcal{S}})\}/\|\boldsymbol{\Delta}^*_{k,\mathcal{S}} - \boldsymbol{\Delta}^{*'}_{k,\mathcal{S}}\|_{\bar{\Sigma}}} \leq e^{C\nu_0^4\overline{f}/(nh)}.$$

Then, we can use Corollary 2.2 in [31] since above result satisfies condition $(\mathcal{E}d)$ of [31]. Thus, with probability at least $1 - e^{-u}$,

$$\sup_{\|\boldsymbol{\Delta}^*_{k,\mathcal{S}}\| \leq r} |W_{kj}(\boldsymbol{\Delta}^*_{k,\mathcal{S}})| \lesssim \nu_0^2 \overline{f}^{1/2} \sigma_{\boldsymbol{x}} r \sqrt{\frac{s+u}{nh}}$$

provided $nh \gtrsim (s+u)^{1/2}$ Taking $u = \log(2p)$ and combining bounds for $k = 1, \ldots, q$, we get

$$\sup_{\boldsymbol{\Theta}^*_{\mathcal{S}}(r)} \|\boldsymbol{\omega}_h(\boldsymbol{\alpha}, \boldsymbol{\beta}) - \boldsymbol{\omega}_h(\boldsymbol{\alpha}^*_h, \boldsymbol{\beta}^*)\|_\infty \lesssim \sigma_{\boldsymbol{x}} r \sqrt{\frac{s + \log p}{nh}} \tag{A.66}$$

with probability at least $1 - q/(2p)$ as long as $nh \gtrsim (s + \log p)^{1/2}$.

Now, we need to bound $\|\nabla_{\boldsymbol{\beta}} Q_h(\widehat{\boldsymbol{\alpha}}^o, \widehat{\boldsymbol{\beta}}^o)\|_\infty$. Since $\nabla_{\boldsymbol{\beta}} Q_h(\widehat{\boldsymbol{\alpha}}^o, \widehat{\boldsymbol{\beta}}^o)_{\mathcal{S}} = \mathbf{0}$, only need to bound $\mathcal{S}^c$ part. For $\boldsymbol{\gamma} \in \boldsymbol{\Theta}^*_{\mathcal{S}}(r)$, we have

$$\nabla_{\boldsymbol{\beta}} Q_h(\boldsymbol{\alpha}, \boldsymbol{\beta})_{\mathcal{S}^c} - \nabla_{\boldsymbol{\beta}} Q_h(\boldsymbol{\alpha}^*_h, \boldsymbol{\beta}^*)_{\mathcal{S}^c}$$
$$= \frac{1}{q}\sum_{k=1}^q \mathbb{E}\int_{-\infty}^\infty K(u)\big\{F_\varepsilon\big(\bar{\boldsymbol{x}}_{\mathcal{S}}^{\mathrm{T}}\boldsymbol{\Delta}^*_{k,\mathcal{S}} + F_\varepsilon^{-1}(\tau_k) - hu\big) - F_\varepsilon\big(F_\varepsilon^{-1}(\tau_k) - hu\big)\big\}\boldsymbol{x}_{\mathcal{S}^c}\mathrm{d}u$$

Using Taylor expansion, we get

$$F_\varepsilon\big(\bar{\boldsymbol{x}}_{\mathcal{S}}^{\mathrm{T}}\boldsymbol{\Delta}^*_{k,\mathcal{S}} + F_\varepsilon^{-1}(\tau_k) - hu\big) - F_\varepsilon\big(F_\varepsilon^{-1}(\tau_k) - hu\big)$$
$$= f_\varepsilon\big(F_\varepsilon^{-1}(\tau_k)\big) \cdot \bar{\boldsymbol{x}}_{\mathcal{S}}^{\mathrm{T}}\boldsymbol{\Delta}^*_{k,\mathcal{S}} + \int_0^{\bar{\boldsymbol{x}}_{\mathcal{S}}^{\mathrm{T}}\boldsymbol{\Delta}^*_{k,\mathcal{S}}} \{f_\varepsilon(t - hu + F_\varepsilon^{-1}(\tau_k)) - f_\varepsilon(F_\varepsilon^{-1}(\tau_k))\}\mathrm{d}t.$$

Let $\boldsymbol{J}_{\mathcal{S}^c\mathcal{S}} = q^{-1}\sum_{k=1}^q f_\varepsilon(F^{-1}(\tau_k))\mathbb{E}(\boldsymbol{x}_{\mathcal{S}^c}\boldsymbol{x}_{\mathcal{S}}^{\mathrm{T}})$, note that $\mathbb{E}\boldsymbol{x}_{\mathcal{S}^c} = 0$, then above displays imply

$$\|\nabla_{\boldsymbol{\beta}} Q_h(\boldsymbol{\alpha}, \boldsymbol{\beta})_{\mathcal{S}^c} - \nabla_{\boldsymbol{\beta}} Q_h(\boldsymbol{\alpha}^*_h, \boldsymbol{\beta}^*)_{\mathcal{S}^c} - \boldsymbol{J}_{\mathcal{S}^c\mathcal{S}}(\boldsymbol{\beta} - \boldsymbol{\beta}^*)\|_\infty$$

$$\begin{aligned}&\leq 0.5 l_0 \max_{j \in \mathcal{S}^c} \mathbb{E}\left\{\frac{1}{q}\sum_{k=1}^{q}(\bar{\boldsymbol{x}}_{\mathcal{S}}^{\mathrm{T}}\boldsymbol{\Delta}_{k,\mathcal{S}^*}^2)^2|x_j|\right\} + \max_{j \in \mathcal{S}^c}\left(\frac{1}{q}\sum_{k=1}^{q}\kappa_1 h \mathbb{E}|x_j \bar{\boldsymbol{x}}_{\mathcal{S}}^{\mathrm{T}}\boldsymbol{\Delta}_{k,\mathcal{S}}^*|\right)\\ &\qquad \leq 0.5 l_0 \sigma_{\boldsymbol{x}} \mu_4^{1/2}\left\|\begin{matrix}(\boldsymbol{\alpha}-\boldsymbol{\alpha}_h^*)/\sqrt{q}\\ \boldsymbol{\beta}_{\mathcal{S}}-\boldsymbol{\beta}_{\mathcal{S}}^*\end{matrix}\right\|_{\Omega_{\mathcal{S}}}^2 + l_0\kappa_1 h \sigma_{\boldsymbol{x}} \cdot \frac{1}{q}\sum_{k=1}^{q}\|\boldsymbol{\Delta}_{k,\mathcal{S}}^*\|_{\bar{\boldsymbol{S}}}.\end{aligned}$$

Thus, it gives

$$\|\nabla_{\boldsymbol{\beta}} Q_h(\widehat{\boldsymbol{\alpha}}^o, \widehat{\boldsymbol{\beta}}^o)_{\mathcal{S}^c} - \boldsymbol{J}_{\mathcal{S}^c\mathcal{S}}(\widehat{\boldsymbol{\beta}}^o - \boldsymbol{\beta}^*)_{\mathcal{S}}\|_\infty \leq 0.5 l_0 \sigma_{\boldsymbol{x}} \mu_4^{1/2} r^2 + l_0 \kappa_1 h \sigma_{\boldsymbol{x}} r. \tag{A.67}$$

Now. it remains to bound. $\|\boldsymbol{J}_{\mathcal{S}^c\mathcal{S}}(\widehat{\boldsymbol{\beta}}^o - \boldsymbol{\beta}^*)_{\mathcal{S}}\|_\infty$. Using the condition given in the statement of the theorem, we obtain

$$\begin{aligned}&\|\boldsymbol{J}_{\mathcal{S}^c\mathcal{S}}(\widehat{\boldsymbol{\beta}}^o - \boldsymbol{\beta}^*)_{\mathcal{S}}\|_\infty = \|\boldsymbol{J}_{\mathcal{S}^c\mathcal{S}}(\boldsymbol{J}_{\mathcal{S}\mathcal{S}})^{-1}\boldsymbol{J}_{\mathcal{S}\mathcal{S}}(\widehat{\boldsymbol{\beta}}^o - \boldsymbol{\beta}^*)_{\mathcal{S}}\|_\infty\\ &\leq \max_{j \in \mathcal{S}^c}\|\boldsymbol{J}_{j\mathcal{S}}(\boldsymbol{J}_{\mathcal{S}^c\mathcal{S}})^{-1}\|_1 \cdot \|\boldsymbol{J}_{\mathcal{S}\mathcal{S}}(\widehat{\boldsymbol{\beta}}^o - \boldsymbol{\beta}^*)_{\mathcal{S}}\|_\infty \leq A_0 \cdot \|\boldsymbol{J}_{\mathcal{S}\mathcal{S}}(\widehat{\boldsymbol{\beta}}^o - \boldsymbol{\beta}^*)_{\mathcal{S}}\|_\infty.\end{aligned} \tag{A.68}$$

Instead of using the trivial ℓ_2 bound for ℓ_∞-norm, we have Proposition 3.4, which gives a Bahadur representation of the oracle estimator

$$\begin{aligned}&\left\|\boldsymbol{D}(\widehat{\boldsymbol{\beta}}^o - \boldsymbol{\beta}^*)_{\mathcal{S}} + \frac{1}{nq}\sum_{i=1}^{n}\sum_{k=1}^{q}\{\bar{K}((\alpha_k^* - \varepsilon_i)/h) - \tau_k\}\boldsymbol{x}_{i,\mathcal{S}}\right\|_2\\ &\qquad \lesssim \frac{(s+t)}{h^{1/2}n} + h^{3/2}\sqrt{\frac{q(s+t)}{n}} + h^4\end{aligned} \tag{A.69}$$

with probability at least $1 - 3qe^{-t}$, where $\boldsymbol{D} = \boldsymbol{J}_{\mathcal{S}\mathcal{S}}$. This gives

$$\begin{aligned}&\|\boldsymbol{D}(\widehat{\boldsymbol{\beta}}^o - \boldsymbol{\beta}^*)_{\mathcal{S}}\|_\infty\\ &\leq \|\boldsymbol{D}(\widehat{\boldsymbol{\beta}}^o - \boldsymbol{\beta}^*)_{\mathcal{S}} + \nabla_{\boldsymbol{\beta}}\widehat{Q}_h(\boldsymbol{\alpha}_h^*, \boldsymbol{\beta}^*)_{\mathcal{S}}\|_\infty + \|\nabla_{\boldsymbol{\beta}}\widehat{Q}_h(\boldsymbol{\alpha}_h^*, \boldsymbol{\beta}^*)_{\mathcal{S}}\|_\infty\\ &\lesssim \frac{(s+t)}{h^{1/2}n} + h^{3/2}\sqrt{\frac{q(s+t)}{n}} + h^4 + \sqrt{\frac{\log(s)+t}{n}},\end{aligned} \tag{A.70}$$

where

$$\|\nabla_{\boldsymbol{\beta}}\widehat{Q}_h(\boldsymbol{\alpha}_h^*, \boldsymbol{\beta}^*)_{\mathcal{S}}\|_\infty \lesssim \sqrt{\frac{\log(s)+t}{n}}$$

with probability at least $1 - e^{-t}$ by using the proof of Proposition 3.2. Then, combining all results above, we get

$$\begin{aligned}\|\nabla_{\boldsymbol{\beta}}\widehat{Q}_h(\widehat{\boldsymbol{\alpha}}^o, \widehat{\boldsymbol{\beta}}^o)\|_\infty &\lesssim \sqrt{\frac{\log(2p)}{n}} + \sqrt{\frac{s+\log p}{nh}}\sqrt{\frac{q(s+t)}{n}}\\ &\quad + A_0\left\{\sqrt{\frac{\log(s)+t}{n}} + \frac{(s+t)}{h^{1/2}n} + h^{3/2}\sqrt{\frac{q(s+t)}{n}} + h^4\right\}\end{aligned} \tag{A.71}$$

with probability at least $1-q/p-(5q+1)e^{-t}$, provided that $\sqrt{(s \vee \log p + t)/n} \lesssim h \lesssim 1$.

Now, take $t = \log(n)$. Using the conditions from Theorem 3.4 and Proposition 3.5, set $r = h/(24\nu_0^2)$, $l = \sqrt{2}\{2 + 2/P'(a_0)\} \cdot \big[\max\{q, (1+b^2)s/\gamma_p\}\big]^{1/2}$, $c = 0.5\underline{\kappa f}$, and choose the bandwidth parameter $h \asymp \{\log(p)/n\}^{1/4}$, we get, with probability at least $1 - 2q/p - (5q+1)/n$,

$$\|\nabla_{\boldsymbol{\beta}} \widehat{Q}_h(\widehat{\boldsymbol{\alpha}}^o, \widehat{\boldsymbol{\beta}}^o)\|_\infty \lesssim \sqrt{\frac{\log(p)}{n}}, \|\boldsymbol{\theta}^o\|_\Omega \lesssim \sqrt{\frac{s+\log(n)}{n}}, \|\widehat{\boldsymbol{\beta}}^o - \boldsymbol{\beta}^*\|_\infty \lesssim \sqrt{\frac{\log(p)}{n}}$$

provided that $n \gtrsim \max\{s^{8/3}/(\log p)^{5/3}, \log(p)\}$. Then, as in (3.27) required in Theorem 3.4, choosing $\lambda = C\sqrt{\log(p)/n}$ with sufficiently large $C > 0$, we have (3.27) with probability at least $1 - 2q/p - (5q+1)/n$, provided that $n \gtrsim \max\{s^{8/3}/(\log p)^{5/3}, s^{4/3}\log(p)\}$, thus provind the strong oracle property.

A.13. Lemma F.2 of [14]

Lemma F.2 of [14] has been used multiple times throughout the technical proofs. Here, we include the statement of the Lemma.

Lemma F.2. Let $D_{\mathcal{L}}(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2) = \mathcal{L}(\boldsymbol{\beta}_1) - \mathcal{L}(\boldsymbol{\beta}_2) - \langle \mathcal{L}(\boldsymbol{\beta}_2), \boldsymbol{\beta}_1 - \boldsymbol{\beta}_2\rangle$ and $D_c L^s(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2) = D_{\mathcal{L}}(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2) + D_{\mathcal{L}}(\boldsymbol{\beta}_2, \boldsymbol{\beta}_1)$. For $\boldsymbol{\beta}(t) = \boldsymbol{\beta}^* + t(\boldsymbol{\beta} - \boldsymbol{\beta}^*)$ with $t \in (0, 1]$, we have that

$$D^s_{\mathcal{L}}(\boldsymbol{\beta}(t), \boldsymbol{\beta}^*) \le t D^s_{\mathcal{L}}(\boldsymbol{\beta}, \boldsymbol{\beta}^*).$$

The above lemma applies to general differentiable convex loss functions, so we can apply in our smoothed loss function as well.

References

[1] Avella-Medina, M. and Ronchetti, E. (2018). Robust and consistent variable selection in high-dimensional generalized linear models. *Biometrika* **105** 31–44. MR3768863

[2] Bach, F., Jenatton, R., Mairal, J. and Obozinski, G. (2012). Optimization with sparsity-inducing penalties. *Foundations and Trends in Machine Learning* **4** 1–106.

[3] Belloni, A. and Chernozhukov, V. (2011). ℓ_1-penalized quantile regression in high-dimensional sparse models. *Ann. Statist.* **39** 82–130. MR2797841

[4] Bickel, P. J., Ritov, Y. and Tsybakov, A. (2009). Simultaneous analysis of Lasso and Dantzig selector. *Ann. Statist.* **37** 1705–1732. MR2533469

[5] Boucheron, S., Lugosi, G. and Massart, P. (2013). *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford University Press, Oxford. MR3185193

[6] Bousquet, O. (2003). Concentration inequalities for sub-additive functions using the entropy method. In *Stochastic Inequalities and Applications* **56** 213–247. Birkhäuser, Basel. MR2073435

[7] Bradic, J., Fan, J. and Wang, W. (2011). Penalized composite quasi-likelihood for ultrahigh dimensional variable selection. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **73** 325–349. MR2815779

[8] Bühlmann, P. and van de Geer, S. (2011). *Statistics for High-Dimensional data: Methods, Theory and Applications.* Springer, Heidelberg. MR2807761

[9] Chen, S. S., Donoho, D. L. and Saunder, M. A. (1999). Atomic decomposition by basis pursuit. *SIAM Journal on Scientific Computing* **20** 33–61. MR1639094

[10] Clémençon, S., Bellet, A. and Colin, I. (2016). Scaling-up empirical risk minimization: Optimization of incomplete U-statistics. *J. Mach. Learn. Res.* **17**(76): 1–36. MR3517099

[11] 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. MR3597972

[12] Fan, J. and Li, R. (2001). Variable selection via nonconcave regularized likelihood and its oracle properties. *J. Amer. Statist. Assoc.* **96** 1348–1360. MR1946581

[13] Fan, J., Li, R., Zhang, C.-H. and Zou, H. (2020). *Statistical Foundations of Data Science.* Chapman and Hall/CRC, New York.

[14] 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. MR3782385

[15] Fernandes, M., Guerre, E. and Horta, E. (2021). Smoothing quantile regressions. *J. Bus. Econ. Statist.* **39** 338–357. MR4187194

[16] Gu, Y., Fan, J., Kong, L., Ma, S. and Zou, H. (2018). ADMM for high-dimensional sparse regularized quantile regression. *Technometrics* **60** 319–331. MR3847169

[17] Gu, Y. and Zou, H. (2020). Sparse composite quantile regression in ultrahigh dimensions with tuning parameter calibration. *IEEE Transactions on Information Theory* **66** 7132–7154. MR4173632

[18] Hastie, T., Tibshirani, R. and Wainwright, M. (2015). *Statistical Learning with Sparsity: The Lasso and Generalizations.* CRC Press, Boca Raton. MR3616141

[19] He, X., Pan, X., Tan, K. M. and Zhou, W.-X. (2022). Smoothed quantile regression with large-scale inference. *Journal of Econometrics*, to appear. MR4539491

[20] Huangfu, Q. and Hall, J. A. J. (2018). Parallelizing the dual revised simplex method. *Mathematical Programming Computation* **10** 119–142. MR3773090

[21] Kai, B., Li, R. and Zou, H. (2010). Local composite quantile regression smoothing: an efficient and safe alternative to local polynomial regression. *J. R. Statist. Soc.* B **72** 49–69. MR2751243

[22] Koenker, R. and Bassett, G (1978). Regression quantiles. *Econometrica* **46** 33-50. MR0474644

[23] Ledoux, M. and Talagrand, M. (1991). *Probability in Banach Spaces: Isoperimetry and Processes.* Springer-Verlag, Berlin. MR1102015

[24] Lee, E. R., Noh, H. and Park, B. U. (2014). Model selection via Bayesian information criterion for quantile regression models. *J. Amer. Statist. Assoc.* **109** 216–229. MR3180558

[25] Li, Y. and Zhu, J. (2008). ℓ_1-norm quantile regression. *J. Comp. Graph. Statist.* **17** 163–185. MR2424800

[26] Loh, P.-L. (2017). Statistical consistency and asymptotic normality for high-dimensional robust M-estimators. *Ann. Statist.* **45** 866–896. MR3650403

[27] Lozano, A. C., Meinshausen, N. and Yang, E. (2016). Minimum distance lasso for robust high-dimensional regression. *Electron. J. Statist.* **10** 1296–1340. MR3504182

[28] Rudelson, M. and Zhou, S. (2013). Reconstruction from anisotropic random measurements. *IEEE Transactions on Information Theory* **59** 3434–3447. MR3061256

[29] She, Y., Wang, Z. and Shen, J. (2021). Gaining outlier resistence with progressive quantile: fast algorithms and theoretical studies. *J. Amer. Statist. Assoc.*, in press. MR4480712

[30] Silverman, B. W. (1986). *Density Estimation for Statistics and Data Analysis.* Boca Raton, FL: CRC/Chapman and Hall. MR0848134

[31] Spokoiny, V. (2012). Parametric estimation. Finite sample theory *Ann. Statist.* **40** 2877–2909 MR3097963

[32] Sun, Q., Zhou, W.-X. and Fan, J. (2020). Adaptive Huber regression. *J. Amer. Statist. Assoc.* **115** 254–265. MR4078461

[33] Tan, K. M., Wang, L. and Zhou, W.-X. (2022). High-dimensional quantile regression: convolution smoothing and concave regularization. *J. R. Statist. Soc.* B **84** 205–233. MR4400395

[34] Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *J. R. Statist. Soc.* B **58** 267–288. MR1379242

[35] van der Vaart, A. W. and Wellner, J. A. (1996). *Weak Convergence and Empirical Processes: With Applications to Statistics.* Springer, New York. MR1385671

[36] Vershynin, R. (12018). *High-Dimensional Probability.* Cambridge University Press, Cambridge. MR3837109

[37] Wainwright, M. (2019). *High-Dimensional Statistics: A Non-asymptotic Viewpoint.* Cambridge University Press, Cambridge. MR3967104

[38] Wang, H., Li, G. and Jiang, G. (2007). Robust regression shrinkage and consistent variable selection through the LAD-Lasso. *J. Bus. Econ. Statist.* **25** 347–355. MR2380753

[39] Wang, L., Peng, B., Bradic, J., Li, R. and Wu, Y. (2020). A tuning-free robust and efficient approach to high-dimensional regression. *J. Amer. Statist. Assoc.* **115** 1700–1714. MR4189748

[40] Wang, X., Jiang, Y., Huang, M. and Zhang, H. (2013). Robust vari-

able selection with exponential squared loss. *J. Amer. Statist. Assoc.* **108** 632–643. MR3174647

[41] Yan, Y., Wang, X. and Zhang, R. (2023). Composite smoothed quantile regression. *Stat,* **12(1)** e542. MR4567856

[42] Yu, L., Lin, N. and Wang, L. (2017). A parallel algorithm for large-scale nonconvex penalized quantile regression. *J. Comp. Graph. Statist.* **26** 935–939. MR3765357

[43] Zhang, C.-H. (2010). Nearly unbiased variable selection under minimax concave penalty. *Ann. Statist.* **38** 894–942. MR2604701

[44] Zhang, C.-H. and Zhang, T. (2012). A general theory of concave regularization for high-dimensional sparse estimation problems. *Statist. Sci.* **27** 576–593. MR3025135

[45] Zou, H. and Li, R. (2008). One-step sparse estimates in nonconcave penalized likelihood models. *Ann. Statist.* **36** 1509–1533. MR2435443

[46] Zou, H. and Yuan, M. (2008). Composite quantile regression and the oracle model selection theory. *Ann. Statist.* **36** 1108–1126. MR2418651