Received: 2 January 2020 | Accepted: 23 October 2021

DOI: 10.1111/rssb.12485

ORIGINAL ARTICLE

High-dimensional quantile regression: Convolution smoothing and concave regularization

Kean Ming Tan[1] | **Lan Wang**[2] | **Wen-Xin Zhou**[3]

[1]Department of Statistics, University of Michigan, Ann Arbor, Michigan, USA

[2]Miami Herbert Business School, University of Miami, Coral Gables, Florida, USA

[3]Department of Mathematics, University of California, San Diego, La Jolla, California, USA

Correspondence
Wen-Xin Zhou, Department of Mathematics, University of California, San Diego, La Jolla, CA 92093, USA.
Email: wez243@ucsd.edu

Funding information
NSF, Grant/Award Numbers: DMS-1949730, DMS-2113346, FRGMS-1952373, OAC-2023755, DMS-2113409; NIH, Grant/Award Number: RF1-MH122833

Abstract

ℓ_1-penalized quantile regression (QR) is widely used for analysing high-dimensional data with heterogeneity. It is now recognized that the ℓ_1-penalty introduces non-negligible estimation bias, while a proper use of concave regularization may lead to estimators with refined convergence rates and oracle properties as the signal strengthens. Although folded concave penalized M-estimation with strongly convex loss functions have been well studied, the extant literature on QR is relatively silent. The main difficulty is that the quantile loss is piecewise linear: it is non-smooth and has curvature concentrated at a single point. To overcome the lack of smoothness and strong convexity, we propose and study a convolution-type smoothed QR with iteratively reweighted ℓ_1-regularization. The resulting smoothed empirical loss is twice continuously differentiable and (provably) locally strongly convex with high probability. We show that the iteratively reweighted ℓ_1-penalized smoothed QR estimator, after a few iterations, achieves the optimal rate of convergence, and moreover, the oracle rate and the strong oracle property under an almost necessary and sufficient minimum signal strength condition. Extensive numerical studies corroborate our theoretical results.

KEYWORDS

concave regularization, convolution, minimum signal strength, oracle property, quantile regression

1 | INTRODUCTION

Massive complex datasets bring challenges to data analysis due to the presence of outliers and heterogeneity. Consider regression of a scalar response y on a p-dimensional predictor $\boldsymbol{x} \in \mathbb{R}^p$. The least squares method focuses on the conditional mean of the outcome given the predictor. Despite its popularity in the statistical and econometric literature, it is sensitive to outliers and fails to capture heterogeneity in the set of important features. Moreover, in many applications, the scientific question of interest may not be fully addressed by inferring the conditional mean. Since the seminal work of Koenker and Bassett (1978), quantile regression (QR) has gained increasing attention by offering a set of complementary methods designed to explore data features invisible to the inveiglements of least squares methods. Quantile regression is robust to data heterogeneity and outliers, and also offers unique insights into the entire conditional distribution of the outcome given the predictor. We refer to Koenker (2005) and Koenker et al. (2017) for an overview of QR theory, methods and applications.

In the high-dimensional setting in which the number of features, p, exceeds the number of observations, n, it is often the case that only a small subset of a large pool of features influences the conditional distribution of the outcome. To perform estimation and variable selection simultaneously, the standard approach is to minimize the empirical loss plus a penalty on the model complexity. The ℓ_1-penalty is arguably the most commonly used penalty function that induces sparsity (Tibshirani, 1996). Least squares methods with ℓ_1-regularization have been extensively studied in the past two decades. Because of the extremely long list of relevant literature, we refer the reader to the monographs Bühlmann and van de Geer (2011), Hastie et al. (2015), Wainwright (2019), Fan et al. (2020), and the references therein. In the context of QR, Belloni and Chernozhukov (2011) provided a comprehensive analysis of the ℓ_1-penalized QR as well as post-penalized QR estimator. Since then, the literature on high-dimensional QR has grown rapidly, and we refer to Chapter 15 of Koenker et al. (2017) for an overview.

It is now a consensus that the ℓ_1-penalty induces non-negligible bias (Fan & Li, 2001; Zhang & Zhang, 2012; Zou, 2006), due to which the selected model tends to include spurious variables unless stringent conditions are imposed on the design matrix, such as the strong irrepresentable condition (Meinshausen & Bühlmann, 2006; Zhao & Yu, 2006). To reduce the bias induced by the ℓ_1-penalty when the signal is sufficiently strong, various concave penalty functions have been designed (Fan & Li, 2001; Zhang, 2010a,b). For concave penalized M-estimation with convex and locally strongly convex losses, a large body of literature has shown that there exists a local solution that possesses the oracle property, that is, a solution that is as efficient as the oracle estimator obtained by assuming the true active set is known a priori, under certain minimum signal strength condition, also known as the beta-min condition. We refer the reader to Fan and Li (2001), Zou and Li (2008), Kim et al. (2008), Zhang (2010b), Fan and Lv (2011), Zhang and Zhang (2012), Kim and Kwon (2012), Loh and Wainwright (2015), and Loh (2017) for more details.

Comparably, QR with concave regularization is much less understood theoretically primarily due to the challenges in analysing the piecewise linear quantile loss and the concave penalty simultaneously. Let $\boldsymbol{\beta}^* \in \mathbb{R}^p$ be the s-sparse underlying parameter vector with support $\mathcal{S} = \{1 \le j \le p : \beta_j^* \neq 0\}$, and define the minimum signal strength $\|\boldsymbol{\beta}^*_{\mathcal{S}}\|_{\min} = \min_{j \in \mathcal{S}} |\beta_j^*|$. Under a beta-min condition $\|\boldsymbol{\beta}^*_{\mathcal{S}}\|_{\min} \gg n^{-1/2} \max\{s, \sqrt{\log(p)}\}$, Wang et al. (2012) showed that the oracle QR estimator belongs to the set of local minima of the non-convex penalized quantile objective function with probability approaching one. From a different angle, Fan et al. (2014) proved that the oracle QR estimator can be obtained via the one-step local linear approximation (LLA) algorithm (Zou & Li, 2008) under a beta-min condition $\|\boldsymbol{\beta}^*_S\|_{\min} \gtrsim \sqrt{s\log(p)/n}$, that is, the minimal non-zero coefficient is of order $\sqrt{s\log(p)/n}$ in magnitude. We refer to Chapter 16 of Koenker et al. (2017) for an overview of the existing results on non-convex regularized QR. Existing work on folded concave penalized QR either impose stringent signal strength assumptions or only establish theoretical guarantees for some local optimum which, due to non-convexity, is not necessarily the solution obtained by any practical algorithm. In other words, there is no guarantee that the solution obtained from a given algorithm will satisfy the desired statistical properties, leaving a gap between theory and practice.

A natural way to resolve the non-differentiability issue is to smooth the piecewise linear quantile loss using a kernel. The idea of kernel smoothing was first considered by Horowitz (1998) in the context of bootstrap inference for median regression. Horowitz (1998) showed that the estimator obtained from the smoothed quantile loss is asymptotically equivalent to that of the standard QR estimator. This motivates a series of work on smoothed QR when the number of features is fixed (Galvao & Kato, 2016; Whang, 2006; Wu et al., 2015). However, smoothing the piecewise linear loss directly yields a non-convex function for which global minimum is not guaranteed. This poses even more challenges in the high-dimensional setting.

In this paper, we propose and study a new method for QR in high-dimensional sparse models, which is based on convolution smoothing and iteratively reweighted ℓ_1-penalization. To deal with non-smoothness, we smooth the piecewise linear quantile loss via convolution. The idea is to smooth the subgradient of the quantile loss, and then integrate it to obtain a smoothed loss function that is also convex. See Figure 1 for a visualization of Horowitz's and convolution smoothing methods. Fernandes et al. (2021) developed the traditional asymptotic theory for convolution smoothing in the context of linear QR when the sample size n tends to infinity while p is kept fixed. For high-dimensional sparse models, we extend the one-step LLA algorithm proposed by Zou and Li (2008), and propose a multi-step, iterative procedure which solves a weighted ℓ_1-penalized smoothed quantile objective function at each iteration. This multi-step procedure consists of a sequence of convex programs, which is similar to the multi-stage convex relaxation method for sparse regularization (Fan et al., 2018; Zhang, 2010b). Computationally, for different smoothing kernels, typified by the uniform and Gaussian kernels, we propose efficient algorithms to minimize the weighted ℓ_1-penalized smoothed quantile objective function at each stage. Comparing with existing methods for fitting high-dimensional QR, the proposed gradient-based algorithms are more scalable to large-scale problems with either large sample size or high dimensionality.

Since the proposed multi-step procedure delivers a sequence of solutions iteratively, to understand how these estimators evolve statistically, we provide a delicate analysis of the estimator at each stage whose overall estimation error consists of three components: shrinkage bias, oracle rate and smoothing bias. The theoretical analysis in Zhang (2010b) and Fan et al. (2018) is primarily suited for the quadratic case, although the method applies to more general loss functions. In this work, we aim at establishing theoretical underpinnings of why and how convolution smoothing and iteratively reweighted ℓ_1-penalization help with achieving oracle properties for QR.

In particular, we show that the solution for the first iteration, that is, the ℓ_1-penalized smoothed QR, is near minimax optimal, and coincide with those of existing results for ℓ_1-penalized QR estimator. Moreover, our analysis reveals that the multi-step, iterative algorithm refines the statistical rate in a sequential manner: every relaxation step shrinks the estimation error from the previous step by a δ-fraction for some predetermined $\delta \in (0, 1)$. All the results are non-asymptotic with explicit errors depending on (s, p, n), including the deterministic smoothing bias and stochastic statistical errors. With a minimal requirement on the

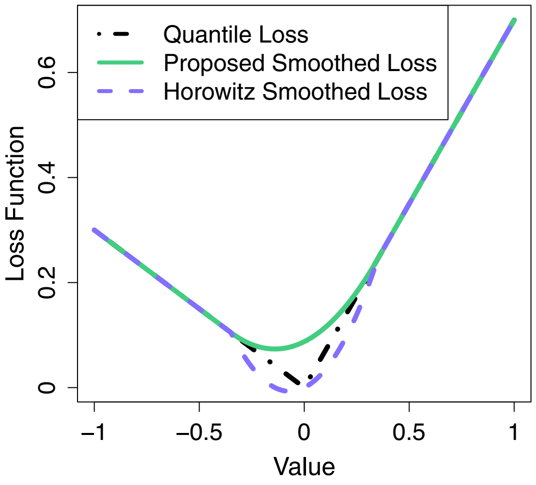

FIGURE 1 Plots of a standard quantile loss, Horowitz's smoothed quantile loss (Horowitz, 1998), and a convolution-type smoothed quantile loss [Colour figure can be viewed at wileyonlinelibrary.com]

signal strength—$\|\boldsymbol{\beta}^*_{\mathcal{S}}\|_{\min} \gtrsim \sqrt{\log(p)/n}$, we show that after as many as $\ell \gtrsim \lceil \log(\max\{\log(p), s\})\rceil$ iterations, the multi-step algorithm will deliver an estimator that achieves the oracle rate of convergence as well as the strong oracle property. The latter implies variable selection consistency as a byproduct. To our knowledge, these are the first statistical characterizations of computationally feasible concave regularized QR estimators.

The rest of the paper is organized as follows. In Section 2, we describe the convolution-type smoothing approach for QR, followed by an iteratively reweighted ℓ_1-penalized procedure for fitting high-dimensional sparse models. At each stage, the problem boils down to minimizing a weighted ℓ_1-penalized smoothed quantile objective function, for which we propose efficient and scalable algorithms in Section 3 with a particular focus on uniform and Gaussian kernels. In Section 4, we provide theoretical guarantees for the sequence of estimators obtained by the multi-step method, including estimation error bounds (in high probability) and strong oracle property. A numerical demonstration of the proposed method on simulated data and a real data application are provided in Sections 5 and 6, respectively. The proofs of all theoretical results are given in the online supplementary material. The Python code that implements the proposed iteratively reweighted regularized QR procedure is available at https://github.com/WenxinZhou/conquer.

Notation: For every integer $k \ge 1$, we use $\mathbb{R}^k$ to denote the k-dimensional Euclidean space, and write $[k] = \{1, \dots, k\}$. The inner product of any two vectors $\boldsymbol{u} = (u_1, \dots, u_k)^{\mathrm{T}}$, $\boldsymbol{v} = (v_1, \dots, v_k)^{\mathrm{T}} \in \mathbb{R}^k$ is defined by $\boldsymbol{u}^{\mathrm{T}}\boldsymbol{v} = \langle \boldsymbol{u}, \boldsymbol{v} \rangle = \sum_{i=1}^k u_i v_i$. Moreover, let $\boldsymbol{u} \circ \boldsymbol{v} = (u_1 v_1, \dots, u_k v_k)^{\mathrm{T}}$ denote the Hadamard product of $\boldsymbol{u}$ and $\boldsymbol{v}$. For a subset $\mathcal{S} \subseteq [k]$ with cardinality $|\mathcal{S}|$, we write $\boldsymbol{u}_{\mathcal{S}} \in \mathbb{R}^{|\mathcal{S}|}$ as the subvector of $\boldsymbol{u}$ that consists of the entries of $\boldsymbol{u}$ indexed by $\mathcal{S}$. We use $\|\cdot\|_p$ $(1 \le q \le \infty)$ to denote the ℓ_q-norm in $\mathbb{R}^k$: $\|\boldsymbol{u}\|_q = (\sum_{i=1}^k |u_i|^q)^{1/q}$ and $\|\boldsymbol{u}\|_\infty = \max_{1 \le i \le k} |u_i|$. For $k \ge 2$, $\mathbb{S}^{k-1} = \{\boldsymbol{u} \in \mathbb{R}^k : \|\boldsymbol{u}\|_2 = 1\}$ denotes the unit sphere in $\mathbb{R}^k$. For any function $f : \mathbb{R} \mapsto \mathbb{R}$ and vector $\boldsymbol{u} = (u_1, \dots, u_k)^{\mathrm{T}} \in \mathbb{R}^k$, we write $f(\boldsymbol{u}) = (f(u_1), \dots, f(u_k))^{\mathrm{T}} \in \mathbb{R}^k$.

Throughout this paper, we use bold uppercase letters to represent matrices. For $k \ge 2$, $\mathbf{I}_k$ represents an $k \times k$ identity matrix. For any $k \times k$ symmetric, positive semidefinite matrix $\mathbf{A} \in \mathbb{R}^{k \times k}$, we use $\gamma(\mathbf{A}) \in \mathbb{R}^k$ to denote its vector of eigenvalues, ordered as $\gamma_1(\mathbf{A}) \ge \cdots \ge \gamma_p(\mathbf{A}) \ge 0$, and let

$\|\mathbf{A}\|_2 = \gamma_1(\mathbf{A})$ be the operator norm of $\mathbf{A}$. Moreover, let $\|\cdot\|_{\mathbf{A}}$ denote the vector norm induced by $\mathbf{A} : \|\boldsymbol{u}\|_{\mathbf{A}} = \|\mathbf{A}^{1/2}\boldsymbol{u}\|_2$ for $\boldsymbol{u} \in \mathbb{R}^k$. For any two real numbers u and v, we write $u \vee v = \max(u,v)$ and $u \wedge v = \min(u, v)$. For two sequences of non-negative numbers $\{a_n\}_{n\geq 1}$ and $\{b_n\}_{n\geq 1}$, $a_n \lesssim b_n$ indicates that there exists a constant $C > 0$ independent of n such that $a_n \geq Cb_n$; $a_n \gtrsim b_n$ is equivalent to $b_n \lesssim a_n$; $a_n \asymp b_n$ is equivalent to $a_n \lesssim b_n$ and $b_n \lesssim a_n$. For two numbers C_1 and C_2, we write $C_2 = C_2(C_1)$ if C_2 depends only on C_1.

2 | SPARSE QUANTILE REGRESSION: CONVOLUTION SMOOTHING AND ITERATIVE REGULARIZATION

2.1 | Penalized quantile regression

We consider a scalar response variable $y \in \mathbb{R}$ and a p-dimensional feature vector $\boldsymbol{x} = (x_1, \dots, x_p)^{\mathrm{T}} \in \mathbb{R}^p$ such that the τth conditional quantile of y given $\boldsymbol{x}$ is modelled as $F^{-1}_{y|\boldsymbol{x}}(\tau|\boldsymbol{x}) = \boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta}^*$ for some $0 < \tau < 1$, where $\boldsymbol{\beta}^* = (\beta_1^*, \dots, \beta_p^*)^{\mathrm{T}} \in \mathbb{R}^p$. Let $\{(y_i, \boldsymbol{x}_i)\}_{i=1}^n$ be a random sample from $(y, \boldsymbol{x})$. The preceding model assumption is equivalent to

$$y_i = \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}^* + \varepsilon_i \ \text{ and } \ \mathbb{P}(\varepsilon_i \leq 0|\boldsymbol{x}_i) = \tau. \tag{1}$$

Throughout the paper, we set $x_1 \equiv 1$ so that β_1^* denotes the intercept. To avoid notational clutter, the dependence of $\boldsymbol{\beta}^*$ and ε_i on τ will be assumed without displaying.

Given a random sample $\{(y_i, \boldsymbol{x}_i)\}_{i=1}^n$, a penalized QR estimator is generally defined as either the global optimum or one of the local optima to the optimization problem

$$\underset{\boldsymbol{\beta}=(\beta_1,\dots,\beta_p)^{\mathrm{T}}\in\mathbb{R}^p}{\text{minimize}} \left\{ \underbrace{\frac{1}{n}\sum_{i=1}^n \rho_\tau(y_i - \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta})}_{=:\hat{Q}(\boldsymbol{\beta})} + \sum_{j=1}^p q_\lambda(|\beta_j|) \right\}, \tag{2}$$

where $\rho_\tau(u) = u\{\tau - \mathbb{1}(u < 0)\}$ is the τ-quantile function, also referred to as the check function, and $q_\lambda(\cdot) : [0, \infty) \to [0, \infty)$ is a sparsity-inducing penalty function parametrized by $\lambda > 0$.

Due to convexity, the ℓ_1-penalized method for which $q_\lambda(t) = \lambda t\,(t \geq 0)$ has dominated the literature on high-dimensional statistics. Work in the context of QR include that of Wang et al. (2007), Belloni and Chernozhukov (2011), Bradic et al. (2011), Wang (2013), and Zheng et al. (2015), Sivakumar and Banerjee (2017), among others. Various algorithms can be employed to solve the resulting ℓ_1-penalized problem (Bach et al., 2012; Boyd et al., 2010; Gu et al., 2018; Koenker et al., 2017). To alleviate the non-negligible bias induced by the ℓ_1 penalty, folded concave penalties have been used in, for example, Wang et al. (2012) and Fan et al. (2014), leading to non-convex optimization problems. Together, the non-differentiable quantile loss and the non-convex penalty bring fundamental statistical and computational challenges.

Statistical theory of non-convex regularized QR is relatively underdeveloped. Most of the existing results are developed either under stringent minimum signal strength conditions, or for the hypothetical global optimum (or one of the local optima). Motivated from the algorithmic approaches developed by Zou and Li (2008) and Fan et al. (2018), we consider a multi-step iterative method that solves a sequence of convex problems, which bypasses the computational issues from solving the non-convex problem (2) directly. Theoretically, a major difficulty is that the quantile loss is piecewise linear, so that its 'curvature energy' is concentrated in a single point. This is in contrast to many popular loss functions considered in the statistical literature, such as the squared, logistic or Huber loss, which are at least locally strongly convex. Therefore, a proper smoothing scheme that creates smoothness and local strong convexity is the key to the success of the proposed framework.

2.2 | Convolution-type smoothing approach

Let $F_{\varepsilon|\boldsymbol{x}}(\cdot)$ be the conditional distribution of ε given $\boldsymbol{x}$. The population quantile loss can then be written as

$$Q(\boldsymbol{\beta}) = \mathbb{E}_{\boldsymbol{x}}\left\{\int_{-\infty}^{\infty} \rho_\tau(u - \langle \boldsymbol{x}, \boldsymbol{\beta} - \boldsymbol{\beta}^* \rangle)\, \mathrm{d}F_{\varepsilon|\boldsymbol{x}}(u)\right\},$$

where $\mathbb{E}_{\boldsymbol{x}}(\cdot)$ is the expectation taken with respect to $\boldsymbol{x}$. Provided that the conditional distribution $F_{\varepsilon|\boldsymbol{x}}(\cdot)$ is sufficiently smooth, $Q(\boldsymbol{\beta})$ is twice differentiable and strongly convex in a neighbourhood of $\boldsymbol{\beta}^*$. For every $\boldsymbol{\beta} \in \mathbb{R}^p$, let $\hat{F}(\cdot;\boldsymbol{\beta})$ be the empirical cumulative distribution function (ECDF) of the residuals $\{r_i(\boldsymbol{\beta}) := y_i - \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}\}_{i=1}^n$, that is, $\hat{F}(u;\boldsymbol{\beta}) = (1/n)\sum_{i=1}^n \mathbb{1}\{r_i(\boldsymbol{\beta}) \le u\}$ for any $u \in \mathbb{R}$. Then, the empirical quantile loss $\hat{Q}(\cdot)$ in Equation (2) can be expressed as

$$\hat{Q}(\boldsymbol{\beta}) = \int_{-\infty}^{\infty} \rho_\tau(u)\, \mathrm{d}\hat{F}(u;\boldsymbol{\beta}). \tag{3}$$

Since the ECDF $\hat{F}(\cdot;\boldsymbol{\beta})$ is discontinuous, the standard empirical quantile loss $\hat{Q}(\cdot)$ has the same degree of smoothness as $\rho_\tau(\cdot)$. This motivates Fernandes et al. (2021) to use a kernel CDF estimator. Given the residuals $r_i(\boldsymbol{\beta}) = y_i - \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}$ and a smoothing parameter/bandwidth $h = h_n > 0$, let $\hat{F}_h(\cdot;\boldsymbol{\beta})$ be the distribution function of the classical Rosenblatt–Parzen kernel density estimator:

$$\hat{F}_h(u;\boldsymbol{\beta}) = \int_{-\infty}^{u} \hat{f}_h(t;\boldsymbol{\beta})\, \mathrm{d}t \text{ with } \hat{f}_h(t;\boldsymbol{\beta}) = \frac{1}{n}\sum_{i=1}^n K_h(t - r_i(\boldsymbol{\beta})),$$

where $K : \mathbb{R} \to [0,\infty)$ is a symmetric, non-negative kernel that integrates to one, and $K_h(u) := (1/h)K(u/h)$ for $u \in \mathbb{R}$. Replacing $\hat{F}(u;\boldsymbol{\beta})$ in Equation (3) with its kernel-smoothed counterpart $\hat{F}_h(u;\boldsymbol{\beta})$ yields the following smoothed empirical quantile loss

$$\hat{Q}_h(\boldsymbol{\beta}) := \int_{-\infty}^{\infty} \rho_\tau(u)\, \mathrm{d}\hat{F}_h(u;\boldsymbol{\beta}) = \frac{1}{nh}\sum_{i=1}^n \int_{-\infty}^{\infty} \rho_\tau(u) K\left(\frac{u + \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta} - y_i}{h}\right) \mathrm{d}u. \tag{4}$$

Define the integrated kernel function $\overline{K} : \mathbb{R} \to [0,1]$ as $\overline{K}(u) = \int_{-\infty}^{u} K(t)\,\mathrm{d}t$. As will be shown in Section 4.1, the smoothed empirical quantile objective function $\hat{Q}_h(\boldsymbol{\beta})$ is twice continuously differentiable with gradient $\nabla\hat{Q}_h(\boldsymbol{\beta}) = (1/n)\sum_{i=1}^n \{\overline{K}(-r_i(\boldsymbol{\beta})/h) - \tau\}\boldsymbol{x}_i$ and Hessian matrix $\nabla^2\hat{Q}_h(\boldsymbol{\beta}) = (1/n)\sum_{i=1}^n K_h(-r_i(\boldsymbol{\beta}))\boldsymbol{x}_i\boldsymbol{x}_i^{\mathrm{T}}$. Moreover, we will show that the smoothed objective function $\hat{Q}_h(\cdot)$ is strongly convex in a cone local neighbourhood of $\boldsymbol{\beta}^*$ with high probability; see Proposition 2.

Remark 1 For a given kernel function $K(\cdot)$ and bandwidth $h > 0$, the smoothed quantile loss $\hat{Q}_h(\cdot)$ defined in Equation (4) can be equivalently written as $\hat{Q}_h(\boldsymbol{\beta}) = (1/n)\sum_{i=1}^n \ell_h(y_i - \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta})$, where

$$\ell_h(u) = (\rho_\tau * K_h)(u) = \int_{-\infty}^{\infty} \rho_\tau(v) K_h(v-u)\,\mathrm{d}v, \quad u \in \mathbb{R}. \tag{5}$$

Here * denotes the convolution operator. To better understand this smoothing mechanism, we compute the smoothed loss $\ell_h = \rho_\tau * K_h$ explicitly for several widely used kernel functions. Recall that $\rho_\tau(u) = |u|/2 + (\tau - 1/2)u$.

1. (Uniform kernel) For the uniform kernel $K(u) = (1/2)\mathbb{1}(|u| \le 1)$, which is the density function of the uniform distribution on $[-1, 1]$, the resulting smoothed loss takes the form $\ell_h(u) = (h/2)U(u/h) + (\tau - 1/2)u$, where $U(u) = (u^2/2 + 1/2)\mathbb{1}(|u| \le 1) + |u|\mathbb{1}(|u| > 1)$ is a Huber-type loss. Convolution plays a role of random smoothing in the sense that $\ell_h(u) = (1/2)\mathbb{E}(|Z_u|) + (\tau - 1/2)u$, where for every $u \in \mathbb{R}$, Z_u denotes a random variable uniformly distributed between $u - h$ and $u + h$.
2. (Gaussian kernel) For the Gaussian kernel $K(u) = \phi(u)$, the density function of a standard normal distribution, the resulting smoothed loss is $\ell_h(u) = (1/2)\mathbb{E}(|G_u|) + (\tau - 1/2)u$, where $G_u \sim N(u, h^2)$. Note that $|G_u|$ follows a folded normal distribution (Leone et al., 1961) with mean $\mathbb{E}|G_u| = (2/\pi)^{1/2} h e^{-u^2/(2h^2)} + u\{1 - 2\Phi(-u/h)\}$. Hence, the smoothed loss can be written as $\ell_h(u) = (h/2)G(u/h) + (\tau - 1/2)u$, where $G(u) = (2/\pi)^{1/2} e^{-u^2/2} + u\{1 - 2\Phi(-u)\}$.
3. (Laplacian kernel) In the case of the Laplacian kernel $K(u) = e^{-|u|}/2$, we have $\ell_h(u) = \rho_\tau(u) + h e^{-|u|/h}/2$.
4. (Logistic kernel) In the case of the logistic kernel $K(u) = e^{-u}/(1 + e^{-u})^2$, the resulting smoothed loss is $\ell_h(u) = \tau u + h \log(1 + e^{-u/h})$.
5. (Epanechnikov kernel) For the Epanechnikov kernel $K(u) = (3/4)(1 - u^2)\mathbb{1}(|u| \le 1)$, the resulting smoothed loss is $\ell_h(u) = (h/2)E(u/h) + (\tau - 1/2)u$, where $E(u) = (3u^2/4 - u^4/8 + 3/8)\mathbb{1}(|u| \le 1) + |u|\mathbb{1}(|u| > 1)$.

2.3 | Iteratively reweighted ℓ_1-penalized method

Let $\{(y_i, \boldsymbol{x}_i)\}_{i=1}^n$ be independent data vectors from the conditional quantile model (1) with a sparse target parameter $\boldsymbol{\beta}^* \in \mathbb{R}^p$. Extending the one-step LLA algorithm proposed by Zou and Li (2008), we consider a multi-step, iteratively regularized method as follows. Let $q_\lambda(\cdot)$ be a prespecified penalty function that is differentiable almost everywhere. Starting at iteration 0 with an initial estimator $\hat{\boldsymbol{\beta}}^{(0)}$, for $\ell = 1,2,\ldots$, we iteratively update the previous estimator $\hat{\boldsymbol{\beta}}^{(\ell-1)}$ by solving

$$\hat{\boldsymbol{\beta}}^{(\ell)} = (\hat{\beta}_1^{(\ell)}, \ldots, \hat{\beta}_p^{(\ell)})^{\mathrm{T}} \in \underset{\beta=(\beta_1,\ldots,\beta_p)^{\mathrm{T}}}{\operatorname{argmin}} \left\{ \hat{Q}_h(\boldsymbol{\beta}) + \sum_{j=1}^p q'_\lambda(|\hat{\beta}_j^{(\ell-1)}|)|\beta_j| \right\}, \tag{6}$$

where $q'_\lambda(\cdot)$ is the first-order derivative of $q_\lambda(\cdot)$, and $\hat{Q}_h(\cdot)$ is the convolution smoothed quantile objective function defined in Equation (4). To avoid notational clutter, we suppress the dependence of $\{\hat{\boldsymbol{\beta}}^{(\ell)} = \hat{\boldsymbol{\beta}}_h^{(\ell)}(\tau, \lambda)\}_{\ell \ge 0}$ on the quantile index τ, bandwidth h, and penalty level λ.

The penalty function $q_\lambda(\cdot)$, or its derivative to be exact, plays the role of producing sparse solutions. We consider a class of penalty functions that satisfies the following conditions.

(A1) The penalty function q_λ is of the form $q_\lambda(t) = \lambda^2 q(t/\lambda)$ for $t \geq 0$, where $q{:}[0,\infty) \mapsto [0,\infty)$ satisfies: (i) q is non-decreasing on $[0, \infty)$ with $q(0) = 0$; (ii) $q(\cdot)$ is differentiable almost everywhere on $(0, \infty)$, $0 \leq q'(t) \leq 1$ and $\lim_{t\downarrow 0} q'(t) = 1$; (iii) $q'(t_1) \leq q'(t_2)$ for all $t_1 \geq t_2 \geq 0$.

Examples of penalties that satisfy Condition (A1) include:

1. ℓ_1-penalty: $q(t) = |t|$. In this case, $q'(t) = 1$ for all $t > 0$. Therefore, $\hat{\boldsymbol{\beta}}^{(1)}$ defined in Equation (6) with $\ell = 1$ is the ℓ_1-penalized SQR estimator, and the procedure stops after the first step.
2. Smoothly clipped absolute deviation (SCAD) penalty (Fan & Li, 2001): The function $q(\cdot)$ is defined through its derivative $q'(t) = \mathbb{1}(t \leq 1) + \frac{(a-t)_+}{a-1}\mathbb{1}(t > 1)$ for $t \geq 0$ and some $a > 2$, and $q(0) = 0$. Fan and Li (2001) suggested $a = 3.7$ by a Bayesian argument.
3. Minimax concave penalty (MCP) (Zhang, 2010a): The function $q(\cdot)$ is defined through its derivative $q'(t) = (1 - t/a)_+$ for $t \geq 0$ and some $a \geq 1$, and $q(0) = 0$.
4. Capped-ℓ_1 penalty (Zhang, 2010b): $q(t) = \min(a/2, t)$ and $q'(t) = \mathbb{1}(t \leq a/2)$ for $t \geq 0$ and some $a \geq 1$.

If we start the multi-step procedure using any penalty q_λ that satisfies Condition (A1) and a trivial initialization $\hat{\boldsymbol{\beta}}^{(0)} = \mathbf{0}$, then $q'_\lambda(|\hat{\beta}_j^{(0)}|) = q'_\lambda(0) = \lambda$ for $j = 1, \ldots, p$, and hence the first step is essentially computing an ℓ_1-penalized smoothed QR estimator. At each subsequent iteration, the subproblem (6) can be expressed as a weighted ℓ_1-penalized smoothed quantile loss minimization:

$$\underset{\boldsymbol{\beta}\in\mathbb{R}^p}{\text{minimize}}\{\hat{Q}_h(\boldsymbol{\beta}) + \|\lambda\circ\boldsymbol{\beta}\|_1\}, \tag{7}$$

where $\lambda = (\lambda_1, \ldots, \lambda_p)^{\mathrm{T}}$ is a p-vector of regularization parameters with $\lambda_j \geq 0$, and $\circ$ denotes the Hadamard product. We summarize this iteratively reweighted ℓ_1-penalized method in Algorithm 1.

Algorithm 1. Iteratively Reweighted ℓ-Penalized Smoothed QR

```
Input: Data vectors $\{(y_i, \boldsymbol{x}_i)\}_{i=1}^n$, quantile index $\tau \in (0, 1)$, bandwidth $h > 0$, and an initial estimator $\widehat{\beta}^{(0)} \in \mathbb{R}^p$.
For $\ell = 1, 2, \ldots$, repeat
    1. Set $\lambda_j^{(\ell-1)} = q'_\lambda(|\widehat{\beta}_j^{(\ell-1)}|)$ for $j = 1, \ldots, p$;
    2. Compute
        $\widehat{\beta}^{(\ell)} \in \operatorname*{argmin}_{\beta\in\mathbb{R}^p} \{\widehat{Q}_h(\beta) + \|\boldsymbol{\lambda}^{(\ell-1)} \circ \beta\|_1\};$    (8)
until convergence.
```

In Section 4, we will establish non-asymptotic statistical theory for the sequence of estimators $\{\hat{\boldsymbol{\beta}}^{(\ell)}\}_{\ell\geq 0}$ initialized with $\hat{\boldsymbol{\beta}}^{(0)} = \mathbf{0}$ when the penalty $q_\lambda(t) = \lambda^2 q(t/\lambda)$ obeys Condition (A1). In order to reduce the (regularization) bias when the signal is sufficiently strong, we are particularly interested in the concave penalty $q(\cdot)$, which not only satisfies Condition (A1) but also has a redescending derivative, that is, $q'(t) = 0$ for all sufficiently large t.

Another widely applicable idea for bias reduction is adaptive Lasso (Zou, 2006), which is a one-step procedure that solves, in the context of QR,

$$\tilde{\boldsymbol{\beta}} \in \underset{\boldsymbol{\beta}\in\mathbb{R}^p}{\operatorname{argmin}} \left\{ \hat{Q}(\boldsymbol{\beta}) + \lambda \sum_{j=1}^{p} w(|\tilde{\beta}_j^{(0)}|)|\beta_j| \right\}, \tag{9}$$

where $\tilde{\boldsymbol{\beta}}^{(0)} = (\tilde{\beta}_1^{(0)}, \dots, \tilde{\beta}_p^{(0)})^{\mathrm{T}}$ is an initial estimator of $\boldsymbol{\beta}^*$, say the ℓ_1-QR (or QR-Lasso) estimator (Belloni & Chernozhukov, 2011), and $w(t) := t^{-\gamma}$ for $t > 0$ and some $\gamma > 0$. Note that the weight function $\lambda w(\cdot)$ for adaptive Lasso is quite different from $q'_\lambda(\cdot) = \lambda q'(\cdot/\lambda)$ in Equation (6). As discussed in Fan and Lv (2008), an advantage of the concave penalty, such as SCAD and MCP, is that zero is not an absorbing state: once a coefficient is shrunk to zero, it will remain zero throughout the remaining iterations. As a result, any true positive that is left out by the initial Lasso estimator will be missed in the second stage as well. The aforementioned is an important phenomenon which was empirically verified by Fan et al. (2018).

Remark 2 In practice, it is common to leave a subset of parameters, such as the intercept and coefficients which correspond to features that are already viewed relevant, unpenalized throughout the multi-step procedure (6). Given a predetermined index set $\mathcal{R} \subseteq [p]$, we can modify Algorithm 1 by taking $\boldsymbol{\lambda}^{(\ell)} = (\lambda_1^{(\ell)}, \dots, \lambda_p^{(\ell)})^{\mathrm{T}}$ ($\ell \geq 0$) to be $\lambda_j^{(\ell)} = 0$ for $j \in \mathcal{R}$ and $\lambda_j^{(\ell)} = q'_\lambda(|\hat{\beta}_j^{(\ell)}|)$ for $j \notin \mathcal{R}$. Theoretically, we will study the sequence of estimates $\{\hat{\boldsymbol{\beta}}^{(\ell)}\}_{\ell\geq 1}$ obtained from Algorithm 1 because a special treatment of leaving parameters indexed by $\mathcal{R}$ unpenalized only makes things more convoluted and does not bring new insights from a theoretical viewpoint.

3 | ALGORITHM

As discussed in Section 2.3, the multi-step convex relaxation method leads to a sequence of iteratively reweighted ℓ_1-penalized problems. Computationally, it suffices to develop efficient algorithms for solving the convex problem (8). For several commonly used kernels, explicit forms of the smoothed check loss functions are given in Remark 1. In the following sections, we present specialized algorithms for two representative kernel functions: the uniform kernel and the Gaussian kernel.

3.1 | A coordinate descent algorithm for uniform kernel

First we describe a coordinate descent algorithm for solving Equation (8) with the uniform kernel, that is, $K(u) = 1/2$ for $|u| \leq 1$. The coordinate descent algorithm is an iterative method that minimizes the objective function with respect to one variable at a time while fixing the other variables. To implement the algorithm, we calculate the partial derivative of the loss function in Equation (8) with respect to each variable, and derive the corresponding update for each variable while keeping the others fixed.

The gradient of the loss function in Equation (8) involves $\overline{K}(\cdot)$. For the uniform kernel, we have

$$\overline{K}\left(\frac{\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}-y_i}{h}\right)=\begin{cases}1 & \text{if } \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}-y_i\geq h,\\ \frac{1}{2}\left(\frac{\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}-y_i}{h}+1\right) & \text{if } |\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}-y_i|\leq h,\\ 0 & \text{if } \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}-y_i\leq -h.\end{cases}$$

Let $C_1=\{i:\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}-y_i\leq -h\}$, $C_2=\{i:|\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}-y_i|\leq h\}$, and $C_3=\{i:\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}-y_i\geq h\}$. Then, the first-order optimality condition of minimizing $\beta_j\rightarrow \hat{Q}_h(\boldsymbol{\beta})+\|\lambda^{(\ell-1)}\circ\boldsymbol{\beta}_-\|_1$ can be written as

$$-\tau\sum_{i=1}^{n}x_{ij}+\frac{1}{2}\sum_{i\in C_2}x_{ij}+\sum_{i\in C_3}x_{ij}+\frac{1}{2h}\sum_{i\in C_2}(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}-y_i)x_{ij}+n\lambda_j^{(\ell-1)}\hat{z}_j=0,$$

where $\hat{z}_j\in\partial|\hat{\beta}_j|$ is the subgradient. This leads to the following closed-form solution for $\hat{\beta}_j$:

$$\hat{\beta}_j=S\left\{\frac{2h\tau\sum_{i=1}^{n}x_{ij}-2h\sum_{i\in C_3}^{n}x_{ij}-h\sum_{i\in C_2}^{n}x_{ij}+\sum_{i\in C_2}x_{ij}(y_i-\langle\boldsymbol{x}_{i,-j},\boldsymbol{\beta}_{-j}\rangle)}{\sum_{i\in C_2}x_{ij}^2},\frac{2nh\lambda_j^{(\ell-1)}}{\sum_{i\in C_2}x_{ij}^2}\right\},$$

where $S(a,b)=\mathrm{sign}(a)\max(|a|-b{,}0)$ denotes the soft-thresholding operator. Therefore, a solution of Equation (8) can be obtained by iteratively updating each $\hat{\beta}_j$ until convergence. The details are summarized in Algorithm 2.

Algorithm 2. Coordinate Descent Algorithm for Solving (2.8) with Uniform Kernel.

Input quantile level τ, smoothing parameter h, regularization parameter $\boldsymbol{\lambda}^{(\ell-1)}$, and convergence criterion ϵ.
Initialization $\widehat{\boldsymbol{\beta}}^{(0)}=\mathbf{0}$.
Iterate the following until the stopping criterion $\|\widehat{\boldsymbol{\beta}}^{(t)}-\widehat{\boldsymbol{\beta}}^{(t-1)}\|_2\leq\epsilon$ is met, where $\widehat{\boldsymbol{\beta}}^{(t)}$ is the value of $\boldsymbol{\beta}$ obtained at the tth iteration. That is, for each $j=1,\ldots,p$:

1. Set $C_1=\{i:\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}-y_i\geq h\}$, $C_2=\{i:|\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}-y_i|\leq h\}$, and $C_3=\{i:\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}-y_i\leq -h\}$, where we use $\boldsymbol{\beta}$ to denote the updated solution at the current iteration.

2. Set

$$\widehat{\beta}_j^{(t)}=S\left\{\frac{2h\tau\sum_{i=1}^{n}x_{ij}-2h\sum_{i\in C_3}^{n}x_{ij}-h\sum_{i\in C_2}^{n}x_{ij}+\sum_{i\in C_2}x_{ij}(y_i-\langle\boldsymbol{x}_{i,-j},\boldsymbol{\beta}_{-j}\rangle)}{\sum_{i\in C_2}x_{ij}^2},\frac{2nh\lambda_j^{(\ell-1)}}{\sum_{i\in C_2}x_{ij}^2}\right\},$$

   where $S(a,b)=\mathrm{sign}(a)\max(|a|-b,0)$ is the soft-thresholding operator.

Output the estimated parameter $\widehat{\boldsymbol{\beta}}^{(t)}$.

Compared to the existing algorithms for solving ℓ_1-regularized QR, Algorithm 2 is computationally efficient especially for large-scale problems. The computational complexity is similar to that of the coordinate descent algorithm for Lasso.

3.2 | An alternating direction method of multiplier algorithm for Gaussian kernel

Next we consider the case of smoothing via the Gaussian kernel function. In this case, we have

$$\overline{K}\left(\frac{\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}-y_i}{h}\right)=\Phi\left(\frac{\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}-y_i}{h}\right),$$

where $\Phi(\cdot)$ is the cumulative distribution function of the standard normal distribution. The coordinate descent approach in the previous section can no longer be employed, at least trivially, to solve Equation (8) since there is no closed-form solution of minimizing $\beta_j \rightarrow \hat{Q}_h(\boldsymbol{\beta})+\|\lambda^{(\ell-1)}\circ\boldsymbol{\beta}_-\|_1$ with the Gaussian kernel. To address this issue, we introduce an alternating direction method of multiplier (ADMM) algorithm to solve Equation (8) by decoupling terms that are difficult to optimize jointly. A similar approach has been considered in Gu et al. (2018) for solving standard QR with ℓ_1-regularization. Let $\boldsymbol{r}=(r_1,\ \dots\ ,r_n)^{\mathrm{T}}$ with $r_i=y_i-\langle\boldsymbol{x}_i,\boldsymbol{\beta}\rangle$. Optimization problem (8) can then be rewritten as

$$\begin{aligned}&\underset{\boldsymbol{\beta}\in\mathbb{R}^p,\boldsymbol{r}\in\mathbb{R}^n}{\text{minimize}}\left\{\hat{Q}_h(\boldsymbol{r})+\|\lambda^{(\ell-1)}\circ\boldsymbol{\beta}_-\|_1\right\},\\&\text{subject to } \boldsymbol{r}=\mathbf{y}-\mathbf{X}\boldsymbol{\beta}.\end{aligned}\tag{10}$$

The augmented Lagrangian for Equation (10) is

$$\mathcal{L}_\rho(\boldsymbol{\beta},\boldsymbol{r},\boldsymbol{\eta})=\hat{Q}_h(\boldsymbol{r})+\|\lambda^{(\ell-1)}\circ\boldsymbol{\beta}_-\|_1+\langle\boldsymbol{\eta},\boldsymbol{r}-\mathbf{y}+\mathbf{X}\boldsymbol{\beta}\rangle+\frac{\rho}{2}\|\boldsymbol{r}-\mathbf{y}+\mathbf{X}\boldsymbol{\beta}\|_2^2,\tag{11}$$

where $\boldsymbol{\eta}$ is the Lagrange multiplier and ρ is a tuning parameter for the ADMM algorithm. Updates for the ADMM can be derived by minimizing each parameter while keeping the others fixed. We summarize the details in Algorithm 3.

Algorithm 3. ADMM Algorithm for Solving (2.8) with Gaussian Kernel.

Input quantile parameter τ, smoothing parameter h, regularization parameter $\boldsymbol{\lambda}^{(\ell-1)}$, and the convergence criterion ϵ.

Initialize the primal variables $\widehat{\boldsymbol{\beta}}^{(0)}=\widehat{\boldsymbol{r}}^{(0)}=\mathbf{0}$ and the dual variable $\widehat{\boldsymbol{\eta}}^{(0)}=\mathbf{0}$.

Iterate the following until the stopping criterion $\|\widehat{\boldsymbol{\beta}}^{(t)}-\widehat{\boldsymbol{\beta}}^{(t-1)}\|_2\leq\epsilon$ is met:

1. Update $\boldsymbol{\beta}$ as

$$\widehat{\boldsymbol{\beta}}^{(t)}=\underset{\boldsymbol{\beta}\in\mathbb{R}^p}{\text{argmin}}\left\{\frac{\rho}{2}\left\|\mathbf{y}-\widehat{\boldsymbol{r}}^{(t-1)}-\frac{1}{\sqrt{\rho}}\widehat{\boldsymbol{\eta}}^{(t-1)}-\mathbf{X}\boldsymbol{\beta}\right\|_2^2+\|\boldsymbol{\lambda}^{(\ell-1)}\circ\boldsymbol{\beta}_-\|_1\right\}.$$

2. Iterate the following until convergence: for each $i=1,\dots,n$, update r_i by solving

$$\tau-\Phi\left(\frac{-r_i}{h}\right)+\widehat{\eta}_i^{(t-1)}+\rho(r_i-y_i+\langle\boldsymbol{x}_i,\widehat{\boldsymbol{\beta}}^{(t)}\rangle)=0.$$

3. Update $\boldsymbol{\eta}$ as

$$\widehat{\boldsymbol{\eta}}^{(t)}=\widehat{\boldsymbol{\eta}}^{(t-1)}+\rho(\widehat{\boldsymbol{r}}^{(t)}-\mathbf{y}+\mathbf{X}\widehat{\boldsymbol{\beta}}^{(t)}).$$

Output the estimated parameter $\widehat{\boldsymbol{\beta}}^{(t)}$.

The updates for $\boldsymbol{\beta}$ involve solving a Lasso regression problem for which efficient software is available. Alternatively, one can also linearize the loss function as in Gu et al. (2018) to obtain a closed-form solution. The updates for $\boldsymbol{r}$ can be obtained using coordinate descent algorithm by updating each coordinate of $\boldsymbol{r}$ using standard numerical methods such as the bisection method. See Algorithm 3 for details.

4 | STATISTICAL THEORY

In this section, we provide a comprehensive analysis of the sequence of regularized QR estimators $\{\hat{\boldsymbol{\beta}}^{(\ell)}\}_{\ell \geq 1}$ obtained by solving Equation (6) iteratively, initialized with $\hat{\boldsymbol{\beta}}^{(0)} = \mathbf{0}$. For simplicity, we restrict our attention to a fixed quantile level $\tau \in (0,1)$ of interest. We first characterize the (deterministic) bias induced by convolution smoothing described in Section 4.1. In Section 4.2, we provide high probability bounds (under ℓ_1- and ℓ_2-errors) for the one-step estimator $\hat{\boldsymbol{\beta}}^{(1)}$, that is, the ℓ_1-penalized smoothed QR estimator (ℓ_1-SQR) which is of independent interest. With a flexible choice of the bandwidth h, these error bounds for $\hat{\boldsymbol{\beta}}^{(1)}$ are near-minimax optimal (Wang & He, 2021), and coincide with those of the ℓ_1-QR estimator Belloni and Chernozhukov (2011). In Section 4.3, we analyse $\hat{\boldsymbol{\beta}}^{(\ell)}$ ($\ell \geq 2$) whose overall estimation error consists of three parts: shrinkage bias, oracle rate, and smoothing bias. Our analysis reveals that the multi-step iterative algorithm refines the statistical rate in a sequential manner: every relaxation step shrinks the estimation error from the previous step by a δ-fraction for some $\delta \in (0,1)$. Under a necessary beta-min condition, we show that the multi-step estimator $\hat{\boldsymbol{\beta}}^{(\ell)}$ with $\ell \gtrsim \log\{\log(p)\}$ achieves the oracle rate of convergence, that is, it shares the convergence rate of the oracle estimator that has access to the true active set. Under a sub-Gaussian condition on the feature vector and a stronger sample size requirement, we further show in Section 4.4 that the multi-step estimator $\hat{\boldsymbol{\beta}}^{(\ell)}$ with $\ell \gtrsim \log(s)$ coincides with the oracle estimator with high probability, and hence achieves variable selection consistency. Throughout, we use the notation '$\lesssim$' to indicate ' $\leq$ ' up to constants that are independent of (s, p, n).

4.1 | Smoothing bias

To begin with, note that the smoothed quantile objective $\hat{Q}_h(\cdot)$ defined in Equation (4) can be written as

$$\hat{Q}_h(\boldsymbol{\beta}) = (1-\tau)\int_{-\infty}^{0} \hat{F}_h(u;\boldsymbol{\beta})\,\mathrm{d}u + \tau\int_{0}^{\infty} \{1-\hat{F}_h(u;\boldsymbol{\beta})\}\,\mathrm{d}u.$$

Recall the integrated kernel function $\overline{K}(u) = \int_{-\infty}^{u} K(t)\,\mathrm{d}t$, which is non-decreasing and takes values in [0, 1]. With $r_i(\boldsymbol{\beta}) = y_i - \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}$, the gradient vector and Hessian matrix of $\hat{Q}_h(\boldsymbol{\beta})$ are, respectively,

$$\nabla\hat{Q}_h(\boldsymbol{\beta}) = \frac{1}{n}\sum_{i=1}^{n}\left\{\overline{K}\left(-r_i(\boldsymbol{\beta})/h\right) - \tau\right\}\boldsymbol{x}_i \quad \text{and} \quad \nabla^2\hat{Q}_h(\boldsymbol{\beta}) = \frac{1}{n}\sum_{i=1}^{n} K_h(-r_i(\boldsymbol{\beta}))\boldsymbol{x}_i\boldsymbol{x}_i^{\mathrm{T}}. \qquad (12)$$

To examine the bias induced by smoothing, define the expected smoothed loss function $Q_h(\boldsymbol{\beta}) = \mathbb{E}\{\hat{Q}_h(\boldsymbol{\beta})\}$, $\boldsymbol{\beta} \in \mathbb{R}^p$, and the pseudo parameter

$$\boldsymbol{\beta}_h^* = (\beta_{h,1}^*, \dots, \beta_{h,p}^*)^{\mathrm{T}} \in \operatorname*{argmin}_{\boldsymbol{\beta} \in \mathbb{R}^p} Q_h(\boldsymbol{\beta}), \tag{13}$$

which is the population minimizer of the smoothed quantile loss and varies with h. In general, $\boldsymbol{\beta}_h^*$ differs from $\boldsymbol{\beta}^*$—the unknown parameter vector in model (1). The latter is identified as the unique minimizer of the population quantile objective $Q(\boldsymbol{\beta}) := \mathbb{E}\{\hat{Q}(\boldsymbol{\beta})\}$. However, as the smoothed quantile loss $\ell_h(\cdot)$ in Equation (5) approximates the quantile loss $\rho_\tau(\cdot)$ as $h = h_n \to 0$, $\boldsymbol{\beta}_h^*$ is expected to converge to $\boldsymbol{\beta}^*$, and we refer to $\|\boldsymbol{\beta}_h^* - \boldsymbol{\beta}^*\|_2$ as the approximation error or bias due to smoothing.

The following result provides upper bounds of the smoothing bias under mild conditions on the random covariates $\boldsymbol{x} \in \mathbb{R}^p$, the conditional density of ε given $\boldsymbol{x}$, and the kernel function. Throughout Section 4, we assume that the second moment $\boldsymbol{\Sigma} = (\sigma_{jk})_{1 \le j,k \le p} = \mathbb{E}(\boldsymbol{x}\boldsymbol{x}^{\mathrm{T}})$ of $\boldsymbol{x} = (x_1, \dots, x_p)^{\mathrm{T}}$ (with $x_1 \equiv 1$) exists and is positive definite. Moreover, let $\gamma_1 = \gamma_1(\boldsymbol{\Sigma}) \ge 1$, $\gamma_p = \gamma_p(\boldsymbol{\Sigma}) \in (0, 1]$, and $\sigma_{\boldsymbol{x}}^2 = \max_{1 \le j \le p} \sigma_{jj}$.

(B1) The conditional density of ε given $\boldsymbol{x}$, denoted by $f_{\varepsilon|\boldsymbol{x}}$, satisfies $f_l \le f_{\varepsilon|\boldsymbol{x}}(0) \le f_u$ almost surely (over $\boldsymbol{x}$) for some $f_u \ge f_l > 0$. Moreover, there exists a constant $l_0 > 0$ such that $|f_{\varepsilon|\boldsymbol{x}}(u) - f_{\varepsilon|\boldsymbol{x}}(v)| \le l_0 |u - v|$ for all $u, v \in \mathbb{R}$ almost surely (over $\boldsymbol{x}$).

(B2) The kernel function $K : \mathbb{R} \to [0, \infty)$ is symmetric around zero, and satisfies $\int_{-\infty}^{\infty} K(u)\,\mathrm{d}u = 1$ and $\int_{-\infty}^{\infty} u^2 K(u)\,\mathrm{d}u < \infty$. For $\ell = 1, 2, \dots$, let $\kappa_\ell = \int_{-\infty}^{\infty} |u|^\ell K(u)\,\mathrm{d}u$ be the ℓth absolute moment of $K(\cdot)$.

Proposition 1 *Assume that Conditions* (B1) *and* (B2) *hold, and* $\mu_3 := \sup_{\boldsymbol{u} \in \mathbb{S}^{p-1}} \mathbb{E}|\boldsymbol{z}^{\mathrm{T}}\boldsymbol{u}|^3 < \infty$ *with* $\boldsymbol{z} = \boldsymbol{\Sigma}^{-1/2}\boldsymbol{x}$. *Provided* $0 < h < f_l/(c_0 l_0)$, $\boldsymbol{\beta}_h^*$ *is the unique minimizer of* $\boldsymbol{\beta} \mapsto Q_h(\boldsymbol{\beta})$ *and satisfies*

$$\|\boldsymbol{\beta}_h^* - \boldsymbol{\beta}^*\|_{\boldsymbol{\Sigma}} \le c_0 l_0 f_l^{-1} h^2, \tag{14}$$

where $c_0 = (\mu_3 + \kappa_2)/2 + \kappa_1$. *In addition, assume* $\kappa_3 < \infty$ *and* $f_{\varepsilon|\boldsymbol{x}}$ *has an* l_1*-Lipschitz continuous derivative almost everywhere for some* $l_1 > 0$. *Then*

$$\left\| \boldsymbol{\Sigma}^{-1}\mathbf{J}(\boldsymbol{\beta}_h^* - \boldsymbol{\beta}^*) + \frac{1}{2}\kappa_2 h^2 \cdot \boldsymbol{\Sigma}^{-1}\mathbb{E}\left\{ f_{\varepsilon|\boldsymbol{x}}'(0)\boldsymbol{x} \right\} \right\|_{\boldsymbol{\Sigma}} \le C h^3, \tag{15}$$

where $\mathbf{J} = \mathbb{E}\{f_{\varepsilon|\boldsymbol{x}}(0) \cdot \boldsymbol{x}\boldsymbol{x}^{\mathrm{T}}\}$, *and* $C > 0$ *depends only on* (f_l, l_0, l_1, μ_3) *and the kernel* K.

Proposition 1 is a non-asymptotic version of Theorem 1 in Fernandes et al. (2021), and explicitly captures the dependence of the bias on several model-based quantities. Note that the $p \times p$ matrix $\mathbf{J} = \mathbb{E}\{f_{\varepsilon|\boldsymbol{x}}(0) \cdot \boldsymbol{x}\boldsymbol{x}^{\mathrm{T}}\}$ is the Hessian of the population quantile objective $Q(\cdot)$ evaluated at $\boldsymbol{\beta}^*$, that is, $\mathbf{J} = \nabla^2 Q(\boldsymbol{\beta}^*)$. Under Condition (B1), $f_l \gamma_p(\boldsymbol{\Sigma}) \le \gamma_p(\mathbf{J}) \le \gamma_1(\mathbf{J}) \le f_u \gamma_1(\boldsymbol{\Sigma})$. An interesting implication of Proposition 1 is that, when both $f_{\varepsilon|\boldsymbol{x}}(0)$ and $f_{\varepsilon|\boldsymbol{x}}'(0)$ are independent of $\boldsymbol{x}$ (i.e., $f_{\varepsilon|\boldsymbol{x}}(0) = f_\varepsilon(0)$ and $f_{\varepsilon|\boldsymbol{x}}'(0) = f_\varepsilon'(0)$), the bias decomposition bound (15) simplifies to

$$\left\| f_\varepsilon(0)(\boldsymbol{\beta}_h^* - \boldsymbol{\beta}^*) + 0.5 f_\varepsilon'(0)\kappa_2 h^2 \begin{bmatrix} 1 \\ \mathbf{0}_{p-1} \end{bmatrix} \right\|_{\boldsymbol{\Sigma}} \le C h^3.$$

In other words, the smoothing bias is concentrated primarily on the intercept. To some extent, this observation further certifies the benefit of smoothing in variable selection of which the main focus is on the slope coefficients rather than the intercept.

4.2 | ℓ_1-penalized smoothed quantile regression

Given a bandwidth $h > 0$ and a regularization parameter $\lambda > 0$, let $\hat{\boldsymbol{\beta}}_h = \hat{\boldsymbol{\beta}}_h(\tau, \lambda)$ be the ℓ_1-penalized SQR (ℓ_1-SQR) estimator, defined as the solution to the following convex optimization problem:

$$\min_{\boldsymbol{\beta}\in\mathbb{R}^p} \left\{ \hat{Q}_h(\boldsymbol{\beta}) + \lambda\|\boldsymbol{\beta}\|_1 \right\}. \tag{16}$$

In this section, we characterize the estimation error of $\hat{\boldsymbol{\beta}}_h \in \mathbb{R}^p$ under ℓ_2- and ℓ_1-norms. First we impose a moment condition on the (random) covariate vector $\boldsymbol{x} = (x_1, \ldots, x_p)^{\mathrm{T}} \in \mathbb{R}^p$ with $x_1 \equiv 1$. Without loss of generality, assume $\mu_j = \mathbb{E}(x_j) = 0$ for $2 \le j \le p$; otherwise, consider a change of variable $(\beta_1, \beta_2, \ldots, \beta_p)^{\mathrm{T}} \mapsto (\beta_1 + \sum_{j=2}^p \mu_j\beta_j, \beta_2, \ldots, \beta_p)^{\mathrm{T}}$ so that the obtained results apply to model $F^{-1}_{y|\boldsymbol{x}}(\tau) = \beta_0^\flat + \sum_{j=2}^p (x_j - \mu_j)\beta_j^*$, where $\beta_0^\flat = \beta_0^* + \sum_{j=2}^p \mu_j\beta_j^*$.

(B3) $\boldsymbol{\Sigma} = \mathbb{E}(\boldsymbol{x}\boldsymbol{x}^{\mathrm{T}})$ is positive definite and $\boldsymbol{z} = \boldsymbol{\Sigma}^{-1/2}\boldsymbol{x} \in \mathbb{R}^p$ is sub-exponential: there exist constants $v_0, c_0 \ge 1$ such that $\mathbb{P}(|\boldsymbol{z}^{\mathrm{T}}\boldsymbol{u}| \ge v_0\|\boldsymbol{u}\|_2 \cdot t) \le c_0 e^{-t}$ for all $\boldsymbol{u} \in \mathbb{R}^p$ and $t \ge 0$. For convenience, we assume $c_0 = 1$, and write $\sigma_{\boldsymbol{x}}^2 = \max_{1\le j\le p} \mathbb{E}(x_j^2)$.

Moreover, for $r, l > 0$, define the (rescaled) ℓ_2-ball and ℓ_1-cone as

$$\mathbb{B}_{\boldsymbol{\Sigma}}(r) = \{\boldsymbol{\delta} \in \mathbb{R}^p : \|\boldsymbol{\delta}\|_{\boldsymbol{\Sigma}} \le r\} \text{ and } \mathbb{C}_{\boldsymbol{\Sigma}}(l) = \left\{\boldsymbol{\delta} \in \mathbb{R}^p : \|\boldsymbol{\delta}\|_1 \le l\|\boldsymbol{\delta}\|_{\boldsymbol{\Sigma}}\right\}. \tag{17}$$

Our theoretical analysis of the ℓ_1-SQR estimator depends crucially on the following 'good' event, which is related to the local restricted strong convexity (RSC) of the empirical smoothed quantile loss function. We refer the reader to Negahban et al. (2012) and Loh and Wainwright (2015) for detailed discussions of the restricted strong convexity for regularized M-estimation in high dimensions.

Definition 1 (Local restricted strong convexity). Given radius parameters $r, l > 0$ and a curvature parameter $\kappa > 0$, define the event

$$\mathcal{E}_{\mathrm{rsc}}(r, l, \kappa) = \left\{ \frac{\langle \nabla\hat{Q}_h(\boldsymbol{\beta}) - \nabla\hat{Q}_h(\boldsymbol{\beta}^*), \boldsymbol{\beta} - \boldsymbol{\beta}^* \rangle}{\|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_{\boldsymbol{\Sigma}}^2} \ge \kappa \text{ for all } \boldsymbol{\beta} \in \boldsymbol{\beta}^* + \mathbb{B}_{\boldsymbol{\Sigma}}(r) \cap \mathbb{C}_{\boldsymbol{\Sigma}}(l) \right\}. \tag{18}$$

Our first result shows that, with suitably chosen (r,l,κ), the event $\mathcal{E}_{\mathrm{rsc}}(r, l, \kappa)$ occurs with high probability. In order for the local RSC condition to hold, the radius parameter r has to be of the same order as, or possibly smaller than the bandwidth h.

Proposition 2 *Assume Conditions* (B1)–(B3) *hold, and* $\kappa_l = \min_{|u|\le 1} K(u) > 0$. *Moreover, let* (r, l, h) *and* n *satisfy*

$$20v_0^2 r \le h \le f_l/(2l_0) \quad \text{and} \quad n \ge C\sigma_{\boldsymbol{x}}^2 f_u f_l^{-2}(l/r)^2 h \log(2p) \tag{19}$$

for a sufficiently large constant C. *Then, the local RSC event* $\mathcal{E}_{\mathrm{rsc}}(r, l, \kappa)$ *with* $\kappa = (\kappa_l f_l)/2$ *occurs with probability at least* $1 - (2p)^{-1}$.

Remark 3 We do not claim that the values of the constants appearing in Proposition 2 are optimal. They result from non-asymptotic probabilistic bounds which reflect worst-case scenarios. The condition $\min_{|u|\le 1} K(u) > 0$ is only for theoretical and notational convenience. If the kernel $K(\cdot)$ is compactly supported on $[-1, 1]$, we may rescale it to obtain $K_a(u) = (1/a)K(u/a)$ for some $a > 1$. Then, $K_a(\cdot)$ is supported on $[-a, a]$ with $\min_{|u|\le 1} K(u) > 0$. For example,

1. (Gaussian kernel) if $K(u) = (2\pi)^{-1/2}e^{-u^2/2}$ is the Gaussian kernel, we have $\kappa_l = (2\pi e)^{-1/2} \approx 0.242$ and $\kappa_2 = 1$;
2. (Uniform kernel) if $K(u) = (1/2)\mathbb{1}(|u| \le 1)$ is the uniform kernel, we may consider its rescaled version $K_{3/2}(u) = (1/3)\mathbb{1}(|u| \le 3/2)$. In this case, $\kappa_l = 1/3$ and $\kappa_2 = 3/4$.

Throughout, we view (κ_l, κ_2) as absolute constants.

Theorem 1 *Under the conditional quantile model* (1) *with $\boldsymbol{\beta}^* \in \mathbb{R}^p$ being s-sparse, assume Conditions (B1)–(B3) hold with* $\kappa_l = \min_{|u|\le 1} K(u) > 0$. *Then, the ℓ_1-SQR estimator $\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}_h$ with* $\lambda \asymp \sigma_{\boldsymbol{x}}\sqrt{\tau(1-\tau)\log(p)/n}$ *satisfies the bounds*

$$\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2 \le C_1 f_l^{-1} s^{1/2}\lambda \text{ and } \|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_1 \le C_2 f_l^{-1} s\lambda \tag{20}$$

with probability at least $1 - p^{-1}$, *provided that the bandwidth satisfies*

$$\max\left(\frac{\sigma_{\boldsymbol{x}}}{f_l}\sqrt{\frac{s\log p}{n}}, \frac{\sigma_{\boldsymbol{x}}^2 f_u}{f_l^2}\frac{s\log p}{n}\right) \lesssim h \le \min\{f_l/(2l_0), (s^{1/2}\lambda)^{1/2}\},$$

where the constants $C_1, C_2 > 0$ *depend only on* $(l_0, v_0, \gamma_p, \kappa_l, \kappa_2)$.

The above theorem shows that with a proper yet flexible choice of the bandwidth, the ℓ_1-penalized smoothed QR estimator achieves the same rate of convergence as the ℓ_1-QR estimator under both ℓ_1- and ℓ_2-errors (Belloni & Chernozhukov, 2011). Technically, we assume the random feature vector is sub-exponential, which is arguably the weakest moment condition in high-dimensional regression analysis under random design (Wainwright, 2019). This preliminary result is of independent interest, and more importantly, it paves the way for further analysis of smoothed QR with iteratively reweighted ℓ_1-regularization.

4.3 | Concave regularization and oracle rate of convergence

In this section, we derive rates of convergence for the solution path $\{\hat{\boldsymbol{\beta}}^{(\ell)}\}_{\ell=1,2,\ldots}$ of the multi-step iterative algorithm defined in Equation (6). Starting from $\hat{\boldsymbol{\beta}}^{(0)} = \mathbf{0}$, we note that $\hat{\boldsymbol{\beta}}^{(1)}$ is exactly the ℓ_1-SQR estimator studied in the previous section; see Theorem 1. For subsequent $\hat{\boldsymbol{\beta}}^{(\ell)}$'s, we first state the result as a deterministic claim in Theorem 2, but conditioned on some 'good' event regarding the local RSC property and the gradient of $\hat{Q}_h(\cdot)$ at $\boldsymbol{\beta}^*$. Under Condition (B3) on the random covariate vector, probabilistic claims enter in certifying that this 'good' event holds with high probability with a suitable choice of λ and h; see Theorem 3.

Recall the event $\mathcal{E}_{\mathrm{rsc}}(r, l, \kappa)$ defined in Equation (18) on which a local RSC property of the smoothed quantile objective $\hat{Q}_h(\cdot)$ holds, where κ is a curvature parameter. Moreover, define

$$\boldsymbol{w}_h^* = \boldsymbol{w}_h(\boldsymbol{\beta}^*) \in \mathbb{R}^p \text{ and } b_h^* = \|\boldsymbol{\Sigma}^{-1/2}\nabla Q_h(\boldsymbol{\beta}^*)\|_2, \tag{21}$$

where $\boldsymbol{w}_h(\boldsymbol{\beta}) = \nabla \hat{Q}_h(\boldsymbol{\beta}) - \nabla Q_h(\boldsymbol{\beta})$ is the centred score function, and $b_h^* \geq 0$ quantifies the bias induced by smoothing. For the standard quantile loss, we have $\nabla Q(\boldsymbol{\beta}^*) = \mathbf{0}$. Under Conditions (B1) and (B2), examine the proof of Proposition 1 yields $b_h^* \leq l_0 \kappa_2 h^2/2$, that is, the smoothing bias has magnitude of the order h^2. To refine the statistical rate obtained in Theorem 1, which is near-minimax optimal for estimating sparse targets, we need an additional beta-min condition on $\|\boldsymbol{\beta}_{\mathcal{S}}^*\|_{\min} = \min_{j\in\mathcal{S}} |\beta_j^*|$, where $\mathcal{S} = \{1 \leq j \leq p : \beta_j^* \neq 0\}$ is the active set of $\boldsymbol{\beta}^*$. For a deterministic analysis, we first derive the contraction property of the solution path $\{\hat{\boldsymbol{\beta}}^{(\ell)}\}_{\ell\geq 1}$ conditioned on some 'good' event.

Theorem 2 *Given $\kappa > 0$ and a penalty function $q(\cdot)$ satisfying (A1), assume that there exists some constant $\alpha_0 > 0$ such that*

$$\frac{\alpha_0}{\sqrt{1 + \{q'(\alpha_0)/2\}^2}} > \frac{1}{\kappa\gamma_p} \text{ and } q'(\alpha_0) > 0. \tag{22}$$

Let the penalty level λ and bandwidth h satisfy $b_h^* \leq (s/\gamma_p)^{1/2}\lambda$. Moreover, define $r_{\rm opt} = \gamma_p^{1/2}\alpha_0 c s^{1/2}\lambda$ and $l = \{(2 + \frac{2}{q'(\alpha_0)})(c^2+1)^{1/2} + \frac{2}{q'(\alpha_0)}\}(s/\gamma_p)^{1/2}$, where the constant $c > 0$ is defined through the equation

$$0.5q'(\alpha_0)(c^2+1)^{1/2} + 2 = \alpha_0\kappa\gamma_p \cdot c. \tag{23}$$

Then, for any $r \geq r_{\rm opt}$, conditioned on the event $\mathcal{E}_{\rm rsc}(r, l, \kappa) \cap \{\|\boldsymbol{w}_h^*\|_\infty \leq 0.5q'(\alpha_0)\lambda\}$, the sequence of solutions $\{\hat{\boldsymbol{\beta}}^{(\ell)}\}_{\ell\geq 1}$ to programs (6) satisfies

$$\|\hat{\boldsymbol{\beta}}^{(\ell)} - \boldsymbol{\beta}^*\|_{\boldsymbol{\Sigma}} \leq \delta \cdot \|\hat{\boldsymbol{\beta}}^{(\ell-1)} - \boldsymbol{\beta}^*\|_{\boldsymbol{\Sigma}} + \underbrace{\kappa^{-1}\gamma_p^{-1/2}\Big\{\|q_\lambda'((|\boldsymbol{\beta}_{\mathcal{S}}^*| - \alpha_0\lambda)_+)\|_2 + \|\boldsymbol{w}_{h,\mathcal{S}}^*\|_2\Big\}}_{=:\, r_{\rm ora}} + \kappa^{-1}b_h^*, \tag{24}$$

where $\delta = \sqrt{1 + \{q'(\alpha_0)/2\}^2}/(\alpha_0\kappa\gamma_p) \in (0,1)$ and $u_+ = \max(u, 0)$. In addition,

$$\|\hat{\boldsymbol{\beta}}^{(\ell)} - \boldsymbol{\beta}^*\|_{\boldsymbol{\Sigma}} \leq \delta^{\ell-1} r_{\rm opt} + (1-\delta)^{-1}(r_{\rm ora} + \kappa^{-1}b_h^*) \text{ for any } \ell \geq 2. \tag{25}$$

Theorem 2 reveals how iteratively reweighted ℓ_1-penalization refines the statistical rate in a sequential manner: every relaxation step shrinks the estimation error from the previous step by a δ-fraction. The error term that does not vary with reweighted penalization consists of

$$\underbrace{\left\|q_\lambda'((|\boldsymbol{\beta}_{\mathcal{S}}^*| - \alpha_0\lambda)_+)\right\|_2}_{\text{shrinkage bias}}, \quad \underbrace{\left\|\boldsymbol{w}_{h,\mathcal{S}}^*\right\|_2}_{\text{oracle rate}}, \quad \text{and} \quad \underbrace{b_h^*}_{\text{smoothing bias}}.$$

The first term $\|q_\lambda'((|\boldsymbol{\beta}_{\mathcal{S}}^*| - \alpha_0\lambda)_+)\|_2$ is known as the shrinkage bias induced by the folded-concave penalty function (Fan et al., 2018). For the ℓ_1-norm penalty, that is, $q_\lambda(t) = \lambda|t|$ and$q_\lambda'(t) =$

$\lambda\, sign(t)$, the shrinkage bias can be as large as $s^{1/2}\lambda$. Without any prior knowledge on the signal strength, we have $\|q'_\lambda((|\boldsymbol{\beta}^*_{\mathcal{S}}| - \alpha_0\lambda)_+)\|_2 \le \|q'_\lambda(\mathbf{0}_{\mathcal{S}})\|_2 = s^{1/2}\lambda$ for any penalty q_λ satisfying Condition (A1). Assume $q_\lambda(t) = \lambda^2 q(t/\lambda)$ is a concave penalty defined on $\mathbb{R}^+$ with $\alpha_* := \inf\{\alpha > 0 : q'(\alpha) = 0\} < \infty$. Given a regularization parameter $\lambda > 0$, consider the decomposition $\mathcal{S} = \mathcal{S}_0 \cup \mathcal{S}_1$, where

$$\mathcal{S}_0 = \left\{ j \in \mathcal{S} : |\beta_j| < (\alpha_0 + \alpha_*)\lambda \right\} \text{ and } \mathcal{S}_1 = \left\{ j \in \mathcal{S} : |\beta_j| \ge (\alpha_0 + \alpha_*)\lambda \right\}$$

have cardinalities s_0 and s_1 respectively. The shrinkage bias term can then be bounded by

$$\|q'_\lambda((|\boldsymbol{\beta}^*_{\mathcal{S}}| - \alpha_0\lambda)_+)\|_2 \le \|q'_\lambda(\mathbf{0}_{\mathcal{S}_0})\|_2 = s_0^{1/2}\lambda.$$

Under the beta-min condition $\|\boldsymbol{\beta}^*_{\mathcal{S}}\|_{\min} \ge (\alpha_0 + \alpha_*)\lambda$, the shrinkage bias vanishes, and hence the final rate of convergence is determined by $\|\boldsymbol{w}^*_{h,\mathcal{S}}\|_2$ and b^*_h. As previously noted, the latter is the smoothing bias term, and satisfies $b^*_h \le l_0\kappa_2 h^2/2$.

The terminology 'oracle' stems from the 'oracle estimator', defined as the QR estimator that knows in advance the true subset of the important features. For a better comparison, we define the oracle smoothed QR estimator as

$$\hat{\boldsymbol{\beta}}^{\text{ora}} = \underset{\boldsymbol{\beta}\in\mathbb{R}^p : \boldsymbol{\beta}_{\mathcal{S}^c}=\mathbf{0}}{\operatorname{argmin}} \hat{Q}_h(\boldsymbol{\beta}) = \underset{\boldsymbol{\beta}\in\mathbb{R}^p : \boldsymbol{\beta}_{\mathcal{S}^c}=\mathbf{0}}{\operatorname{argmin}} \frac{1}{n}\sum_{i=1}^n \ell_h(y_i - \boldsymbol{x}^{\mathrm{T}}_{i,\mathcal{S}}\boldsymbol{\beta}_{\mathcal{S}}), \tag{26}$$

where $\ell_h(\cdot)$ is the smoothed quantile loss given in Equation (5). As we will show in Section 4.4, the oracle SQR estimator $\hat{\boldsymbol{\beta}}^{\text{ora}}$ satisfies the bound

$$\|\hat{\boldsymbol{\beta}}^{\text{ora}} - \boldsymbol{\beta}^*\|_2 \lesssim \|\boldsymbol{w}^*_{h,\mathcal{S}}\|_2 + h^2$$

with high probability, and $\|\boldsymbol{w}^*_{h,\mathcal{S}}\|_2$ is of order $\sqrt{s/n}$.

Theorem 2 is a deterministic result. Probabilistic claims enter in certifying that the local RSC condition holds with high probability (see Proposition 2), and in verifying that the 'good' event $\{\|\boldsymbol{w}^*_h\|_\infty \le 0.5q'(\alpha_0)\lambda\}$ occurs with high probability with a specified choice of λ. The following theorem states, under a necessary beta-min condition, the iteratively reweighted ℓ_1-penalized SQR (IRW-ℓ_1-SQR) estimator $\hat{\boldsymbol{\beta}}^{(\ell)}$, after a few iterations, achieves the estimation error of the oracle that knows the sparsity pattern of $\boldsymbol{\beta}^*$.

Theorem 3 *In addition to Conditions* (A1), (B1)–(B3), *assume there exist* $\alpha_1 > \alpha_0 > 0$ *such that*

$$q'(\alpha_0) > 0, \quad \frac{\alpha_0}{\sqrt{4 + \{q'(\alpha_0\}^2}} > (\kappa_l f_l \gamma_p)^{-1} \quad \text{and} \quad q'(\alpha_1) = 0, \tag{27}$$

where $\kappa_l = \min_{|u|\le 1} K(u) > 0$. *Moreover, let the regularization parameter* λ *and bandwidth* h *satisfy* $\lambda \asymp \sigma_{\boldsymbol{x}}\sqrt{\tau(1-\tau)\log(p)/n}$ *and*

$$\max\left(\frac{\sigma_{\boldsymbol{x}}}{f_l}\sqrt{\frac{s\log p}{n}}, \frac{\sigma^2_{\boldsymbol{x}} f_u}{f_l^2}\frac{s\log p}{n}\right) \lesssim h \lesssim (s^{1/2}\lambda)^{1/2}.$$

For any $t \geq 0$, under the beta-min condition $\|\boldsymbol{\beta}^*_S\|_{\min} \geq (\alpha_0 + \alpha_1)\lambda$ and scaling $n \gtrsim \max\{s \log(p), s + t\}$, the IRW-$\ell_1$-SQR *estimator $\hat{\boldsymbol{\beta}}^{(\ell)}$ with $\ell \gtrsim \lceil \log\{\log(p)\}/\log(1/\delta) \rceil$ satisfies the bounds*

$$\|\hat{\boldsymbol{\beta}}^{(\ell)} - \boldsymbol{\beta}^*\|_2 \lesssim f_l^{-1}\left(\sqrt{\frac{s+t}{n}} + h^2\right) \text{ and } \|\hat{\boldsymbol{\beta}}^{(\ell)} - \boldsymbol{\beta}^*\|_1 \lesssim f_l^{-1} s^{1/2}\left(\sqrt{\frac{s+t}{n}} + h^2\right) \tag{28}$$

with probability at least $1 - p^{-1} - e^{-t}$, *where* $\delta = \sqrt{4 + \{q'(\alpha_0)\}^2}/(\alpha_0 \kappa_l f_l \gamma_p) \in (0, 1)$.

Remark 4 (Oracle rate of convergence and high-dimensional scaling). The conclusion of Theorem 3 is referred to as the weak oracle property: the IRW-ℓ_1-SQR estimator achieves the convergence rate of the oracle $\hat{\boldsymbol{\beta}}^{\mathrm{ora}}$ when the support set S were known a priori. Starting from $\hat{\boldsymbol{\beta}}^{(0)} = \mathbf{0}$, the one-step estimator $\hat{\boldsymbol{\beta}}^{(1)}$ (ℓ_1-SQR) has an estimation error (under ℓ_2-norm) of order $\sqrt{s \cdot \log(p)/n}$ (see Theorem 1). Under an almost necessary and sufficient beta-min condition—$\|\boldsymbol{\beta}^*_S\|_{\min} \gtrsim \sqrt{\log(p)/n}$, a refined near-oracle statistical rate $\sqrt{s/n} + h^2$ can be attained by a multi-step iterative procedure, which solves a sequence of convex programs. Here, $\sqrt{s/n}$ is referred to as the oracle rate, and the h^2-term quantifies the smoothing bias (Proposition 1). In order to certify the local RSC property of the smoothed objective function, the bandwidth should have magnitude at least of the order $\sqrt{s \log(p)/n}$. If we choose a bandwidth $h \asymp \sqrt{s \log(p)/n}$, the ℓ_2-error of the multi-step estimator will be of order $\sqrt{s/n} + s \log(p)/n$ under the high-dimensional scaling $n \gtrsim s \log(p)$. Intuitively, the main reason for having an extra term $s \log(p)/n$ is that even if the underlying vector $\boldsymbol{\beta}^*$ is s-sparse, the population parameter $\boldsymbol{\beta}^*_h \in \mathbb{R}^p$ corresponding to the smoothed objective function (see Equation (13)) may be denser. As a result, there is a statistical price to pay for smoothing.

Remark 5 (Minimum signal strength and oracle rate). In a linear regression model $y = \boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta}^* + \varepsilon$ with a Gaussian error $\varepsilon \sim N(0, \sigma^2)$, consider the parameter space $\Omega_{s,a} = \{\boldsymbol{\beta} \in \mathbb{R}^p : \|\boldsymbol{\beta}\|_0 \leq s, \min_{j:\beta_j \neq 0} |\beta_j| \geq a\}$ for $a > 0$. Assuming that the design matrix $\mathbb{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)^{\mathrm{T}} \in \mathbb{R}^{n \times p}$ satisfies a restricted isometry property and has normalized columns (each column has an ℓ_2-norm equal to $\sqrt{n}$), Ndaoud (2019) derived the following sharp lower bounds for the minimax risk $\psi(s, a) := \inf_{\hat{\beta}} \sup_{\boldsymbol{\beta}^* \in \Omega_{s,a}} \mathbb{E}\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2^2$: for any $\varepsilon \in (0,1)$,

$$\psi(s, a) \geq \{1 + o(1)\}\frac{2\sigma^2 s \log(ep/s)}{n} \quad \text{for any } a \leq (1 - \varepsilon)\sigma\sqrt{\frac{2 \log(ep/s)}{n}}$$

and

$$\psi(s, a) \geq \{1 + o(1)\}\frac{\sigma^2 s}{n} \quad \text{for any } a \geq (1 + \varepsilon)\sigma\sqrt{\frac{2 \log(ep/s)}{n}},$$

where the limit corresponds to $s/p \to 0$ and $s \log(ep/s)/n \to 0$. The minimax rate $2\sigma^2 s \log(ep/s)/n$ can be attained by both Lasso and Slope (Bellec et al., 2018), while the oracle rate $\sigma^2 s/n$ can only be achieved when the magnitude of the minimum signal is of order $\sigma\sqrt{\log(p/s)/n}$. For estimating an s-sparse vector $\boldsymbol{\beta}^* \in \mathbb{R}^p$ in the conditional quantile model (1), Wang and He (2021) proved the lower bound $\sqrt{s \log(p/s)/n}$ for the minimax estimation error under ℓ_2-norm. In order to achieve the refined oracle rate, Fan et al. (2014)

required a stronger beta-min condition, that is, $\|\boldsymbol{\beta}^*_{\mathcal{S}}\|_{\min} \gtrsim \sqrt{s \log(p)/n}$, and a stringent independence assumption between ε and $\boldsymbol{x}$ in the conditional quantile model (1). The beta-min condition imposed in Theorems 2 and 3 is almost necessary and sufficient, and is the weakest possible up to constant factors.

4.4 | Strong oracle property

In this section, we establish the strong oracle property for the multi-step estimator $\hat{\boldsymbol{\beta}}^{(\ell)}$ when ℓ is sufficiently large, that is, $\hat{\boldsymbol{\beta}}^{(\ell)}$ equals the oracle estimator $\hat{\boldsymbol{\beta}}^{\mathrm{ora}}$ with high probability (Fan & Lv, 2011). To this end, we define a similar local RSC event to $\mathcal{E}_{\mathrm{rsc}}(r, l, \kappa)$ given in Equation (18). Recall that $\mathcal{S} \subseteq [p]$ is the support of $\boldsymbol{\beta}^*$. Given radius parameters $r, l > 0$ and a curvature parameter $\kappa > 0$, define

$$\mathcal{G}_{\mathrm{rsc}}(r, l, \kappa) = \left\{ \frac{\langle \nabla \hat{Q}_h(\boldsymbol{\beta}_1) - \nabla \hat{Q}_h(\boldsymbol{\beta}_2), \boldsymbol{\beta}_1 - \boldsymbol{\beta}_2 \rangle}{\|\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2\|_{\boldsymbol{\Sigma}}^2} \geq \kappa \text{ for all } (\boldsymbol{\beta}_1, \boldsymbol{\beta}_2) \in \Lambda(r, l) \right\}, \tag{29}$$

where $\Lambda(r, l) := \{(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2) : \boldsymbol{\beta}_1 \in \boldsymbol{\beta}_2 + \mathbb{B}_{\boldsymbol{\Sigma}}(r) \cap \mathbb{C}_{\boldsymbol{\Sigma}}(l), \boldsymbol{\beta}_2 \in \boldsymbol{\beta}^* + \mathbb{B}_{\boldsymbol{\Sigma}}(r/2),\ \mathrm{supp}(\boldsymbol{\beta}_2) \subseteq \mathcal{S}\}$. Similarly to Equation (21), we define the oracle score

$$\boldsymbol{w}_h^{\mathrm{ora}} = \nabla \hat{Q}_h(\hat{\boldsymbol{\beta}}^{\mathrm{ora}}) \in \mathbb{R}^p, \tag{30}$$

where $\hat{\boldsymbol{\beta}}^{\mathrm{ora}}$ is defined in Equation (16). By the optimality of $\hat{\boldsymbol{\beta}}^{\mathrm{ora}}$, we have $\boldsymbol{w}_{h,\mathcal{S}}^{\mathrm{ora}} = (-1/n)\sum_{i=1}^n \ell_h'(y_i - \boldsymbol{x}_{i,\mathcal{S}}^{\mathrm{T}} \hat{\boldsymbol{\beta}}_{\mathcal{S}}^{\mathrm{ora}}) \boldsymbol{x}_{i,\mathcal{S}} = \mathbf{0}_s$. Like Theorem 2, the following result is also deterministic given the stated conditioning.

Theorem 4 *Assume Condition* (A1) *holds, and for some predetermined* $\delta \in (0, 1)$ *and* $\kappa > 0$, *there exist constants* $\alpha_1 > \alpha_0 > 0$ *such that*

$$q'(\alpha_0) > 0, \quad \frac{\alpha_0}{\sqrt{1 + \{q'(\alpha_0)/2\}^2}} > \frac{1}{\delta \kappa \gamma_p} \text{ and } q'(\alpha_1) = 0. \tag{31}$$

Moreover, let $r \geq \gamma_p^{1/2} \alpha_0 c_1 s^{1/2} \lambda$ *and* $l = \{2 + \frac{2}{q'(\alpha_0)}\}(c_1^2 + 1)^{1/2}(s/\gamma_p)^{1/2}$, *where* $c_1 > 0$ *is a constant determined by*

$$0.5q'(\alpha_0)(c_1^2 + 1)^{1/2} + 1 = \alpha_0 \kappa \gamma_p c_1. \tag{32}$$

Assume the beta-min condition $\|\boldsymbol{\beta}^*_{\mathcal{S}}\|_{\min} \geq (\alpha_0 + \alpha_1)\lambda$ *holds. Then, conditioned on the event*

$$\begin{aligned} &\left\{\|\boldsymbol{w}_h^{\mathrm{ora}}\|_\infty \leq 0.5q'(\alpha_0)\lambda\right\} \cap \left\{\|\hat{\boldsymbol{\beta}}^{\mathrm{ora}} - \boldsymbol{\beta}^*\|_{\boldsymbol{\Sigma}} \leq r/2\right\} \cap \mathcal{G}_{\mathrm{rsc}}(r, l, \kappa) \\ &\quad \cap \left\{ \|\hat{\boldsymbol{\beta}}^{\mathrm{ora}} - \boldsymbol{\beta}^*\|_\infty \leq \left[\alpha_0 - \frac{\sqrt{1 + \{q'(\alpha_0)/2\}^2}}{\delta \kappa \gamma_p}\right] \lambda \right\}, \end{aligned} \tag{33}$$

the strong oracle property holds: $\hat{\boldsymbol{\beta}}^{(\ell)} = \hat{\boldsymbol{\beta}}^{\mathrm{ora}}$ *provided* $\ell \geq \lceil \log(s^{1/2}/\delta)/\log(1/\delta) \rceil$.

Our next goal is to control the probability of the events in Equation (33). To this end, we need the following statistical properties of the oracle estimator $\hat{\boldsymbol{\beta}}^{\mathrm{ora}}$, including a deviation bound and a non-asymptotic Kiefer–Bahadur representation that are of independent interest. The latter requires a slightly stronger moment condition on the random feature.

(B1′) In addition to Condition (B1), assume $\sup_{u\in\mathbb{R}} |f_{\varepsilon|\boldsymbol{x}}(u)| \le \bar{f}_u < \infty$ almost surely over $\boldsymbol{x}$.
(B2′) In addition to Condition (B2), assume $\sup_{u\in\mathbb{R}} K(u) \le \kappa_u$ for some $\kappa_u \in (0, 1]$.
(B3′) The (random) covariate vector $\boldsymbol{x} = \boldsymbol{\Sigma}^{1/2}\boldsymbol{z} \in \mathbb{R}^p$ is sub-Gaussian: there exists some $\upsilon_1 \ge 1$ such that $\mathbb{P}(|\boldsymbol{z}^{\mathrm{T}}\boldsymbol{u}| \ge \upsilon_1 \|\boldsymbol{u}\|_2 \cdot t) \le 2e^{-t^2/2}$ for all $\boldsymbol{u} \in \mathbb{R}^p$ and $t \ge 0$.

Note that the oracle $\hat{\boldsymbol{\beta}}^{\mathrm{ora}} \in \mathbb{R}^p$ with $\hat{\boldsymbol{\beta}}^{\mathrm{ora}}_{\mathcal{S}^{\mathrm{c}}} = \mathbf{0}$ is essentially an unpenalized smoothed QR estimator in the low-dimensional regime '$s \ll n$'. We refer to Fernandes et al. (2021) for a comprehensive asymptotic analysis when s is fixed, and He et al. (2021) for a finite sample theory when s is allowed to grow with n. This paper concerns the case where both s (intrinsic dimension) and p (ambient dimension) can grow with sample size n. We therefore summarize the estimation bound and Bahadur representation for $\hat{\boldsymbol{\beta}}^{\mathrm{ora}}_{\mathcal{S}}$ by He et al. (2021) in the following proposition. Let

$$\mathbf{S} = \mathbb{E}(\boldsymbol{x}_{\mathcal{S}}\boldsymbol{x}_{\mathcal{S}}^{\mathrm{T}}) \text{ and } \mathbf{D} = \mathbb{E}\{f_{\varepsilon|\boldsymbol{x}}(0) \cdot \boldsymbol{x}_{\mathcal{S}}\boldsymbol{x}_{\mathcal{S}}^{\mathrm{T}}\} \tag{34}$$

be, respectively, the $s \times s$ sub-matrices of $\boldsymbol{\Sigma}$ and $\mathbf{J}$ indexed by the true support $\mathcal{S} \subseteq [p]$.

Proposition 3 *Assume Conditions* (B1′)–(B3′) *hold. For any* $t \ge 0$, *suppose the sample size* n *and the bandwidth* $h = h_n$ *are such that* $n \gtrsim s + t$ *and* $\sqrt{(s+t)/n} \lesssim h \lesssim 1$. *Then, the oracle estimator* $\hat{\boldsymbol{\beta}}^{\mathrm{ora}}$ *defined in Equation* (16) *satisfies*

$$\|\hat{\boldsymbol{\beta}}^{\mathrm{ora}} - \boldsymbol{\beta}^*\|_{\boldsymbol{\Sigma}} = \|(\hat{\boldsymbol{\beta}}^{\mathrm{ora}} - \boldsymbol{\beta}^*)_{\mathcal{S}}\|_{\mathbf{S}} \lesssim f_l^{-1}\left(\sqrt{\frac{s+t}{n}} + h^2\right) \tag{35}$$

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

$$\left\|\mathbf{D}(\hat{\boldsymbol{\beta}}^{\mathrm{ora}} - \boldsymbol{\beta}^*)_{\mathcal{S}} + \frac{1}{n}\sum_{i=1}^{n}\left\{\overline{K}(-\varepsilon_i/h) - \tau\right\}\boldsymbol{x}_{i,\mathcal{S}}\right\|_{\mathbf{S}^{-1}} \lesssim \frac{s+t}{h^{1/2}n} + h\sqrt{\frac{s+t}{n}} + h^3 \tag{36}$$

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

Finally, with the above preparations, we are able to establish the strong oracle property of $\hat{\boldsymbol{\beta}}^{(\ell)}$ when ℓ is sufficiently large.

Theorem 5 *Assume Conditions* (B1′)–(B3′) *and* (A1) *hold with* $\kappa_l = \min_{|u|\le 1} K(u) > 0$ *and*

$$\max_{j\in\mathcal{S}^{\mathrm{c}}} \|\mathbf{J}_{j\mathcal{S}}(\mathbf{J}_{\mathcal{S}\mathcal{S}})^{-1}\|_1 \le A_0. \tag{37}$$

for some $A_0 \ge 1$. *For a prespecified* $\delta \in (0, 1)$, *suppose there exist constants* $\alpha_1 > \alpha_0$ *satisfying Equation* (31) *with* $\kappa = \kappa_l f_l/2$, *and the beta-min condition* $\|\boldsymbol{\beta}^*_{\mathcal{S}}\|_{\min} \ge (\alpha_0 + \alpha_1)\lambda$. *Choose the bandwidth* h *and penalty level* λ *as* $h \asymp \{\log(p)/n\}^{1/4}$ *and* $\lambda \asymp \sqrt{\log(p)/n}$. *Then, with probability at least* $1 - 2p^{-1} - 5n^{-1}$, $\hat{\boldsymbol{\beta}}^{(\ell)} = \hat{\boldsymbol{\beta}}^{\mathrm{ora}}$ *for all* $\ell \ge \lceil \log(s^{1/2}/\delta)/\log(1/\delta) \rceil$, *provided that the sparsity s and ambient dimension p obey the growth condition* $\max\{s^2 \log(p), s^{8/3}/(\log p)\} \lesssim n$.

As stated in Theorem 5, in addition to the beta-min condition $\|\boldsymbol{\beta}^*_{\mathcal{S}}\|_{\min} \gtrsim \sqrt{\log(p)/n}$, we need an extra assumption (37) to establish the strong oracle property. Informally speaking, if we regress every spurious (density-weighted) feature $f_{\varepsilon|\boldsymbol{x}}(0) \cdot x_j$ ($j \in \mathcal{S}^c$) on the important (density-weighted) features $f_{\varepsilon|\boldsymbol{x}}(0) \cdot x_{\mathcal{S}}$, Equation (37) requires the ℓ_1-norm of the resulting regression coefficient vector to be bounded by A_0. It is worth noting that assumption (37) is much weaker than the irrepresentable condition, which is sufficient and nearly necessary for model consistency of the Lasso (Lahiri, 2021; Meinshausen & Bühlmann, 2006; Zhao & Yu, 2006) in the conditional mean model. A population version of the irrepresentable condition is that, for some $\alpha \in (0,1)$, $\max_{j\in\mathcal{S}^c} \|\boldsymbol{\Sigma}_{j\mathcal{S}}(\boldsymbol{\Sigma}_{\mathcal{S}\mathcal{S}})^{-1}\|_1 \leq \alpha$.

For conditional mean regression with heavy-tailed errors, Loh (2017) established the strong oracle property for any local stationary point of the folded concave penalized optimization problem (2) subject to an ℓ_1-ball constraint, when the loss function is twice differentiable. The required growth condition on (s, p) is $\max\{s \log(p), s^2\} \lesssim n$; see Theorem 2 in Loh (2017). For sparse QR, our result requires a slightly stronger scaling $\max\{s^2 \log(p), s^{8/3}/(\log p)\} \lesssim n$ due to the non-smoothness of the quantile loss. Intuitively, the strong oracle property is related to the second-order accuracy and efficiency: the oracle estimator is asymptotically normal provided that the sparsity s does not grow too fast with the sample size. For Huber's M-estimator, He and Shao (2000) proved the asymptotic normality for its linear functionals under the scaling $s^2 \log(s) = o(n)$; while in the context of QR, the same asymptotic results usually hold under stronger growth conditions due to both non-linearity and non-smoothness of the problem, such as $s^3(\log n)^2 = o(n)$ (He & Shao, 2000; Welsh, 1989) and $s^{8/3} = o(n)$ (He et al., 2021). To some extent, this explains why the high-dimensional scaling in our Theorem 5 is slightly stronger than those needed for regularized M-estimators with smooth loss functions.

5 | NUMERICAL STUDY

We perform numerical studies to assess the performance of the proposed regularized QR method using ℓ_1 and SCAD penalties. The SCAD penalty (Fan & Li, 2001) is defined through its derivative that takes the form $q'_\lambda(t) = \lambda \mathbb{1}(t \leq \lambda) + (a-1)^{-1}(a\lambda - t)_+ \mathbb{1}(t > \lambda)$ for $t \geq 0$, where we pick $a = 3.7$ as suggested in Fan and Li (2001), although it may not be the optimal value for QR. We use uniform and Gaussian kernels to smooth the quantile loss, and then employ the multi-stage convex relaxation method described in Algorithm 1 with $\ell = 3$ iterations. We will show later in this section that for moderately large p, $\ell = 3$ iterations is often sufficient and that more iterations will lead to little to no improvement in terms of estimation accuracy.

We compare our proposal—iteratively reweighted ℓ_1-penalized smoothed QR, with the standard Lasso implemented by the `R` packageg `glmnet`, and both ℓ_1- and folded concave penalized QRs implemented by the `R` package `FHDQR` (Gu et al., 2018). As a benchmark, we also compute the oracle estimator by fitting unpenalized QR using the important covariates. The regularization parameter λ for Lasso and penalized QR is selected via fivefold cross-validation; for the latter, we use the check loss to define the validation error. Specifically, we choose the λ value that yields the minimum cross-validation error under the ℓ_2-loss and check loss for Lasso and penalized QR respectively. The proposed method involves a smoothing parameter h, which can also be tuned via cross-validation in practice. Recall that convolution smoothing facilitates optimization through a balanced trade-off between statistical accuracy and computational complexity. Our numerical experiments show that the results are rather insensitive to the choice of the bandwidth provide that it is in a reasonable range (neither too small nor too large). The default value of h is set to be $\max\{0.05, \sqrt{\tau(1-\tau)}\{\log(p)/n\}^{1/4}\}$. We note that this particular choice of h is by no means optimal numerically.

For all the numerical experiments, we generate synthetic data $\{(y_i, \boldsymbol{x}_i)\}_{i=1}^n$ from a linear model $y_i = \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}^* + \varepsilon_i$ with $\boldsymbol{\beta}^* = (1.8, 0, 1.6, 0, 1.4, 0, 1.2, 0, 1, 0, -1, 0, -1.2, 0, -1.4, 0, -1.6, 0, -1.8, \mathbf{0}_{p-19})^{\mathrm{T}}$, and $\boldsymbol{x}_i \sim N_p(\mathbf{0}, \boldsymbol{\Sigma})$ with $\boldsymbol{\Sigma} = (0.7^{|j-k|})_{1\le j,k\le p}$. The random error follows one of the following four distributions: (i) standard normal distribution $N(0,1)$; (ii) t-distribution with 1.5 degrees of freedom; (iii) standard Cauchy distribution; and (iv) a mixture of normal distributions—$0.7N(0, 1)+0.3N(0, 25)$.

To evaluate the performance across different methods, we report the true and false positive rates (TPR and FPR), defined as the proportion of correctly estimated nonzeros and the proportion of falsely estimated nonzeros respectively. We also report the sum of squared errors (SSE), that is, $\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2^2$. Results for four different noise distributions under moderate ($n = 500$, $p = 400$) and high-dimensional settings ($n = 500$, $p = 1000$), averaged over 100 replications, are displayed in Tables 1–4.

Under the Gaussian random noise, we see from Table 1 that all methods have similar TPR and FPR. The Lasso has the lowest SSE compared to QR-Lasso and SQR-Lasso, which coincides with the fact that QR does lose some efficiency in a normal model. For both standard and smoothed QRs, iteratively reweighted regularization with the SCAD penalty considerably reduces the estimation error, is proximate to the oracle procedure. Similar results hold when the MCP is used. This supports our theoretical results on SQR that concave regularization improves the estimation error from $\sqrt{s\log(p)/n}$ to the near-oracle rate $\sqrt{\{s+\log(p)\}/n}$. Among all regularized QR methods, the proposed procedure—iteratively reweighted ℓ_1-penalized SQR with either uniform or Gaussian kernel smoothing—has the best overall performance.

TABLE 1 Numerical comparisons under Gaussian model

	Moderate dimension ($n = 500, p = 400$)			**High dimension ($n = 500, p = 1000$)**		
Methods	**TPR**	**FPR**	**Error**	**TPR**	**FPR**	**Error**
Lasso	1 (0)	0.067 (0.003)	0.147 (0.006)	1 (0)	0.033 (0.001)	0.167 (0.006)
SCAD	1 (0)	0.055 (0.003)	0.062 (0.012)	1 (0)	0.026 (0.001)	0.051 (0.003)
QR-Lasso	1 (0)	0.119 (0.006)	0.240 (0.009)	1 (0)	0.068 (0.003)	0.284 (0.009)
QR-SCAD	1 (0)	0.112 (0.006)	0.183 (0.014)	1 (0)	0.069 (0.004)	0.161 (0.010)
SQR-Lasso (uniform)	1 (0)	0.066 (0.003)	0.224 (0.013)	1 (0)	0.036 (0.002)	0.234 (0.007)
SQR-SCAD (uniform)	1 (0)	0.057 (0.004)	0.129 (0.011)	1 (0)	0.032 (0.002)	0.116 (0.008)
SQR-Lasso (Gaussian)	1 (0)	0.072 (0.004)	0.191 (0.007)	1 (0)	0.034 (0.002)	0.223 (0.007)
SQR-SCAD (Gaussian)	1 (0)	0.056 (0.003)	0.131 (0.010)	1 (0)	0.028 (0.002)	0.108 (0.007)
Oracle	1 (0)	0 (0)	0.049 (0.003)	1 (0)	0 (0)	0.053 (0.003)

Notes: The empirical average (and standard error) of the true and false positive rates (TPR and FPR) as well as the sum of squared errors (SSE), over 100 simulations, are reported.

TABLE 2 Numerical comparisons under $t_{1.5}$ model

	Moderate dimension **($n = 500, p = 400$)**			**High dimension** **($n = 500, p = 1000$)**		
Methods	**TPR**	**FPR**	**Error**	**TPR**	**FPR**	**Error**
Lasso	0.908 (0.016)	0.052 (0.002)	4.615 (0.401)	0.854 (0.022)	0.023 (0.001)	5.668 (0.524)
SCAD	0.842 (0.020)	0.044 (0.002)	7.138 (0.739	0.790 (0.024)	0.019 (0.001)	8.253 (0.762)
QR-Lasso	1 (0)	0.112 (0.005)	0.417 (0.015)	1 (0)	0.065 (0.003)	0.541 (0.021)
QR-SCAD	1 (0)	0.103 (0.005)	0.346 (0.024)	1 (0)	0.062 (0.003)	0.362 (0.022)
SQR-Lasso (uniform)	0.999 (0.001)	0.067 (0.004)	0.387 (0.032)	1 (0)	0.032 (0.002)	0.433 (0.017)
SQR-SCAD (uniform)	0.999 (0.001)	0.055 (0.004)	0.266 (0.028)	1 (0)	0.028 (0.002)	0.230 (0.017)
SQR-Lasso (Gaussian)	1 (0)	0.066 (0.003)	0.332 (0.012)	1 (0)	0.030 (0.001)	0.420 (0.017)
SQR-SCAD (Gaussian)	1 (0)	0.048 (0.003)	0.238 (0.018)	1 (0)	0.024 (0.001)	0.220 (0.015)
Oracle	1 (0)	0 (0)	0.065 (0.004)	1 (0)	0 (0)	0.074 (0.004)

TABLE 3 Numerical comparisons under Cauchy model

	Moderate dimension **($n = 500, p = 400$)**			**High dimension** **($n = 500, p = 1000$)**		
Methods	**TPR**	**FPR**	**Error**	**TPR**	**FPR**	**Error**
Lasso	0.344 (0.032)	0.021 (0.003)	16.799 (0.522)	0.305 (0.033)	0.009 (0.001)	17.479 (0.953)
SCAD	0.297 (0.028)	0.020 (0.002)	20.382 (0.860)	0.272 (0.029)	0.009 (0.001)	19.526 (0.871)
QR-Lasso	1 (0)	0.118 (0.004)	0.546 (0.022)	1 (0)	0.060 (0.002)	0.709 (0.025)
QR-SCAD	1 (0)	0.112 (0.005)	0.585 (0.047)	1 (0)	0.058 (0.002)	0.473 (0.034)
SQR-Lasso (uniform)	0.990 (0.004)	0.054 (0.002)	0.628 (0.070)	0.999 (0.010)	0.030 (0.002)	0.588 (0.042)
SQR-SCAD (uniform)	0.992 (0.004)	0.045 (0.003)	0.391 (0.047)	0.998 (0.002)	0.026 (0.001)	0.308 (0.031)
SQR-Lasso (Gaussian)	1 (0)	0.058 (0.002)	0.434 (0.017)	1 (0)	0.028 (0.001)	0.533 (0.019)
SQR-SCAD (Gaussian)	1 (0)	0.042 (0.002)	0.298 (0.021)	1 (0)	0.022 (0.001)	0.276 (0.021)
Oracle	1 (0)	0 (0)	0.076 (0.004)	1 (0)	0 (0)	0.080 (0.004)

Next, we examine the performance of different methods when outliers are present. From Table 2 we see that the Lasso has the highest SSE with TPR merely above 0.5 in both moderate- and high-dimensional settings. In contrast, regularized QR methods have high TPR while maintain low FPR. The FPR and SSE for SQR are further reduced by a visible margin when the SCAD penalty is used. This corroborates our main message that high-dimensional QR significantly benefits from smoothing and non-convex regularization. Similar results can be found in Tables 3 and 4 for Cauchy and a mixture normal error distributions.

Lastly, we assess more closely the effects of iteratively reweighted ℓ_1-regularization; see Algorithm 1. We keep the above model settings and focus on three different noise distributions: (i) t distribution with 1.5 degrees of freedom; (ii) standard Cauchy distribution; and (iii) a mixture normal distribution. For simplicity, we set the tuning parameter $\lambda = 0.5\sqrt{\log(p)/n}$. We run Algorithm 1 with uniform kernel and stop after seven iterations. Starting with $\hat{\boldsymbol{\beta}}^{(0)} = \mathbf{0}$, recall that

TABLE 4 Numerical comparisons under mixture normal model

	Moderate dimension ($n = 500, p = 400$)			**High dimension ($n = 500, p = 1000$)**		
Methods	**TPR**	**FPR**	**Error**	**TPR**	**FPR**	**Error**
Lasso	0.999 (0.001)	0.062 (0.003)	1.253 (0.058)	1 (0)	0.030 (0.001)	1.346 (0.047)
SCAD	0.996 (0.002)	0.048 (0.002)	0.606 (0.063)	0.995 (0.002)	0.025 (0.001)	0.746 (0.070)
QR-Lasso	1 (0)	0.126 (0.005)	0.507 (0.019)	1 (0)	0.059 (0.002)	0.559 (0.017)
QR-SCAD	1 (0)	0.121 (0.006)	0.546 (0.041)	1 (0)	0.057 (0.002)	0.361 (0.020)
SQR-Lasso (uniform)	0.999 (0.001)	0.070 (0.004)	0.496 (0.040)	1 (0)	0.030 (0.002)	0.462 (0.013)
SQR-SCAD (uniform)	1 (0)	0.060 (0.004)	0.366 (0.029)	1 (0)	0.026 (0.002)	0.244 (0.016)
SQR-Lasso (Gaussian)	1 (0)	0.072 (0.003)	0.405 (0.015)	1 (0)	0.029 (0.001)	0.443 (0.013)
SQR-SCAD (Gaussian)	1 (0)	0.054 (0.003)	0.346 (0.024)	1 (0)	0.024 (0.001)	0.242 (0.015)
Oracle	1 (0)	0 (0)	0.087 (0.005)	1 (0)	0 (0)	0.086 (0.004)

$\hat{\boldsymbol{\beta}}^{(1)}$ is the SQR-Lasso estimator. To quantify the relative performance of the solution path, at ℓth iteration, we define the relative improvement of $\hat{\boldsymbol{\beta}}^{(\ell)}$ with respect to$\hat{\boldsymbol{\beta}}^{(\ell-1)}$ as

$$\frac{\|\hat{\boldsymbol{\beta}}^{(\ell-1)} - \boldsymbol{\beta}^*\|_2^2 - \|\hat{\boldsymbol{\beta}}^{(\ell)} - \boldsymbol{\beta}^*\|_2^2}{\|\hat{\boldsymbol{\beta}}^{(1)} - \boldsymbol{\beta}^*\|_2^2}, \quad \ell \geq 2. \tag{38}$$

The relative improvement is a value between zero and one. A value close to zero indicates that there is little improvement in estimation error and vice versa. The results for $n = 500$ and $p \in \{200, 400, 1000, 2000\}$, averaged over 100 replications, are summarized in Figure 2. We see that running an additional iteration ($\ell = 2$) leads to the most significant improvement. The estimator, after $\ell = 3$ iterations, can still be improved under the t and Cauchy models. In all the (n, p) settings considered, running $\ell \geq 4$ iterations only shows marginal improvement, suggesting that the multi-step procedure with $\ell = 3$ is sufficient for moderate-scale datasets.

6 | AN APPLICATION TO GENE EXPRESSION DATA

We apply the proposed method to an expression quantitative trait locus (eQTL) dataset previously analysed in Scheetz et al. (2006), Kim et al. (2008) and Wang et al. (2012). The dataset was collected on a study that used eQTL mapping in laboratory rats to investigate and identify genetic variation in the mammalian eye that is relevant to human eye disease (Scheetz et al., 2006) Following Wang et al. (2012), we study the association between gene TRIM32, which was found to be associated with human eye disease, and the other expressions at other probes. The data consist of expression values of 31,042 probe sets on 120 rats. After some data pre-processing steps as described in Wang et al. (2012), the number of probes are reduced to 18,958. We further select the top 500 probes that have the highest absolute correlation with the expression of the response. We apply the proposed method using the uniform kernel and SCAD penalty, with

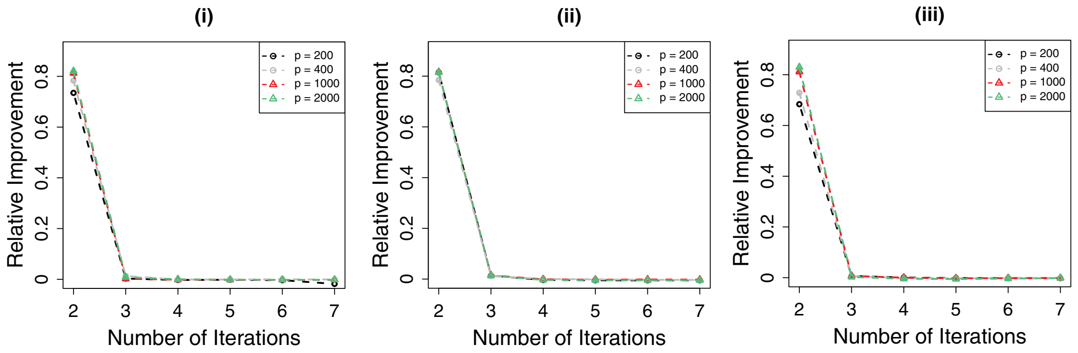

FIGURE 2 Plots of relative improvement defined in Equation (38) versus number of iterations when $n = 500$ and $p \in \{200, 400, 1000, 2000\}$. The three panels correspond to models with different noise distributions: (i) t distribution with 1.5 degrees of freedom; (ii) standard Cauchy distribution; and (iii) a mixture normal distribution [Colour figure can be viewed at wileyonlinelibrary.com]

TABLE 5 The average selected model size and prediction error (under quantile loss), with standard errors in the parenthesis, over 50 random partitions

Methods	Model size	Prediction error
QR-Lasso ($\tau = 0.3$)	38.28 (3.192)	0.225 (0.005)
QR-SCAD ($\tau = 0.3$)	34.66 (3.291)	0.241 (0.006)
SQR-Lasso ($\tau = 0.3$)	45.28 (1.866)	0.118 (0.003)
SQR-SCAD ($\tau = 0.3$)	31.32 (1.827)	0.106 (0.003)
QR-Lasso ($\tau = 0.5$)	33.76 (1.985)	0.222 (0.003)
QR-SCAD ($\tau = 0.5$)	30.28 (2.114)	0.236 (0.004)
SQR-Lasso ($\tau = 0.5$)	36.76 (1.533)	0.142 (0.003)
SQR-SCAD ($\tau = 0.5$)	29.58 (2.006)	0.132 (0.003)
QR-Lasso ($\tau = 0.7$)	29.66 (1.669)	0.195 (0.003)
QR-SCAD ($\tau = 0.7$)	24.22 (1.942)	0.205 (0.003)
SQR-Lasso ($\tau = 0.7$)	41.44 (2.262)	0.124 (0.003)
SQR-SCAD ($\tau = 0.7$)	27.52 (2.269)	0.116 (0.004)

regularization parameter selected by tenfold cross-validation. For comparisons, we also implement the ℓ_1- and concave regularized QR methods, denoted by QR-Lasso and QR-SCAD, using the R package FHDQR.

Similar to Wang et al. (2012), we conduct 50 random partitions of the data by randomly selecting the expression values for 80 rats as the training data and the remaining 40 rats as the testing data. The selected model size and prediction error (under quantile loss), averaged over 50 random partitions, are reported in Table 5. We observe from Table 5 that the SQR has consistently lower prediction errors than the standard QR across all three quantile levels considered. The prediction error is also improved for SQR when the SCAD penalty is used. In contrary, QR-SCAD exhibits no improvement over QR-Lasso in prediction accuracy, which is in line with the observation in Wang et al. (2012). One explanation may be that the lack of smoothness and strong convexity of the quantile loss overshadows the bias-reducing property of the concave penalty. These results suggest that high-dimensional QR considerably benefits from smoothing and concave regularization in terms of model selection ability, prediction accuracy and computational feasibility.

7 | DISCUSSIONS

In this paper we introduced a class of penalized convolution smoothed methods for fitting sparse QR models in high dimensions. Convolution smoothing turns the non-differentiable check loss into a twice-differentiable and convex surrogate, and the resulting empirical loss is proven to be locally strongly convex (with high probability). To reduce the ℓ_1-regularization bias as the signal strengthens, we considered a multi-step, iterative procedure which solves a weighted ℓ_1-penalized smoothed quantile objective function at each iteration. Statistically, we established the oracle-like performance of the output of this procedure, such as the oracle convergence rate and variable selection consistency, under an almost necessary and sufficient minimum signal strength condition. From a computational perspective, together convolution smoothing and convex relaxation enable the use of gradient-based algorithms that are much more scalable to large-scale datasets. In summary, through convolution smoothing with a suitably chosen bandwidth, we aim to seek a better trade-off between statistical accuracy and computational precision for high-dimensional QR. The proposed procedures will be implemented in the R package `conquer`, available at https://cran.r-project.org/web/packages/conquer/index.html.

The Python code is also publicly accessible at https://github.com/WenxinZhou/conquer, with an option to perform post-selection inference (via bootstrap).

There are several avenues for future work. When the parameter of interest arises in a matrix form, the low-rankness is often used to capture its low intrinsic dimension. This falls into the general category of ill-posed inverse problems, where the number of observations/measurements is much smaller than the ambient dimension of the model. See Chandrasekaran et al. (2012) for a general framework to convert notions of simplicity into convex penalty functions, resulting in convex optimization solutions to linear, underdetermined inverse problems. The idea of concave penalization can also be applied to low-rank matrix recovery problems. In essence, one can use a concave function to penalize the vector of singular values of matrix $\mathbf{\Theta} \in \mathbb{R}^{p_1 \times p_2}$. We refer to Wang et al. (2017) for a unified computational and statistical framework for non-convex low-rank matrix estimation when the Frobenius norm is used as the data-fitting measure. We conjecture that the proposed multi-step reweighted convex penalization approach and convolution smoothing will lead to oracle statistical guarantees and fast computational methods for quantile matrix regression and quantile matrix completion problems (Belloni et al., 2019). We leave this as future work.

ACKNOWLEDGEMENTS

We sincerely thank the associate editor and the two reviewers for their many constructive comments and valuable suggestions. Tan was supported by NSF DMS-1949730, NSF DMS-2113346 and NIH RF1-MH122833. Wang was supported by NSF FRGMS-1952373 and NSF OAC-2013755. Zhou acknowledges the support from NSF Grant DMS-2113409 and the UCSD Hellman Fellowship.

ORCID

Kean Ming Tan https://orcid.org/0000-0001-8491-275X
Lan Wang https://orcid.org/0000-0002-3217-0202
Wen-Xin Zhou https://orcid.org/0000-0002-2761-485X

REFERENCES

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

Bellec, P.C., Lecué, G. & Tsybakov, A.B. (2018) Slope meets Lasso: improved oracle bounds and optimality. *The Annals of Statistics*, 46, 3603–3642.

Belloni, A. & Chernozhukov, V. (2011) ℓ_1-penalized quantile regression in high-dimensional sparse models. *The Annals of Statistics*, 39, 82–130.

Belloni, A., Chen, M., Padilla, O.H.M. & Wang, Z. (2019) High dimensional latent pandel quantile regression with an application to asset pricing. *arXiv preprint arXiv:1912.02151*.

Boyd, S., Parikh, N., Chu, E., Peleato, B. & Eckstein, J. (2010) Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends in Machine Learning*, 3, 1–122.

Bradic, J., Fan, J. & Wang, W. (2011) Penalized composite quasi-likelihood for ultrahigh dimensional variable selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73, 325–349.

Bühlmann, P. & van de Geer, S. (2011) *Statistics for high-dimensional data: methods, theory and applications*. Heidelberg: Springer.

Chandrasekaran, V., Recht, B., Parrilo, P.A. & Willsky, A.S. (2012) The convex geometry of linear inverse problems. *Foundations of Computational Mathematics* 12, 805–849.

Fan, J. & Li, R. (2001) Variable selection via nonconcave regularized likelihood and its oracle properties. *Journal of the American statistical Association*, 96, 1348–1360.

Fan, J. & Lv, J. (2008) Sure independence screening for ultrahigh dimensional feature space. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70, 849–911.

Fan, J. & Lv, J. (2011) Nonconcave regularized likelihood with NP-dimensionality. *IEEE Transactions on Information Theory*, 57, 5467–5484.

Fan, J., Xue, L. & Zou, H. (2014) Strong oracle optimality of folded concave regularized estimation. *The Annals of Statistics*, 42, 819–849.

Fan, J., Liu, H., Sun, Q. & Zhang, T. (2018) I-LAMM for sparse learning: simultaneous control of algorithmic complexity and statistical error. *The Annals of Statistics*, 46, 814–841.

Fan, J., Li, R., Zhang, C.-H. & Zou, H. (2020) *Statistical foundations of data science*. Boca Raton: CRC Press.

Fernandes, M., Guerre, E. & Horta, E. (2021) Smoothing quantile regressions. *Journal of Business & Economic Statistics*, 39, 338–357.

Galvao, A.F. & Kato, K. (2016) Smoothed quantile regression for panel data. *Journal of Econometrics*, 193, 92–112.

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

Hastie, T., Tibshirani, R. & Wainwright, M. (2015) *Statistical learning with sparsity: the Lasso and generalizations*. Boca Raton: CRC Press.

He, X. & Shao, Q.-M. (2000) On parameters of increasing dimensions. *Journal of Multivariate Analysis*, 73, 120–135.

He, X., Pan, X., Tan, K.M.& Zhou, W.-X. (2021) Smoothed quantile regression with large-scale inference. *J. Econometrics*, to appear. Available from: https://doi.org/10.1016/j.jeconom.2021.07.010.

Horowitz, J. L. (1998) Bootstrap methods for median regression models. *Econometrica*, 66, 1327–1351.

Kim, Y. & Kwon, S. (2012) Global optimality of nonconvex regularized estimators. *Biometrika*, 99, 315–325.

Kim, Y., Choi, H. & Oh, H. S. (2008) Smoothly clipped absolute deviation on high dimensions. *Journal of the American Statistical Association*, 103, 1665–1673.

Koenker, R. (2005) *Quantile regression*. Cambridge: Cambridge University Press.

Koenker, R. (2015) Quantreg: quantile regression. *R Package Version 5.19*. Available from: https://cran.r-project.org/web/packages/quantreg/index.html.

Koenker, R. & Bassett, G. (1978) Regression quantiles. *Econometrica*, 46, 33-50.

Koenker, R., Chernozhukov, V., He, X. & Peng, L., eds. (2017) *Handbook of quantile regression*. Boca Raton, FL: CRC Press.

Lahiri, S.N. (2021) Necessary and sufficient conditions for variable selection consistency of the LASSO in high dimensions. *The Annals of Statistics*, 49, 820–844.

Leone, F.C., Nelson, L.S. & Nottingham, R.B. (1961) The folded normal distribution. *Technometrics*, 3, 543–550.

Loh, P.-L. (2017) Statistical consistency and asymptotic normality for high-dimensional robust M-estimators. *The Annals of Statistics*, 45, 866–896.

Loh, P.-L. & Wainwright, M.J. (2015) Regularized M-estimators with nonconvexity: statistical and algorithmic theory for local optima. *The Journal of Machine Learning Research*, 16, 559–616.

Meinshausen, N. & Bühlmann, P. (2006) High-dimensional graphs and variable selection with the Lasso. *The Annals of Statistics*, 34, 1436–1462.

Ndaoud, M. (2019) Interplay of minimax estimation and minimax support recovery under sparsity. *Proceedings of Machine Learning Research*, 98, 647–668.

Negahban, S. N., Ravikumar, P., Wainwright, M.J. & Yu, B. (2012) A unified framework for high-dimensional analysis of M-estimators with decomposable regularizers. *Statistical Science*, 27, 538–557.

Scheetz, T., Kim, K.-Y., Swiderski, R., Pilp, A., Braun, T., Knudtson, K. et al. (2006) Regulation of gene expression in the mammalian eye and its relevance to eye disease. *Proceedings of the National Academy of Sciences*, 103, 14429–14434.

Sivakumar, V. & Banerjee, A. (2017) High-dimensional structured quantile regression. *Proceedings of Machine Learning Research*, 70, 3220–3229.

Tibshirani, R. (1996) Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58, 267–288.

Wainwright, M.J. (2009) Sharp thresholds for high-dimensional and noisy recovery using ℓ_1-constrained quadratic programming (Lasso). *IEEE Transactions on Information Theory*, 55, 2183–2202.

Wainwright, M.J. (2019) *High-dimensional statistics: a non-asymptotic viewpoint*. Cambridge: Cambridge University Press.

Wang, L. (2013) The L_1 regularized LAD estimator for high dimensional linear regression. *Journal of Multivariate Analysis*, 120, 135–151.

Wang, L. & He, X. (2021) Analysis of global and local optima of regularized quantile regression in high dimension: a subgradient approach. *Preprint*.

Wang, H., Li, G. & Jiang, G. (2007) Robust regression shrinkage and consistent variable selection through the LAD-Lasso. *The Journal of Business & Economic Statistics*, 25, 347–355.

Wang, L., Wu, Y. & Li, R. (2012) Quantile regression for analyzing heterogeneity in ultra-high dimension. *Journal of the American Statistical Association*, 107, 214–222.

Wang, L., Zhang, X. & Gu, Q. (2017) A unified computational and statistical framework for nonconvex low-rank matrix estimation. *Proceedings of Machine Learning Research*, 54, 981–990.

Welsh, A.H. (1989) On M-processes and M-estimation. *The Annals of Statistics*, 15, 337–361.

Whang, Y.-J. (2006) Smoothed empirical likelihood methods for quantile regression models. *Economic Theory*, 22, 173–205.

Wu, Y., Ma, Y. & Yin, G. (2015) Smoothed and corrected score approach to censored quantile regression with measurement errors. *Journal of the American Statistical Association*, 110, 1670–1683.

Zhang, C.-H. (2010a) Nearly unbiased variable selection under minimax concave penalty. *The Annals of Statistics*, 38, 894–942.

Zhang, T. (2010b) Analysis of multi-stage convex relaxation for sparse regularization. *The Journal of Machine Learning Research*, 11, 1081–1107.

Zhang, C.-H. & Zhang, T. (2012) A general theory of concave regularization for high-dimensional sparse estimation problems. *Statistical Science*, 27, 576–593.

Zhao, P. & Yu, B. (2006) On model selection consistency of Lasso. *The Journal of Machine Learning Research*, 7, 2541–2563.

Zheng, Q., Peng, L. & He, X. (2015) Globally adaptive quantile regression with ultra-high dimensional data. *The Annals of Statistics*, 43, 2225–2258.

Zou, H. (2006) The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, 101, 1418–1429.

Zou, H. & Li, R. (2008) One-step sparse estimates in nonconcave regularized likelihood models. *The Annals of Statistics*, 36, 1509–1533.

How to cite this article: Tan, K.M., Wang, L. & Zhou, W.-X. (2022) High-dimensional quantile regression: Convolution smoothing and concave regularization. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 84(1), 205–233. Available from: https://doi.org/10.1111/rssb.12485