JOURNAL OF INDUSTRIAL AND
MANAGEMENT OPTIMIZATION
Volume **13**, Number **1**, January **2017**

doi:10.3934/jimo.2016006

pp. **93–112**

A SMOOTHING ITERATIVE METHOD FOR QUANTILE REGRESSION WITH NONCONVEX ℓ_p PENALTY

Lianjun Zhang

1: Department of Applied Mathematics
Beijing Jiaotong University
Beijing, 100044, China
2: No.1 High School of Huojia
Xinxiang, Henan, 453800, China

Lingchen Kong*

Department of Applied Mathematics
Beijing Jiaotong University
Beijing, 100044, China

Yan Li

School of Insurance and Economics
University of International Business and Economics
Beijing, 100029, China

Shenglong Zhou

School of Mathematics
University of Southampton
Southampton SO17 1BJ, United Kingdom

(Communicated by Adil Bagirov)

Abstract. The high-dimensional linear regression model has attracted much attention in areas like information technology, biology, chemometrics, economics, finance and other scientific fields. In this paper, we use smoothing techniques to deal with high-dimensional sparse models via quantile regression with the nonconvex ℓ_p penalty ($0 < p < 1$). We introduce two kinds of smoothing functions and give the estimation of approximation by our different smoothing functions. By smoothing the quantile function, we derive two types of lower bounds for any local solution of the smoothing quantile regression with the nonconvex ℓ_p penalty. Then with the help of ℓ_1 regularization, we propose a smoothing iterative method for the smoothing quantile regression with the weighted ℓ_1 penalty and establish its global convergence, whose efficient performance is illustrated by the numerical experiments.

2010 *Mathematics Subject Classification.* Primary: 58F15, 58F17; Secondary: 53C35.

Key words and phrases. High-dimensional linear sparse model, quantile regression, smoothing function, nonconvex ℓ_p penalty, weighted ℓ_1 penalty.

The first author is supported by National Natural Science Foundation of China (11431002,11171018).

* Corresponding author: Lingchen Kong.

1. **Introduction.** High-dimensional linear regression models have attracted much attention in areas like information technology, biology, chemometrics, genomic, economics, finance, functional magnetic resonance imaging and other scientific fields. The word “high-dimensional” refers to the situation where the number of unknown variables is larger than the number of samples in the underlying data. Obviously, it is almost impossible to tackle such kind of data without additional assumptions. One natural option is to utilize sparsity, which assumes that only a small number of unknown variables influence the response vector. Analysis of high-dimensional data poses many challenges for statisticians and calls for new methods and theories. For details, see, e.g., [2, 11] and references therein.

Regularization as one popular way to analyze high-dimensional sparse data have been used for a long time. Since sparse models are becoming increasingly important in statistics, machine learning and signal processing, there are a wealth of researchers working on sparse estimations via ℓ_1 or ℓ_p regularizations, see, e.g., [1, 5, 4, 7, 8, 22, 3, 9, 20, 21]. To be more specific, Aravkin, Kambadur, Lozano [1] used ℓ_1 regularization to solve the high-dimensional sparse linear regression model. Donoho and Elad [8] studied sparse representation in general dictionaries via ℓ_1 penalty. Chen, Xu, Ye [5] derived lower bounds for nonzero entries of local minimizers and also proposed a hybrid orthogonal matching pursuit-smoothing gradient method. Chen and Zhou [6], Lai and Wang [15] considered the different approximation of non-Lipschitz continuous $\|x\|_p^p$, and proposed an iterative reweighted ℓ_1 and ℓ_2 algorithm to solve this approximation problem, respectively. Lu [17] provided a unified convergence analysis for a general problem and developed new iterative reweighted methods.

Quantile regression performs well in the high dimensional sparse model, particularly in situations where noises are heavy-tailed or heterogeneous [12, 18], which was introduced by Koenker and Bassett [13]. Quantile regression becomes a popular and important tool in statistical analysis, which includes the well-known median regression or LAD as a special case. A comprehensive review can be found in Koenker [14]. Recently, regularized quantile regression has been widely studied. For example, Aravkin, Kambadur, Lozano [1] considered quantile regression with ℓ_0 and ℓ_1 penalties and exploited quantile Huber penalty (a smooth function) to substitute the quantile loss function and proposed a generalized orthogonal matching pursuit (OMP) method for variable selection. Fan, Fan and Barut [12] studied the penalized quantile regression with the weighted L_1-penalty (WR-Lasso) and designed a two-step procedure, called adaptive robust Lasso (AR-Lasso). They investigated the model selection oracle property and establish the asymptotic normality of the WR-Lasso. Wang et al. [18] considered the nonconvex penalized quantile regression in the ultra-high dimensional setting and showed that the oracle estimate belongs to the set of local minima of the nonconvex penalized quantile regression, under mild assumptions on the error distribution.

Motivated by the arguments to handle the high dimensional sparse model above, we consider quantile regression with nonconvex ℓ_p penalty $(0 < p < 1)$. We first introduce two sorts of smoothing functions, and give the estimation of approximation by our different smoothing functions. Then we utilize them to smoothing the quantile function $\rho_\tau(\cdot)$. Based on that, two lower bounds for nonconvex smoothing quantile regression are acquired. Moreover, since the nonconvex ℓ_p penalty smoothing model is still NP-hard, we take advantage of weighted ℓ_1 regularization replacing ℓ_p penalty. After that we propose a smoothing iterative method for smoothing quantile regression with weighted ℓ_1 penalty and establish its global convergence. Finally, we stimulate two examples to demonstrate the efficiency of the proposed approach.

The rest of this paper is organised as follows. Section 2 introduces some related preliminaries about the model and two types of smoothing functions. We then establish lower bounds theories for smoothing quantile regression with $\ell_p(0 < p < 1)$ penalty in Section 3. Algorithm based on smoothing quantile regression with weighted ℓ_1 penalty and its corresponding convergence is given in Section 4. We stimulate some numerical experiments in Section 5 and conclude the paper in the last section.

2. **Preliminaries.**

In this section, we give some notations about the mathematical model of the high-dimensional linear quantile regression with nonconvex ℓ_p penalty. By smoothing the quantile function and relaxing ℓ_p penalty into weighted ℓ_1 penalty, we can give corresponding different models for distinct theoretical and methodological purposes.

2.1. **Models.**

Consider the high-dimensional linear regression model

$$y = X\beta + \xi, \tag{1}$$

where $y \in \mathbb{R}^m$ is a response vector, $X = (x_1, x_2, \cdots, x_m)^\top \in \mathbb{R}^{m\times n}$ is a fixed design matrix, $\beta \in \mathbb{R}^n$ is a regression coefficient vector, $\xi \in \mathbb{R}^m$ is an error (noise) vector whose components are independently distributed. Here, high-dimensional means that $n > m$, namely the number of unknown variables is larger than the number of samples in the data. As we all known that without any assumption, $m \ll n$ means the problem is ill-posed, so we assume that β is sparse ($\|\beta\|_0 = k \ll n$) which means most components of β are zeroes.

To estimate sparse regression coefficient vector β, we propose the quantile regression with ℓ_p penalty defined as

$$\min \sum_{i=1}^{m} \rho_\tau(y_i - x_i^\top \beta) + \mu\|\beta\|_p^p, \tag{2}$$

where quantile $\tau \in (0, 1)$, quantile loss function is defined as $\rho_\tau(x) = (\tau - I(x < 0))x$ with $I(\cdot)$ being an indicator function, $\|\beta\|_p^p = \sum_{i=1}^{n} |\beta_i|^p$ with $0 < p < 1$, and penalty parameter $\mu > 0$. Model (2) is nonconvex and nonsmooth. When $\tau = \frac{1}{2}$, $\rho_\tau(x)$ is the so called least absolute deviation (LAD). Thus model (2) reduces to the LAD-ℓ_p model, that is

$$\min \|y - X\beta\|_1 + \mu\|\beta\|_p^p. \tag{3}$$

One way to study nonconvex and nonsmooth model is to adopt a smoothing function to substitute the objective function. In this paper, we use two smoothing functions to replace the quantile regression, one of which is a generalization of that in Chen and Zhou [6], another is the quantile Huber penalty as in [1]. We will interpret some properties of them in the subsequent subsection. Let $\rho_\tau(\cdot, \delta)$ be the smoothing quantile loss function, then model (2) can be smoothed as smoothing quantile regression with ℓ_p penalty

$$\min \sum_{i=1}^{m} \rho_\tau(y_i - x_i^\top \beta, \delta) + \mu\|\beta\|_p^p, \tag{4}$$

where δ is a smoothing parameter. Based on this model, we can easily drive two lower bounds similar to Chen, Xu, and Ye [5], see Section 3. However model (4) is still nonconvex and nonsmooth, and thus it is difficult to solve. Therefore we use weighted ℓ_1 penalty to approximate the ℓ_p penalty in (4), which is the following convex model

$$\min \sum_{i=1}^{m} \rho_\tau(y_i - x_i^\top \beta, \delta) + \mu \|w \circ \beta\|_1. \tag{5}$$

Based on this model, we will propose an effective method to pursue the sparse solution of high dimensional linear regression.

2.2. **Smoothing functions.** In this part, we first give a definition of smoothing functions. Then we will introduce two kinds of functions and prove that they are all smoothing functions of quantile loss function respectively. Smoothing technique has been studied and used in optimization and variational inequalities. Here we take the following definition. For more details, see e.g., [10, 16].

Definition 2.1. Let X, Y be two finite dimensional real Euclidean spaces. Let $f : X \to Y$ be a nondifferentiable function. A function $f_\delta : X \to Y$ with a parameter $\delta \in \mathbb{R}_+$ is called a smoothing function of f if it has the following properties:

(a) f_δ is continuously differentiable for any $\delta \in \mathbb{R}_{++}$;
(b) $\lim_{\delta\to 0} f_\delta(x) = f(x)$ for any $x \in X$, where $\delta \to 0$ for any $\delta \in \mathbb{R}_{++}$.

We say that f_δ is a uniformly smooth approach function of f if there is a scalar $K > 0$ such that

$$\| f_\delta(x) - f(x) \| \leq K \| \delta \|, \ \forall \delta \in \mathbb{R}_{++}, \ \forall x \in X. \tag{6}$$

We now introduce two kinds of smoothing functions and their properties. Then we will prove that they are smoothing functions of quantile loss function respectively. The first smoothing function is defined as:

$$\rho_{\tau,1}(\theta, \delta) = \begin{cases} (\tau - 1)\theta, \ \theta < -\delta, \\ \dfrac{\theta^2}{4\delta} + (\tau - \dfrac{1}{2})\theta + \dfrac{\delta}{4}, \ |\theta| \leq \delta, \\ \tau\theta, \ \theta > \delta. \end{cases} \tag{7}$$

Clearly, when $\tau = \frac{1}{2}$ the function $\rho_{1/2,1}(\theta, \delta)$ reduces to the smoothing function that has been extensively exploited, see [6],

$$\rho_{1/2,1}(\theta, \delta) = \begin{cases} |\theta|/2, \ |\theta| > \delta, \\ \dfrac{\theta^2}{4\delta} + \dfrac{\delta}{4}, \ |\theta| \leq \delta. \end{cases}$$

The graphs of quantile loss function and its smoothing function $\rho_{\tau,1}(\theta, \delta)$ is plotted in Figure 1. Every broken dotted line and the smooth active line above it are $\rho_\tau(\theta)$ and its smoothing function $\rho_{\tau,1}(\theta, \delta)$ respectively. We distinguish different functions with different colors when τ and δ are different.

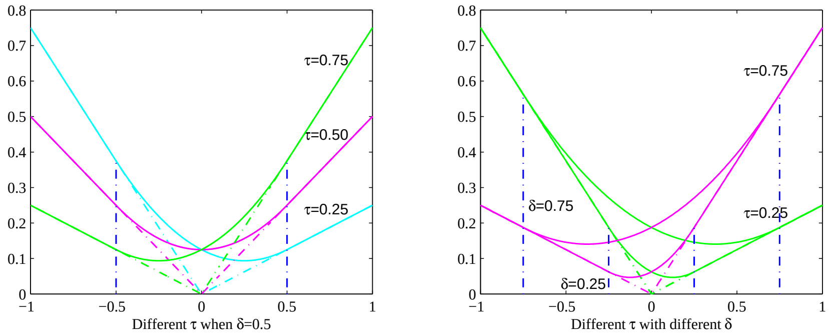


FIGURE 1. Function $\rho_\tau(\theta)$ and its smoothing relaxation $\rho_{\tau,1}(\theta, \delta)$

Theorem 2.2. *For any fixed $\delta > 0$ and $\tau \in (0, 1)$, the function $\rho_{\tau,1}(y_i - x_i^\top \beta, \delta)$ is continuously differentiable with respect to $\beta \in \mathbb{R}^n$. Moreover its first order partial derivative with respect to β is*

$$\nabla \rho_{\tau,1}(\theta_i, \delta) = \begin{cases} (1-\tau)x_i, \ \theta_i < -\delta, \\ -\left(\dfrac{\theta_i}{2\delta} + \left(\tau - \dfrac{1}{2}\right)\right) x_i, \ |\theta_i| \le \delta, \\ -\tau x_i, \ \theta_i > \delta, \end{cases} \tag{8}$$

and the second order partial derivative with respect to β is

$$\nabla^2 \rho_{\tau,1}(\theta_i, \delta) = \begin{cases} \dfrac{x_i x_i^\top}{2\delta}, \ |\theta_i| < \delta, \\ 0, \ |\theta_i| > \delta, \end{cases} \tag{9}$$

where $\theta_i = y_i - x_i^\top \beta$ for all i.

Theorem 2.3. *The function $\rho_{\tau,1}(\theta, \delta)$ is a uniformly smooth approach function of quantile loss function $\rho_\tau(\theta)$.*

Proof. From Proposition 2.2 we know that $\rho_{\tau,1}(\theta, \delta)$ is continuously differentiable for any $\delta \in \mathbb{R}_{++}$. Obviously we have $\lim_{\delta \to 0} \rho_{\tau,1}(\theta, \delta) = \rho_\tau(\theta)$ for any $\theta \in \mathbb{R}, \ \delta \in \mathbb{R}_{++}$. Suppose $h_{\tau,1}(\theta, \delta) = \rho_{\tau,1}(\theta, \delta) - \rho_\tau(\theta)$, which means

$$h_{\tau,1}(\theta, \delta) = \begin{cases} \dfrac{\theta^2}{4\delta} + \dfrac{\theta}{2} + \dfrac{\delta}{4}, \ -\delta < \theta < 0, \\ \dfrac{\theta^2}{4\delta} - \dfrac{\theta}{2} + \dfrac{\delta}{4}, \ 0 \le \theta < \delta, \\ 0, \ \text{otherwise}. \end{cases} \tag{10}$$

When $\theta \in (-\delta, 0)$, according to the Lagrange Mean Value Theorem, for any $\theta \in (-\delta, 0)$, there must be a $\xi \in (-\delta, \theta)$ that satisfies

$$\frac{h_{\tau,1}(\theta, \delta) - h_{\tau,1}(-\delta, \delta)}{\theta + \delta} = \frac{\xi}{2\delta} + \frac{1}{2},$$

which means

$$0 < h_{\tau,1}(\theta, \delta) = (\frac{\xi}{2\delta} + \frac{1}{2})(\theta + \delta) \le \frac{1}{2}\delta.$$

Thus, setting $K_1 = \frac{1}{2}$, we have

$$\| \rho_{\tau,1}(\theta, \delta) - \rho_\tau(\theta) \| \le K_1 \| \delta \|$$

for any $\theta \in (-\delta, 0)$.

We can also get

$$\| \rho_{\tau,1}(\theta,\delta) - \rho_\tau(\theta) \| \leq K_2 \| \delta \|$$

in the same way when $\theta \in [0, \delta)$. Therefore, takeing $K = \max\{K_1, K_2\}$, we obtain

$$\| \rho_{\tau,1}(\theta,\delta) - \rho_\tau(\theta) \| \leq K \| \delta \|$$

for any $\theta \in \mathbb{R}$. □

By applying the first smoothing function (7), we can achieve the lower bounds theories for smoothing quantile regression with $\ell_p (0 < p < 1)$ penalty, see the next part. However when it comes to computing issues, the quantile Huber penalty in [1] performs better than the first smoothing function (7), so we introduce it here

$$\rho_{\tau,2}(\theta,\delta) = \begin{cases} (\tau-1)\theta - \frac{\delta(1-\tau)^2}{2}, & \theta \in (-\infty,\ (\tau-1)\delta), \\ \dfrac{\theta^2}{2\delta}, & \theta \in [(\tau-1)\delta,\ \tau\delta], \\ \tau\theta - \frac{\delta\tau^2}{2}, & \theta \in (\tau\delta,\ +\infty) \end{cases} \tag{11}$$

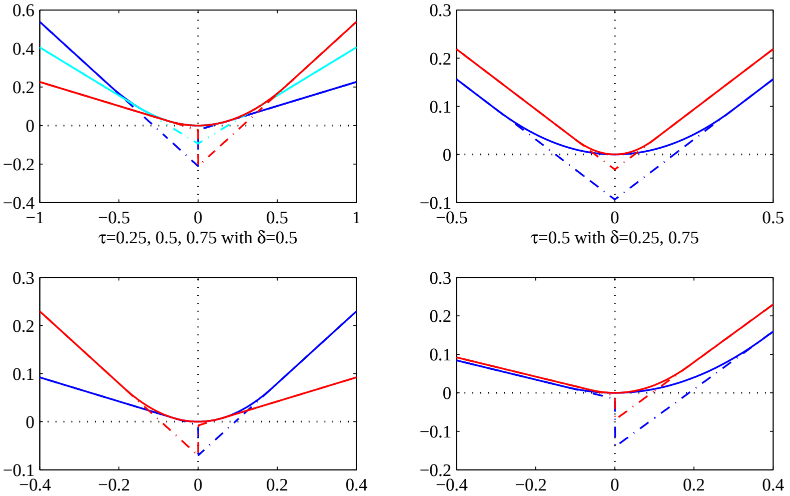
If we let $\tau = \frac{1}{2}$, the function $\rho_{1/2,2}(\theta,\delta)$ deduces to the following smoothing function:

$$\rho_{1/2,2}(\theta,\delta) = \begin{cases} \dfrac{|\theta|}{2} - \dfrac{\delta}{4}, & |\theta| > \delta, \\ \dfrac{\theta^2}{4\delta}, & |\theta| \leq \delta. \end{cases}$$

Actually, for any $\delta > 0$, function $\rho_{\tau,2}(\theta,\delta)$ no longer smooths the quantile function $\rho_\tau(\theta)$. Function $\rho_{\tau,2}(\theta,\delta)$ smoothes function $\varrho_\tau(\cdot,\delta)$ which is defined as

$$\varrho_\tau(\theta,\delta) = \begin{cases} (\tau-1)\theta - \frac{\delta(1-\tau)^2}{2}, & \theta \leq 0, \\ \tau\theta - \frac{\delta\tau^2}{2}, & \theta > 0. \end{cases} \tag{12}$$

Figure 2 presents us the patterns of $\rho_{\tau,2}(\theta,\delta)$ and $\varrho_\tau(\theta,\delta)$ under different $\tau \in (0,1)$ and $\delta > 0$. Every pair of dotted line in the same color and the smooth active line above it are $\rho_\tau(\theta,\delta)$ and its smoothing function $\rho_{\tau,2}(\theta,\delta)$ respectively. We distinguish different functions with different colors when τ and δ are different.



τ=0.25, 0.5, 0.75 with δ=0.5

τ=0.5 with δ=0.25, 0.75

τ=0.25, 0.75 with δ=0.25

τ=0.75 with δ=0.25, 0.75

FIGURE 2. Function $\varrho_\tau(\theta, \delta)$ and its smoothing relaxation $\rho_{\tau,2}(\theta, \delta)$

Moreover for $\delta = 0$, function $\varrho_\tau(\theta, \delta) = \rho_\tau(\theta)$, and the following property holds.

Theorem 2.4. *For any fixed $\delta > 0$ and $\tau \in (0, 1)$, we have*

$$\text{argmin}_{\beta\in\mathbb{R}^n} \sum_{i=1}^{m} \rho_\tau(y_i - x_i^\top \beta) = \text{argmin}_{\beta\in\mathbb{R}^n} \sum_{i=1}^{m} \rho_{\tau,2}(y_i - x_i^\top \beta, \delta). \tag{13}$$

This signifies that quantile Huber penalty has the exactly same solution with the original quantile loss function. Therefore, in order to obtain the optimal solution of the left side of the equality above, we could take advantage of $\rho_{\tau,2}(\cdot, \delta)$ to substitute the quantile loss function $\rho_\tau(\cdot)$.

Even though there are some differences between the two smoothing functions (7) and (11), the quantile Huber penalty can also drive the same lower bounds because they all have the first and second order partial derivatives which will be used in Section 3.

Theorem 2.5. *For any fixed $\delta > 0$ and $\tau \in (0, 1)$, the function $\rho_{\tau,2}(y_i - x_i^\top \beta, \delta)$ is continuously differentiable with respect to $\beta \in \mathbb{R}^n$. Moreover its first order partial derivative with respect to β is*

$$\nabla \rho_{\tau,2}(y_i - x_i^\top \beta, \delta) = \begin{cases} (1-\tau)x_i, & \theta_i \in (-\infty,\ (\tau-1)\delta), \\ -\frac{\theta_i}{\delta} x_i, & \theta_i \in [(\tau-1)\delta,\ \tau\delta], \\ -\tau x_i, & \theta_i \in (\tau\delta,\ +\infty), \end{cases} \tag{14}$$

and the second order partial derivative with respect to β is

$$\nabla^2 \rho_{\tau,2}(y_i - x_i^\top \beta, \delta) = \begin{cases} \frac{x_i x_i^\top}{\delta}, & \theta_i \in [(\tau-1)\delta,\ \tau\delta], \\ 0, & \textit{otherwise}, \end{cases} \tag{15}$$

where $\theta_i = y_i - x_i^\top \beta$ for all i.

Now we prove that $\rho_{\tau,2}(\theta, \delta)$ is a uniformly smooth approach function of $\rho_\tau(\theta)$.

Theorem 2.6. *The function $\rho_{\tau,2}(\theta, \delta)$ is a uniformly smooth approach function of quantile loss function $\rho_\tau(\theta)$.*

Proof. From Proposition 2.5 we know that $\rho_{\tau,2}(\theta, \delta)$ is continuously differentiable for any $\delta \in \mathbb{R}_{++}$. Obviously we have $\lim_{\delta\to 0} \rho_{\tau,2}(\theta, \delta) = \rho_\tau(\theta)$ for any $\theta \in \mathbb{R},\ \delta \in \mathbb{R}_{++}$. Suppose $h_{\tau,2}(\theta, \delta) = \rho_{\tau,2}(\theta, \delta) - \rho_\tau(\theta)$, which means

$$h_{\tau,2}(\theta, \delta) = \begin{cases} -\frac{\delta(1-\tau)^2}{2}, & -\infty < \theta < (\tau-1)\delta, \\ \frac{\theta^2}{2\delta} - (\tau-1)\theta, & (\tau-1)\delta \le \theta < 0, \\ \frac{\theta^2}{2\delta} - (\tau)\theta, & 0 \le \theta < \tau\delta, \\ -\frac{\delta\tau^2}{2}, & \tau\delta \le \theta < +\infty. \end{cases} \tag{16}$$

When $\theta \in (-\infty, (\tau-1)\delta)$,

$$\mid h_{\tau,2}(\theta, \delta) \mid = \mid \rho_{\tau,2}(\theta, \delta) - \rho_\tau(\theta) \mid = \mid \frac{\delta(1-\tau)^2}{2} \mid \le \frac{\delta}{2}.$$

This means taking $K_1 = \frac{1}{2}$, for any $\theta \in (-\infty, (\tau-1)\delta)$ we all have

$$\mid \rho_{\tau,2}(\theta, \delta) - \rho_\tau(\theta) \mid \le K_1\delta.$$

When $\theta \in [(\tau-1)\delta, 0)$, according to the Lagrange Mean Value Theorem, for any $\theta \in [(\tau-1)\delta, 0)$, there must be a $\xi \in (\theta, 0)$ that satisfies

$$| \; h_{\tau,2}(\theta,\delta) - h_{\tau,2}(0,\delta) \; |=| \; (\frac{\xi}{\delta} - (\tau-1))\theta \; | \leq (1-\tau)^2\delta,$$

which means taking $K_2 = (1-\tau)^2$, for any $\theta \in [(\tau-1)\delta, 0)$ we all have

$$| \; \rho_{\tau,2}(\theta,\delta) - \rho_\tau(\theta) \; | \leq K_2\delta.$$

We can also get

$$\| \; \rho_{\tau,1}(\theta,\delta) - \rho_\tau(\theta) \; \| \leq K_3 \; \| \; \delta \; \|$$

when $\theta \in [0, \tau\delta)$ and

$$\| \; \rho_{\tau,1}(\theta,\delta) - \rho_\tau(\theta) \; \| \leq K_4 \; \| \; \delta \; \|$$

when $\theta \in [\tau\delta, +\infty)$ in the same way.

Therefore, taking $K = \max\{K_1, K_2, K_3, K_4\}$, we obtain

$$\| \; \rho_{\tau,1}(\theta,\delta) - \rho_\tau(\theta) \; \| \leq K \; \| \; \delta \; \|$$

for any $\theta \in \mathbb{R}$. □

3. Lower bounds of smoothing quantile regression with ℓ_p penalty.

From the previous analysis, we drive two types of lower bounds results in this section. For notational simplicity, we use $\rho_\tau(\cdot,\delta)$ to represent both $\rho_{\tau,1}(\cdot,\delta)$ and $\rho_{\tau,1}(\cdot,\delta)$ and $S_\tau(\cdot,\delta) = \sum_{i=1}^m \rho_\tau(\cdot,\delta)$, then we can denote the model (4) as

$$\min \; S_\tau(\beta,\delta) + \mu\|\beta\|_p^p =: \Phi_\tau(\beta). \tag{17}$$

Before giving the lower bounds for (17), we represent a derivative property about smoothing functions, which will be given next.

Theorem 3.1. *The first and second order partial derivatives of $\rho_\tau(y_i - x_i^\top\beta, \delta)$ are bounded if they exist, namely $\|\nabla\rho_\tau(y_i - x_i^\top\beta,\delta)\| \leq \max\{\tau, 1-\tau\}\|x_i\| \leq \|x_i\|$ and $\|\nabla^2\rho_\tau(y_i - x_i^\top\beta,\delta)\| \leq \|x_i x_i^\top\|/(2\delta)$.*

We let $|\beta| = (|\beta_1|, |\beta_2|, \cdots, |\beta_n|)^\top$ and $\mathrm{sign}(\beta)$ stands for the signum function of β. A vector $x_\Omega \in \mathbb{R}^n$ denotes the vector equals to x on an index set Ω and zero elsewhere. For a matrix $X \in \mathbb{R}^{m\times n}$, we write $X_{i\cdot}(X_{\cdot j})$ as the i-th row (j-th column) of X, and $X_\Omega \in \mathbb{R}^{m\times k}$ as the sub-matrix of X with $(X_\Omega)_{ij} = (X)_{ij}, j \in \Omega$ and $(X_\Omega)_{ij} = 0$, otherwise, where $k = |\Omega| \leq \min\{m,n\}$. $\lambda_{\max}(X)$ is the maximum eigenvalue of X. The nonzero indices of the true regression vector $\widehat{\beta} \in \mathbb{R}^n$ is denoted as Ω with $|\Omega| = k$, namely $\Omega = supp(\widehat{\beta})$. For any $z \in \mathbb{R}^k$, define

$$\tilde{\rho}_\tau(z) = \begin{pmatrix} \nabla\rho_\tau(y_1 - (X_\Omega)_{1\cdot}^\top z, \delta) \\ \nabla\rho_\tau(y_2 - (X_\Omega)_{2\cdot}^\top z, \delta) \\ \vdots \\ \nabla\rho_\tau(y_m - (X_\Omega)_{m\cdot}^\top z, \delta) \end{pmatrix} \in \mathbb{R}^m,$$

$$A_{\rho_\tau} = \begin{pmatrix} \nabla^2\rho_\tau(y_1 - (X_\Omega)_{1\cdot}^\top z, \delta)a_1^\top \\ \nabla^2\rho_\tau(y_2 - (X_\Omega)_{2\cdot}^\top z, \delta)a_2^\top \\ \vdots \\ \nabla^2\rho_\tau(y_m - (X_\Omega)_{m\cdot}^\top z, \delta)a_m^\top \end{pmatrix} \in \mathbb{R}^{m\times k}, z \in \mathbb{R}^k,$$

where $a_i = (X_\Omega)_{i\cdot}, i = 1, 2, \cdots, m$ and $c_j = (X_\Omega)_{\cdot j}, \; j = 1, 2, \cdots, k$.

Inspired by those in [5], we now drive two theorems about the lower bounds using the first and second optimal conditions. Since $\Phi_\tau(\beta) = S_\tau(\beta,\delta) + \mu\|\beta\|_p^p \geq \mu\|\beta\|_p^p$, the the objective function is bounded below and $\Phi_\tau(\beta) \to \infty$ if $\mu\|\beta\|_p^p \to \infty$. So the local solutions set of (17) defined as $\mathbb{B}$ is nonempty and bounded. Let b be some positive constant such that for any $z \in \mathbb{R}^k$,

$$\|(X_\Omega)^\top \cdot \tilde{\rho}_\tau(z)\| \le b. \tag{18}$$

Remark 1. In practice, we can choose $b = m\|X_\Omega\| \cdot \max_{i=1,2,\cdots,m} \|(X_\Omega)_{i\cdot}\|$, because

$$\|(X_\Omega)^\top \cdot \tilde{\rho}_\tau(z)\| \le \|(X_\Omega)^\top\|\|\tilde{\rho}_\tau(z)\| \le m\|X_\Omega\| \max_{i=1,2,\cdots,m} \|(X_\Omega)_{i\cdot}\|. \tag{19}$$

Theorem 3.2. *(The first order bound) Let $L = \left(\frac{\mu p}{b}\right)^{\frac{1}{1-p}}$. Then for any $\hat{\beta} \in \mathbb{B}$, we have*

$$\hat{\beta}_i \in (-L, L) \;\Rightarrow\; \hat{\beta}_i = 0,\; i = 1, 2, ..., n. \tag{20}$$

Proof. For any $\hat{\beta} \in \mathbb{B}$ with $\|\hat{\beta}\|_0 = k$. Without loss of generality, we assume that $\hat{\beta} = (\hat{\beta}_1, \hat{\beta}_2, ...\hat{\beta}_k, 0, ..., 0)^\top$. Let $\hat{z} = (\hat{\beta}_1, \hat{\beta}_2, ...\hat{\beta}_k)^\top$. Define a function $\Psi_\tau(z)$ by

$$\Psi_\tau(z) := S_\tau(y_i - a_i^\top z, \delta) + \mu\|z\|_p^p. \tag{21}$$

Then we have

$$\Phi_\tau(\hat{\beta}) = S_\tau(y_i - x_i^\top \hat{\beta}, \delta) + \mu\|\hat{\beta}\|_p^p = S_\tau(y_i - a_i^\top \hat{z}, \delta) + \mu\|\hat{z}\|_p^p = \Psi_\tau(\hat{z}). \tag{22}$$

Since $\mid \hat{z}_i \mid > 0, i = 1, 2, ..., k$, $\Psi_\tau(z)$ is continuously differentiable at $\hat{z}$. Moreover

$$\Psi_\tau(\hat{z}) = \Phi_\tau(\hat{\beta}) \le \min\{\Phi_\tau(\beta) \mid \beta_i, i = k+1, ..., n\} = \min\{\Psi_\tau(z) \mid z \in \mathbb{R}^k\} \tag{23}$$

holds in a neighborhood of $\hat{\beta}$, so we find that $\hat{z}$ must be a local minimizer of function $\Psi_\tau(z)$. Hence the first order necessary condition for $\Psi_\tau(z)$ holds at $\hat{z}$. This gives

$$(X_\Omega)^\top \tilde{\rho}_\tau(\hat{z}) + \mu p(\mid \hat{z} \mid^{p-1} \operatorname{sign}(\hat{z})) = 0. \tag{24}$$

Therefore, we obtain

$$\mu p\|\hat{z}\|_{p-1}^{p-1} = \|(X_\Omega)^\top \rho_\tau(\hat{z})\| \le b, \tag{25}$$

which implies

$$b \ge \mu p\|\hat{z}\|_{p-1}^{p-1} > \mu p \min\{\mid \hat{z}_i \mid^{p-1}\},\; i = 1, ..., k. \tag{26}$$

Noting that $0 < p < 1$, we get

$$\min\;\{\mid \hat{z}_i \mid, i = 1, ..., k\} \ge \left(\frac{\mu p}{b}\right)^{\frac{1}{1-p}} = L. \tag{27}$$

Since $\hat{\beta} \in \mathbb{B}$ is chosen arbitrarily, we can claim that for any $\hat{\beta} \in \mathbb{B}$, its nonzero components are no less than L. In other words, if $\hat{\beta}_i \in (-L, L)$, then $\hat{\beta}_i = 0$ for $i = 1, 2, ..., n$. □

We next develop another lower bound by using the second order necessary optimality condition.

Theorem 3.3. *(The second order bound) Let $L_i = \left(\dfrac{2\delta\mu p(1-p)}{\| c_i \|_2^2}\right)^{\frac{1}{2-p}}$, $i = 1, 2, \cdots, n$. We have*

$$\hat{\beta}_i \in (-L_i, L_i) \;\Rightarrow\; \hat{\beta}_i = 0,\; i = 1, 2, ..., n. \tag{28}$$

Proof. Similar to the proof of Theorem 3.2, (23) is still holds at $\hat{z}$, which leads to the matrix

$$(X_\Omega)^\top A_{\rho_\tau} + \mu p(p-1)\text{diag}(\mid \hat{z} \mid^{p-2}) \tag{29}$$

to be positive semi-definite when the second order necessary optimality condition holds. That is

$$e_i^\top (X_\Omega)^\top A_{\rho_\tau} e_i + \mu p(p-1) \mid \hat{z}_i \mid^{p-2} \geq 0, \tag{30}$$

for any $i = 1, 2, \cdots, k$, so

$$\frac{1}{2\delta}\|c_i\|_2^2 \geq \mu p(1-p) \mid \hat{z}_i \mid^{p-2}, \tag{31}$$

which manifests

$$\mid \hat{z}_i \mid \geq \left(\frac{2\delta\mu p(1-p)}{\|c_i\|_2^2}\right)^{\frac{1}{2-p}} = L_i, \ i = 1, 2, \cdots, k. \tag{32}$$

Since $\hat{\beta} \in \mathbb{B}$ is chosen arbitrarily, we can claim that for any $\hat{\beta} \in \mathbb{B}$, its nonzero components are no less than L. In other words, if $\hat{\beta}_i \in (-L_i, L_i)$, then $\hat{\beta}_i = 0$ for $i = 1, 2, \cdots, n$. □

From the above results, we can see that model (17) wouldn't get very small nonzero components in any local solution. This can been seen a kind of variable selecting procedure.

4. Algorithm for smoothing quantile regression.

Even though there are lower bounds for smoothing quantile regression with ℓ_p penalty, it is still NP-hard. Inspired by the approximation

$$\|x^k\|_p^p \approx \sum_{i=1}^n \frac{1}{\left(|x_i^k| + \epsilon\right)^{1-p}}|x_i| =: \sum_{i=1}^n \widetilde{w}_i(x^k)|x_i|$$

where x^k is the result generated in the k-th iteration and sufficient small smoothing factor $\epsilon > 0$, we take advantage of a weighted ℓ_1 minimization instead of ℓ_p regularization directly. In this section we mainly analyze the convex model

$$\min \ S_\tau(\beta, \delta) + \mu\|w \circ \beta\|_1, \tag{33}$$

where $S_\tau(\cdot, \delta) = \sum_{i=1}^m \rho_{\tau,2}(\cdot, \delta)$. We will see that quantile Huber penalty perform better than $\rho_{\tau,1}(\cdot, \delta)$ in numerical experiments, so we let $S_\tau(\beta, \delta)$ represents the sum of quantile Huber penalty functions.

With better comprehending the structure of the algorithm, we define several notations.

$$\begin{aligned} \Omega(\beta, \delta) &:= \left\{i \in \{1, 2, \cdots, n\} \;\middle|\; (\tau-1)\delta \leq y_i - x_i^\top \beta \leq \tau\delta\right\}, & (34)\\ \mathcal{C}_n^s &:= \left\{S \subset \{1, 2, \cdots, n\} \;\middle|\; |S| = s\right\}, & (35)\\ M_{\tau,\epsilon}(\beta, \alpha, \delta) &:= S_\tau(\alpha, \delta) + \langle \nabla_\beta S_\tau(\alpha, \delta), \beta - \alpha\rangle + \frac{L_\alpha + \epsilon}{2}\|\beta - \alpha\|_2^2, & (36) \end{aligned}$$

where $\epsilon > 0$ is sufficiently small, $\nabla_\beta S_\tau(\alpha,\delta) = \sum_{i=1}^m \nabla\rho_\tau\left(y_i - x_i^\top\alpha, \delta\right)$ (if it is not ambiguous, we shortly write $\nabla S_\tau(\alpha,\delta) = \nabla_\beta S_\tau(\alpha,\delta)$) and

$$\begin{aligned} L_\alpha \geq & \lambda_{\max}\left(\nabla^2 S_\tau(\beta^k,\delta)\right) \\ = & \frac{1}{\delta}\lambda_{\max}\left(\sum_{i\in\Omega(\alpha,\delta)} x_i x_i^\top\right) = \frac{1}{\delta}\lambda_{\max}\left(\sum_{i\in\Omega(\alpha,\delta)} x_i x_i^\top\right). \end{aligned} \tag{37}$$

Now we give the iterative formula to pursue the sparse solution of (33). Initialize β^0, we update β^{k+1} as

$$\beta^{k+1} \quad = \quad \operatorname{argmin}_{\beta\in\mathbb{R}^n} \; M_{\tau,\epsilon}(\beta,\beta^k,\delta) + \mu\|w\circ\beta\|_1. \tag{38}$$

Before proving the convergence of our proposed algorithm, we first give a property of $M_{\tau,\epsilon}(\beta,\alpha,\delta)+\mu\|w\circ\beta\|_1$ which is the majorized function of $S_\tau(\beta,\delta)+\mu\|w\circ\beta\|_1$.

Theorem 4.1. *If $\widehat{\beta}$ is a global minimizer of $M_{\tau,\epsilon}(\beta,\alpha,\delta) + \mu\|w\circ\beta\|_1$ for any fixed $\tau \in (0,1), \delta, \mu, \epsilon > 0, w, \alpha \in \mathbb{R}^n$ and $L_\alpha \geq \lambda_{\max}(\sum_{i=1}^n x_i x_i^T)$, then $\widehat{\beta}$ can be analytically expressed by*

$$\widehat{\beta} = \operatorname{sign}(\widetilde{\beta}) \circ \max\left\{|\widetilde{\beta}| - \frac{\mu}{L_\alpha+\epsilon}w, 0\right\},$$

where $\widetilde{\beta} = \alpha - \nabla S_\tau(\alpha,\delta)/(L_\alpha+\epsilon)$.

Proof. Since function $S_\tau(\beta,\delta)$ is smooth with respect to β on $\mathbb{R}^n$ when fixing $\delta > 0$, we acquire the majorized function from its second Taylor expansion at α whose formula is as follows:

$$\begin{aligned} S_\tau(\beta,\delta) \quad &= \quad \sum_{i=1}^m \rho_\tau(y_i - x_i^\top\beta,\delta) \\ &= \quad S_\tau(\alpha,\delta) + \langle\nabla S_\tau(\alpha,\delta), \beta-\alpha\rangle + \frac{1}{2}(\beta-\alpha)^\top\nabla^2 S_\tau(\alpha,\delta)(\beta-\alpha) \\ &\quad\quad +o(\|\beta-\alpha\|_2^2) \\ &\leq \quad S_\tau(\alpha,\delta) + \langle\nabla S_\tau(\alpha,\delta), \beta-\alpha\rangle + \frac{L_\alpha+\epsilon}{2}\|\beta-\alpha\|_2^2 \\ &= \quad M_{\tau,\epsilon}(\beta,\alpha,\delta), \end{aligned}$$

where $L_\alpha \geq \lambda_{\max}\left(\nabla^2 S_\tau(\alpha,\delta)\right)$. Henceforth, for any $\beta \neq \alpha \in \mathbb{R}^n$, we have

$$M_{\tau,\epsilon}(\beta,\alpha,\delta) > S_\tau(\beta,\delta) \text{ and } \; M_{\tau,\epsilon}(\beta,\beta,\delta) = S_\tau(\beta,\delta),$$

which means that $M_{\tau,\epsilon}(\beta,\alpha,\delta)$ is a majorization of $S_\tau(\beta,\delta)$. Using this majorization function, for given $\tau \in (0,1)$ and $\delta > 0$, we start with an initial iteration β^0 and update β^k by solving

$$\begin{aligned} \beta^{k+1} \quad &= \quad \operatorname{argmin}_{\beta\in\mathbb{R}^n} \; M_{\tau,\epsilon}(\beta,\beta^k,\delta) + \mu\|w\circ\beta\|_1 \\ &= \quad \operatorname{argmin}_{\beta\in\mathbb{R}^n} \; \langle\nabla S_\tau(\beta^k,\delta)\beta\rangle + \frac{L_k+\epsilon}{2}\|\beta-\beta^k\|_2^2 + \mu\|w\circ\beta\|_1, \end{aligned} \tag{39}$$

where $L_k \geq \lambda_{\max}\left(\sum_{i=1}^n x_i x_i^T)\right), \nabla S_\tau(\beta^k,\delta) = \sum_{i=1}^m \nabla\rho_\tau\left(y_i - x_i^\top\beta^k,\delta\right)$, which is equivalent to

$$\begin{aligned} \beta^{k+1} \quad &= \quad \operatorname{argmin}_{\beta\in\mathbb{R}^n} \; \frac{L_k+\epsilon}{2}\|\beta-\widetilde{\beta}^k\|_2^2 + \mu\|w\circ\beta\|_1 \\ &= \quad \operatorname{sign}(\widetilde{\beta}^k)\circ\max\left\{|\widetilde{\beta}^k| - \frac{\mu}{L_k+\epsilon}w, 0\right\}, \end{aligned} \tag{40}$$

where

$$\widetilde{\beta}^k := \beta^k - \nabla S_\tau(\beta^k,\delta)/(L_k+\epsilon).$$

Then the proposition 4.1 holds. □

Based on the proposition above, the following theorem establishes the relationship of the optimal solutions between the model (33) and the problem

$$\min \left\{ M_{\tau,\epsilon}(\beta,\widehat{\beta},\delta) + \mu\|w\circ\beta\|_1 \right\}.$$

Theorem 4.2. *Let $\tau\in(0,1),\delta,\epsilon,\mu>0$ and w be given. If $\widehat{\beta}$ is a global minimizer of (33), then $\widehat{\beta}$ is also the global minimizer of $M_{\tau,\epsilon}(\beta,\widehat{\beta},\delta)+\mu\|w\circ\beta\|_1$.*

Proof. For given $\widehat{\beta}\in\mathbb{R}^n$, since $M_{\tau,\epsilon}(\cdot,\widehat{\eta},\delta)$ is the majorization of $S_\tau(\cdot,\delta)$, we have

$$\begin{array}{rcl} M_{\tau,\epsilon}(\beta,\widehat{\beta},\delta)+\mu\|w\circ\beta\|_1 & \geq & S_\tau(\beta,\delta)+\mu\|w\circ\beta\|_1 \\ & \geq & S_\tau(\widehat{\beta},\delta)+\mu\|w\circ\widehat{\beta}\|_1 \\ & = & M_{\tau,\epsilon}(\widehat{\beta},\widehat{\beta},\delta)+\mu\|w\circ\widehat{\beta}\|_1, \end{array}$$

where the second inequality holds due to the fact that $\widehat{\beta}$ is a global minimizer of $S_\tau(\beta,\delta)+\mu\|w\circ\beta\|_1$, and the last equality is also derived from $M_{\tau,\epsilon}(\cdot,\widehat{\beta},\delta)$ being the majorization of $S_\tau(\cdot,\delta)$. □

Then we can built the necessary and sufficient condition, that is the fixed point equation, of the optimal solution of (33) from above theorem and proposition.

Theorem 4.3. *Let $\tau\in(0,1),\delta,\epsilon,\mu>0$ and w be given. Then $\widehat{\beta}$ is a global minimizer of (33) if and only if $\widehat{\beta}$ satisfies the following fixed point equation*

$$\widehat{\beta} = \mathrm{sign}(\widetilde{\beta})\circ\max\left\{|\widetilde{\beta}| - \frac{\mu}{\widehat{L}+\epsilon}w, 0\right\},$$

where $\widetilde{\beta} = \widehat{\beta} - \nabla S_\tau(\widehat{\beta},\delta)/(\widehat{L}+\epsilon)$.

Proof. Necessity: Since $\widehat{\beta}$ is a global minimizer of (33), it also is the global minimizer of $M_{\tau,\epsilon}(\beta,\widehat{\beta},\delta)+\mu\|w\circ\beta\|_1$ from Theorem 4.2. Then by Proposition 4.1 (with $\alpha=\widehat{\beta}$), $\widehat{\beta}$ satisfies the fixed point equation. Sufficiency: Due to the convexity of (33), its global minimizer is its stationary point β^* which contents

$$\langle \nabla S_\tau(\beta^*,\delta)+\mu w\circ\mathrm{sign}(\beta^*),\ \beta-\beta^*\rangle \geq 0,\ \forall\ \beta\in\mathbb{R}^n.$$

From the fixed point function, we have

$$\begin{array}{rcl} \widehat{\beta} & = & \mathrm{sign}(\widetilde{\beta})\circ\max\left\{|\widetilde{\beta}| - \frac{\mu}{\widehat{L}+\epsilon}w, 0\right\} \\ & = & \mathrm{argmin}_{\beta\in\mathbb{R}^n}\ M_{\tau,\epsilon}(\beta,\widehat{\beta},\delta)+\mu\|w\circ\beta\|_1 \\ & = & \mathrm{argmin}_{\beta\in\mathbb{R}^n}\ \frac{\widehat{L}+\epsilon}{2}\|\beta-\widetilde{\beta}\|_2^2+\mu\|w\circ\beta\|_1. \end{array} \tag{41}$$

Similarly, (41) is convex and thus its global minimizer is its stationary point $\widehat{\beta}$ satisfying

$$\left\langle (\widehat{L}+\epsilon)(\widehat{\beta}-\widetilde{\beta})+\mu w\circ\mathrm{sign}(\widehat{\beta}),\ \beta-\widehat{\beta}\right\rangle \geq 0,\ \forall\ \beta\in\mathbb{R}^n,$$

which is also equivalent to

$$\left\langle \nabla\varphi_\tau(\widehat{\beta},\delta)+\mu w\circ\mathrm{sign}(\widehat{\beta}),\ \beta-\widehat{\beta}\right\rangle \geq 0,\ \forall\ \beta\in\mathbb{R}^n,$$

because of $\widetilde{\beta} = \widehat{\beta} - \nabla S_\tau(\widehat{\beta}, \delta)/(\widehat{L} + \epsilon)$. Therefor $\widehat{\beta}$ is also the stationary point (also a global minimizer) of (33). Overall the whole proof is achieved. □

Now we are ready to prove the convergence of the proposed algorithm .

Theorem 4.4. *For given $\tau \in (0,1), \delta, \epsilon, \mu > 0$ and w, let $\{\beta^k\}$ be the sequence generated by (40) and $\Phi_\tau(\beta, \delta) = \{S_\tau(\beta, \delta) + \mu\|w \circ \beta\|_1\}$, then*

(A) $\Phi_\tau(\beta^k, \delta)$ is monotonically non-increasing and converges to $\Phi_\tau(\widehat{\beta}, \delta)$, where $\widehat{\beta}$ is any accumulation point of $\{\beta^k\}$;
(B) $\{\beta^k\}$ is asymptotically regular, namely, $\lim_{k\to\infty} \left\|\beta^{k+1} - \beta^k\right\|_2 = 0$;
(C) $\{\beta^k\}$ converges to the global minimizer of problem (33).

Proof. (A) Since $M_{\tau,\epsilon}(\cdot, \widehat{\eta}, \delta)$ is the majorization of $S_\tau(\cdot, \delta)$, it suffices to

$$\begin{aligned}
\Phi_\tau(\beta^{k+1}, \delta) &= S_\tau(\beta^{k+1}, \delta) + \mu\|w \circ \beta^{k+1}\|_1 \\
&\leq M_{\tau,\epsilon}(\beta^{k+1}, \beta^k, \delta) + \mu\|w \circ \beta^{k+1}\|_1 \\
&\leq M_{\tau,\epsilon}(\beta^k, \beta^k, \delta) + \mu\|w \circ \beta^k\|_1 \\
&= \Phi_\tau(\beta^k, \delta),
\end{aligned}$$

which signifies that $\{\Phi_\tau(\beta^k, \delta)\}$ is monotonically non-increasing. The second inequality derives from $\beta^{k+1} = \operatorname{argmin}_{\beta\in\mathbb{R}^n} M_{\tau,\epsilon}(\beta, \beta^k, \delta) + \mu\|w\circ\beta\|_1$. As $\{\Phi_\tau(\beta^k, \delta)\}$ is bounded from below, $\{\Phi_\tau(\beta^k, \delta)\}$ converges to a constant $\widehat{\Phi}$. From $\{\beta^k\} \subset \{\beta \mid \Phi_\tau(\beta, \delta) \leq \Phi_\tau(\beta^0, \delta)\}$ which is bounded, it yields that $\{\beta^k\}$ is also bounded, and thus $\{\beta^k\}$ has at least one accumulation point. Let $\widehat{\beta}$ be an accumulation point of $\{\beta^k\}$. By the continuity of $\Phi_\tau(\beta, \delta)$ and the convergence of $\{\Phi_\tau(\beta^k, \delta)\}$, we get $\{\Phi_\tau(\beta^k, \delta)\} \to \widehat{\Phi} = \Phi_\tau(\widehat{\beta}, \delta)$ as $k \to \infty$.

(B) Through simply computing, we have

$$\begin{aligned}
&\Phi_\tau(\beta^k, \delta) - \Phi_\tau(\beta^{k+1}, \delta) \\
= \;& S_\tau(\beta^k, \delta) + \mu\|w \circ \beta^k\|_1 - S_\tau(\beta^{k+1}, \delta) + \mu\|w \circ \beta^{k+1}\|_1 \\
= \;& M_{\tau,\epsilon}(\beta^k, \beta^k, \delta) + \mu\|w \circ \beta^k\|_1 - S_\tau(\beta^{k+1}, \delta) + \mu\|w \circ \beta^{k+1}\|_1 \\
\geq \;& M_{\tau,\epsilon}(\beta^{k+1}, \beta^k, \delta) + \mu\|w \circ \beta^{k+1}\|_1 - S_\tau(\beta^{k+1}, \delta) + \mu\|w \circ \beta^{k+1}\|_1 \qquad (42)\\
= \;& M_{\tau,\epsilon}(\beta^{k+1}, \beta^k, \delta) - S_\tau(\beta^{k+1}, \delta) \\
= \;& S_\tau(\beta^k, \delta) + \langle \nabla S_\tau(\beta^k, \delta), \beta^{k+1} - \beta^k\rangle + \frac{L_k + \epsilon}{2}\|\beta^{k+1} - \beta^k\|_2^2 \\
& -S_\tau(\beta^k, \delta) - \langle \nabla S_\tau(\beta^k, \delta), \beta^{k+1} - \beta^k\rangle \\
& -\frac{1}{2}(\beta^{k+1} - \beta^k)^\top \nabla^2 S_\tau(\beta^k, \delta)(\beta^{k+1} - \beta^k) - o(\|\beta^{k+1} - \beta^k\|_2^2) \\
= \;& \frac{\epsilon}{2}\|\beta^{k+1} - \beta^k\|_2^2 + \frac{1}{2}(\beta^{k+1} - \beta^k)^\top \left(L_k I - \nabla^2 S_\tau(\beta^k, \delta)\right)(\beta^{k+1} - \beta^k) \\
& -o(\|\beta^{k+1} - \beta^k\|_2^2) \\
\geq \;& \frac{\epsilon}{2}\|\beta^{k+1} - \beta^k\|_2^2,
\end{aligned}$$

where the first inequality derives from $\beta^{k+1} = \operatorname{argmin}_{\beta\in\mathbb{R}^n} M_{\tau,\epsilon}(\beta, \beta^k, \delta) + \mu\|w \circ \beta\|_1$. The forth equality is from the Taylor expansion of $S_\tau(\beta^{k+1}, \delta)$ at point β^k, and the last inequality holds due to $L_k \geq \lambda_{\max}\left(\sum_{i=1}^n x_i x_i^T\right)$ which suffices to $L_k I - \nabla^2 S_\tau(\beta^k, \delta) \succeq L_k I - \sum_{i=1}^n x_i x_i^T \succeq 0$.

Henceforth, for given $\epsilon > 0$ and any positive integer N,

$$\sum_{k=0}^{N} \|\beta^{k+1} - \beta^k\|_2^2 \le \frac{2}{\epsilon} \sum_{k=0}^{N} \left(\Phi_\tau(\beta^k, \delta) - \Phi_\tau(\beta^{k+1}, \delta)\right) \le \frac{2}{\epsilon} \Phi_\tau(\beta^0, \delta),$$

which implies that $\sum_{k=0}^{\infty} \|\beta^{k+1} - \beta^k\|_2^2 < \infty$ and thus $\lim_{k\to\infty} \|\beta^{k+1} - \beta^k\|_2 = 0$, that is $\{\beta^k\}$ is asymptotically regular.

(C) Let $\left\{\beta^{k_j}\right\}$ be a convergent subsequence of $\left\{\beta^{k_j}\right\}$ and $\widehat{\beta}$ be its limit point, i.e.

$$\beta^{k_j} \to \widehat{\beta}, \text{ as } k_j \to \infty. \tag{43}$$

Since $\nabla S_\tau(\alpha, \delta)$ is continuously differential, combining with the limitation above, one can immediately derive

$$\widetilde{\beta}^{k_j} = \beta^{k_j} - \frac{\nabla S_\tau(\beta^{k_j}, \delta)}{L_{k_j} + \epsilon} \longrightarrow \frac{\widehat{\beta} - \nabla S_\tau(\widehat{\beta}, \delta)}{L_* + \epsilon} =: \widetilde{\beta}, \text{ as } k_j \longrightarrow \infty, \tag{44}$$

where $\widehat{L} \ge \lambda_{\max}\left(\sum_{i=1}^n x_i x_x^T\right)$. Then the limitation (43) and the asymptotical regularity of $\left\{\beta^k\right\}$ imply

$$\|\beta^{k_j+1} - \widehat{\beta}\|_2 \le \|\beta^{k_j+1} - \beta^{k_j}\|_2 + \|\beta^{k_j} - \widehat{\beta}\|_2 \longrightarrow 0, \text{ as } k_j \longrightarrow \infty, \tag{45}$$

which guarantees that $\left\{\beta^{k_j+1}\right\}$ also converges to $\widehat{\beta}$. On the other side, by (40) and (44), we have

$$\begin{aligned} \beta^{k_j+1} &= \operatorname{sign}(\widetilde{\beta}^{k_j}) \circ \max\left\{|\widetilde{\beta}^{k_j}| - \frac{\mu}{L_{k_j} + \epsilon} w, 0\right\} \\ &\longrightarrow \operatorname{sign}(\widetilde{\beta}) \circ \max\left\{|\widetilde{\beta}| - \frac{\mu}{\widehat{L} + \epsilon} w, 0\right\}, \text{ as } k_j \longrightarrow \infty, \end{aligned}$$

which manifests $\widehat{\beta} = \operatorname{sign}(\widetilde{\beta}) \circ \max\left\{|\widetilde{\beta}| - \frac{\mu}{\widehat{L}+\epsilon} w, 0\right\}$. Finally, $\widehat{\beta}$ is the global minimizer of (33) from Theorem 4.3. Since (33) is strongly convex, $\widehat{\beta}$ is unique. So $\{\beta^k\}$ converges to the global minimizer of (33). □

To solve the model (33), we now present the algorithm framework of modified iterative reweighted ℓ_1 minimization (MIRL1) in Table 1.

TABLE 1. The framework of MIRL1 .

```
Modified iterative reweighted $\ell_1$ minimization (MIRL1)
Initialize $\tau, \gamma \in (0,1), \delta^1 > 0, \epsilon > 0, \beta^0, w^1, M, \mu^1$;
For $t = 1 : M$
    Initialize $\beta^{t,1} = \beta^{t-1}$;
    While $\|\beta^{t,k+1} - \beta^{t,k}\|_2 \ge \mu^t \max\left\{1, \|\beta^{t,k}\|_2\right\}$
        Compute $L_{t,k} \ge \frac{1}{2\delta^t} \min\left\{\lambda_{\max}\left(\sum_{\Omega\left(\beta^{t,k}, \delta^\top\right)} x_i x_i^\top\right), \lambda_{\max}(XX^\top)\right\}$;
        Compute $\widetilde{\beta}^{t,k} = \beta^{t,k} - \nabla S_\tau(\beta^{t,k}, \delta^t)/(L_{t,k} + \epsilon)$;
        Compute $\beta^{t,k+1} = \operatorname{sign}\left(\widetilde{\beta}^{t,k}\right) \circ \max\left\{|\widetilde{\beta}^{t,k}| - \frac{\mu^t}{L_{t,k}+\epsilon} w^t, 0\right\}$.
    End
    Update $\delta^{t+1} = \gamma\delta^t$ and $\beta^t = \beta^{t,k+1}$;
    Update $w^{t+1}$ from $\beta^{t-1}$ and $\beta^t$ based on (46), (47) and (48);
End
```

Based on the existing way of giving the weight [21], we introduce it to value the w whose the t-th iteration w^t is updated by

$$T^t \quad = \quad \mathrm{argmax}_{|T|=s_t, T\subseteq\{1,\cdots,n\}} \ \|h_T^t\|_1, \ t = 1, 2, \cdots \tag{46}$$

$$w_i^t = \begin{cases} \left[\dfrac{|h_i^t| + \varepsilon}{\max_{j\notin T^t} |h_j^t|}\right]^{q-1} & , i \in T^t, \\ 1 \ , i \notin T^t, \end{cases} \tag{47}$$

where

$$h^t = \beta^t - \beta^{t-1}, \ s_t = |\mathrm{supp}(\beta^t)| \tag{48}$$

and $0 < q \leq 1, \varepsilon > 0$ is sufficiently small. One can easily find that T^t coincides with the indices of the s_t largest entries of $|h^t|$. See [21] for more details.

Theorem 4.5. *The inter loops of MIRL1 are convergent, that is, for each t, the sequence of $\{\beta^{t,k}\}$ globally converges to*

$$\beta^t = \mathrm{argmin}_{\beta\in\mathbb{R}^n} \ M_{\tau,\epsilon}(\beta, \beta^{t-1}, \delta) + \mu^t \|w^t \circ \beta\|_1.$$

5. Numerical experiments.

In this section, the numerical experiments are presented to demonstrate the performance of the proposed approach. From our experience, we know that quantile Huber penalty performs better than the first smoothing function. So we adopt quantile Huber penalty in our numerical experiments. We will take advantage of smoothing functions $\rho_\tau(\cdot, \delta)$ to relax the regression quantiles function $\rho_\tau(\cdot)$ under selecting different quantiles parameter τ.

5.1. Example I: Simulated experiment.

In the simulation we considered linear model,

$$y = X\beta + \xi$$

with sample matrix $X = (x_1, x_2, \cdots, x_m)^\top \in \mathbb{R}^{m\times n}$ being from Gaussian matrices, and $\mathbb{E}(\xi) = 0$, $\mathrm{Var}(\xi) = \sigma^2$. We here let ξ obeys the Normal distribution and the Log-normal distribution which is one of heavy-tailed distributions. We first design the true regression quantiles estimator β^* from the generated measurements X. We randomly generate 20 samples. For each data set, the random matrix X and vector y are generated by the following MATLAB codes:

$$\begin{aligned} & b = randperm(n), \ \beta^* = zeros(n, 1), \\ & \beta^*(b(1:s)) = randn(s, 1), \ X = randn(m, n), \\ & y1 = X\beta^* + \sigma * randn(0, 1), \ y2 = X\beta^* + \sigma * lognrnd(0, 1, n, 1), \end{aligned}$$

where the sparsity s of the true regression quantiles estimator β^* is always settled as $s = 1\% \times n$. The parameter σ will be taken as $\sigma = 0.01$ or 0.25.

By exploiting method MIRL1 to compute the approximately optimal solution, we denote it as $\widehat{\beta}$. In Table 2, we compare the estimation errors $\|\widehat{\beta} - \beta^*\|_2$, the prediction errors $\frac{1}{\sqrt{m}} \parallel X\widehat{\beta} - X\beta^* \parallel_2$, the CPU time, and we add to another two types of errors which are

$$\mathrm{FPR} := \frac{\mathrm{Card}\left\{ j : \ \beta_j^* \neq 0 \ \& \ \widehat{\beta}_j = 0 \right\}}{\mathrm{Card}\left\{ j : \widehat{\beta}_j = 0 \right\}}, \ \mathrm{TPR} := \frac{\mathrm{Card}\left\{ j : \ \beta_j^* \neq 0 \ \& \ \widehat{\beta}_j \neq 0 \right\}}{\mathrm{Card}\left\{ j : \widehat{\beta}_j \neq 0 \right\}},$$

where FPR stands for the false positive rate, which means the rate of significant variables that are unselected over the whole zero entries, and TPR denotes the true positive rate, which implies the ratio of significant variables that are selected over the entire none zero elements. It is worth mentioning that the smaller FPR and the larger TPR is, the better our approach would perform.

TABLE 2. $n = 512, m = 128$.

τ	Noise	$\|\widehat{\beta} - \beta^*\|_2$	$\frac{1}{\sqrt{m}}\|X\widehat{\beta} - X\beta^*\|_2$	FPR	TPR	Time(s)
0.1	$N(0, \sigma^2)$	0.0063	0.0078	0	1.0000	2.7474
	$LN(0, \sigma^2)$	0.0146	0.0177	0	0.9429	3.0230
0.3	$N(0, \sigma^2)$	0.0067	0.0091	0	1.0000	2.4950
	$LN(0, \sigma^2)$	0.0167	0.0209	0	0.8762	2.2424
0.5	$N(0, \sigma^2)$	0.0082	0.0097	0	1.0000	2.2793
	$LN(0, \sigma^2)$	0.0175	0.0201	0	0.8563	2.1709
0.7	$N(0, \sigma^2)$	0.0069	0.0093	0	1.0000	2.4489
	$LN(0, \sigma^2)$	0.0208	0.0258	0	0.8028	2.4478
0.9	$N(0, \sigma^2)$	0.0062	0.0077	0	1.0000	2.6616
	$LN(0, \sigma^2)$	0.0243	0.0299	0	0.6254	2.7624

TABLE 3. $n = 1024, m = 256$.

τ	Noise	$\|\widehat{\beta} - \beta^*\|_2$	$\frac{1}{\sqrt{m}}\|X\widehat{\beta} - X\beta^*\|_2$	FPR	TPR	Time(s)
0.1	$N(0, \sigma^2)$	0.0069	0.0091	0	1.0000	13.3403
	$LN(0, \sigma^2)$	0.0194	0.0257	0	0.9818	15.4523
0.3	$N(0, \sigma^2)$	0.0070	0.0093	0	1.0000	10.1645
	$LN(0, \sigma^2)$	0.0193	0.0244	0	1.0000	11.2862
0.5	$N(0, \sigma^2)$	0.0076	0.0096	0	1.0000	11.4844
	$LN(0, \sigma^2)$	0.0201	0.0252	0	1.0000	11.0627
0.7	$N(0, \sigma^2)$	0.0074	0.0097	0	1.0000	12.2611
	$LN(0, \sigma^2)$	0.0206	0.0264	0	0.9818	12.3306
0.9	$N(0, \sigma^2)$	0.0070	0.0093	0	1.0000	13.6169
	$LN(0, \sigma^2)$	0.0226	0.0286	0	0.9538	14.1217

Table 2 and Table 3 are the numerical results recording the average of estimation error $\|\widehat{\beta} - \beta^*\|_2$, the prediction error $\frac{1}{\sqrt{m}}\|X\widehat{\beta} - X\beta^*\|_2$, FPR and TPR over 100 simulations under several τ and CPU times when $n = 512$ or $n = 1024$ respectively. As indicated in Tables 2 and 3, the most significant character is that the values of FPR are all as small as zeros and the values of TPR are all almost equal ones (particularly for the Normal distributed case) respectively, which signifies that all significant variables are selected and insignificant ones basically are not chosen. As for the errors, it is not difficult to see that the two types of errors derived from Normal distribution are far smaller than that stemmed from Log-normal distribution. More specifically, for the Log-normal distribution when τ is close to 0.1 with relatively low error, the model performs much better, seeing Figure 3 for more visualized clarity.

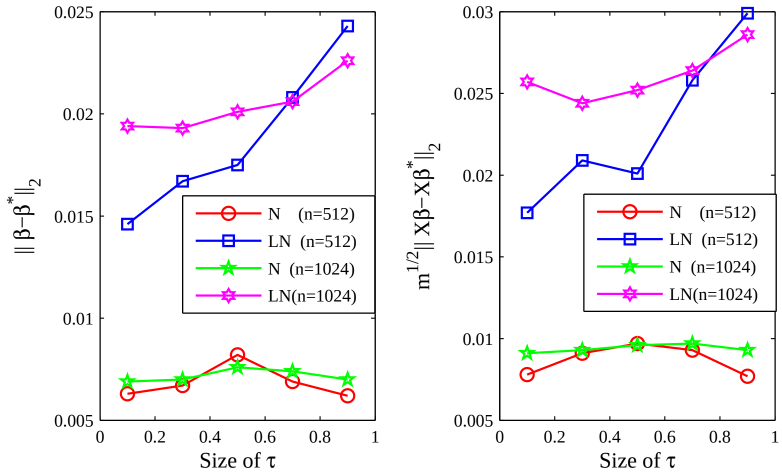


FIGURE 3. Errors for different τ, n, and noise in Example 1

5.2. **Example II: Toeplitz correlation matrix.** The second modified example is from [19], which aims at considering the estimator $\beta^* = (3, ..., 3, 0, ..., 0)^\top$, with s (being taken $s = 1\%n$) none zero entries 3 in β^*. In the simulation study, each row of the design matrix X is generated by $N(0, \Sigma)$ distribution with Toeplitz correlation matrix $\Sigma_{ij} = (1/2)^{|i-j|}$, i.e., $x_i \sim \Sigma^{1/2}N(0, 1)$, $i = 1, 2, ..., n$; and then normalize the columns of X such that each column has L_2 norm $\sqrt{n}$. We use two noise patterns: (a) $N(0, \sigma^2)$-normal noise, (b) $LN(0, \sigma^2)$-lognormal noise. Corresponding MATLAB cades are:

$$y1 = X\beta^* + \sigma * randn(0, 1),$$
$$y2 = X\beta^* + \sigma * lognrnd(0, 1, n, 1).$$

TABLE 4. $n = 512, m = 128$.

τ	Noise	$\|\widehat{\beta} - \beta^*\|_2$	$\frac{1}{\sqrt{m}}\|X\widehat{\beta} - X\beta^*\|_2$	FPR	TPR	Time(s)
0.1	$N(0, \sigma^2)$	0.0071	0.0096	0	1.0000	3.3895
	$LN(0, \sigma^2)$	0.0185	0.0242	0	1.0000	2.5908
0.3	$N(0, \sigma^2)$	0.0071	0.0098	0	1.0000	2.6341
	$LN(0, \sigma^2)$	0.0189	0.0246	0	1.0000	2.1094
0.5	$N(0, \sigma^2)$	0.0071	0.0094	0	1.0000	1.7425
	$LN(0, \sigma^2)$	0.0184	0.0224	0	1.0000	1.3525
0.7	$N(0, \sigma^2)$	0.0070	0.0095	0	1.0000	1.6148
	$LN(0, \sigma^2)$	0.0196	0.0249	0	0.9667	1.3640
0.9	$N(0, \sigma^2)$	0.0072	0.0095	0	1.0000	2.3751
	$LN(0, \sigma^2)$	0.0186	0.0227	0	1.0000	2.8742

Table 4 and Table 5 are the numerical results recording the average of estimation error $\|\widehat{\beta} - \beta^*\|_2$, the prediction error $\frac{1}{\sqrt{m}}\|X\widehat{\beta} - X\beta^*\|_2$, FPR and TPR over 100

TABLE 5. $n = 1024, m = 256$.

τ	Noise	$\|\widehat{\beta}-\beta^*\|_2$	$\frac{1}{\sqrt{m}}\|X\widehat{\beta}-X\beta^*\|_2$	FPR	TPR	Time(s)
0.1	$N(0,\sigma^2)$	0.0071	0.0096	0	1.0000	17.5535
	$LN(0,\sigma^2)$	0.0186	0.0246	0	1.0000	13.5658
0.3	$N(0,\sigma^2)$	0.0074	0.0093	0	1.0000	13.3480
	$LN(0,\sigma^2)$	0.0196	0.0261	0	1.0000	8.8331
0.5	$N(0,\sigma^2)$	0.0076	0.0094	0	1.0000	9.8347
	$LN(0,\sigma^2)$	0.0183	0.0241	0	1.0000	7.6007
0.7	$N(0,\sigma^2)$	0.0069	0.0093	0	1.0000	13.6823
	$LN(0,\sigma^2)$	0.0200	0.0272	0	1.0000	7.9791
0.9	$N(0,\sigma^2)$	0.0072	0.0094	0	1.0000	18.5440
	$LN(0,\sigma^2)$	0.0184	0.0249	0	1.0000	14.3975

simulations under several τ and CPU times when $n = 512$ or $n = 1024$ respectively. As shown in Tables 4 and 5, the most significant character is that the values of FPR are all as small as zeros and the values of TPR are all almost equal ones (particularly for the Normal distributed case) respectively, which signifies that all significant variables are selected and insignificant ones basically are not chosen. As for the errors, it is not difficult to see that the two types of errors derived from Normal distribution are far smaller than that stemmed from Log-normal distribution. One can also see Figure 4 for more visualized clarity.

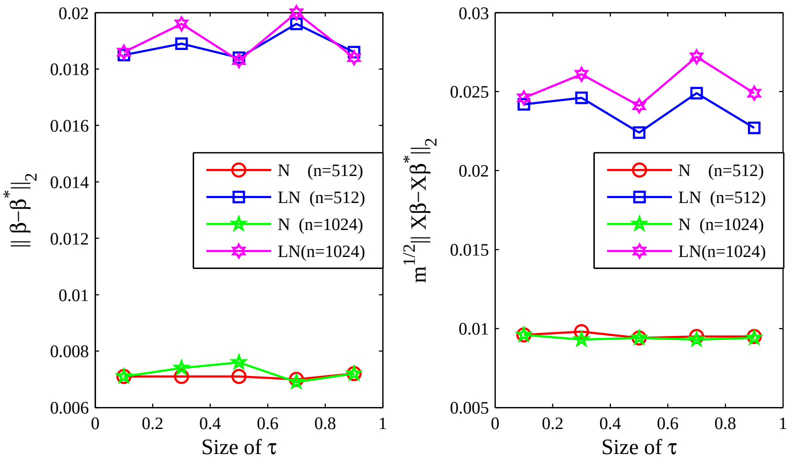


FIGURE 4. Errors for different τ, n, and noise in Example 2

6. **Conclusion.** In this paper we considered the high-dimensional linear regression model with assumption of sparsity. We introduced the quantile regression with the nonconvex ℓ_p penalty $(0 < p < 1)$ for the high-dimensional linear sparse model. Since it is a nonconvex and nonsmooth NP-hard problem, we used smoothing techniques to deal with it. More specifically, we first introduced a definition of smoothing functions. Then we gave two smoothing functions to alter the quantile loss function.

One smoothing function is a generalization of the smoothing function in [6] which is a smoothing of absolute value function. The other smoothing function is called quantile Huber penalty. We also gave the estimation of approximation by our different smoothing functions through Theorem 2.3 and Theorem 2.6. Then, we derived two types of lower bounds for any local solution of the smoothing model with the help of ℓ_p regularization . Moreover, with the help of weighted ℓ_1 regularization, we proposed a smoothing iterative method for the smoothing quantile regression and established its global convergence. Finally, we reported the numerical experiments which illustrate the efficient performance of our method.

Acknowledgments. The authors would like to thank the anonymous referee and the associate editor for their numerous insightful comments and suggestions, which have greatly improved the paper. The work was supported in part by the National Natural Science Foundation of China (11431002,11171018).

REFERENCES

[1] A. Y. Aravkin, A. Kambadur, A. C. Lozano and R. Luss, Sparse quantile huber regression for efficient and robust estimation, preprint, arXiv:1402.4624.

[2] P. Bühlmann and S. V. D. Geer, *Statistics for High-Dimensional Data: Methods, Theory and Applications,* Springer Science & Business Media, 2011.

[3] E. J. Candès, M. B. Wakin and S. P. Boyd, Enhancing sparsity by reweighted ℓ_1 minimization, *Journal of Fourier Analysis and Applications*, **14** (2008), 877–905.

[4] T. T. Cai, L. Wang and G. W. Xu, Shifting inequality and recovery of sparse signals, *IEEE Trans. Signal Process*, **58** (2010), 1300–1308.

[5] X. J. Chen, F. M. Xu and Y. Y. Ye, Lower bound theory of nonzero entries in solutions of ℓ_2-ℓ_p minimization, *SIAM J. Sci. Comput.*, **32** (2010), 2832–2852.

[6] X. J. Chen and W. J. Zhou, Smoothing nonlinear conjugate gradient method for image restoration using nonsmooth nonconvex minimization, *SIAM J. Imaging Sci.*, **3** (2010), 765–790.

[7] T. T. Cai and A. Zhang, Sharp RIP bound for sparse signal and low-rank matrix recovery, *Applied and Computational Harmonic Analysis*, **35** (2013), 74–93.

[8] D. L. Donoho and M. Elad, Optimally sparse representation in general (nonorthogonal) dictionaries via ℓ_1 minimization, *PNAS*, **100** (2003), 2197–2202.

[9] I. Daubechies, R. DeVore, M. Fornasier and C. S. Güntürk, Iteratively reweighted least squares minimization for sparse recovery, *Communications on Pure and Applied Mathematics*, **63** (2010), 1–38.

[10] F. Francisco and J. S. Pang, *Finite-Dimensional Variational Inequalities and Complementarity Problems,* Springer-Verlag, New York, 2007.

[11] J. Q. Fan, F. Han and H. Liu, Challenges of big data analysis, *National Science Review*, **1** (2014), 239–314.

[12] J. Q. Fan, Y. Y. Fan and E. Barut, Adaptive robust variable selection, *Annals of Statistics*, **42** (2014), 324–351.

[13] R. Koenker and G. Bassett, Regression quantiles, *Econometrica*, **46** (1978), 33–50.

[14] R. Koenker, *Quantile Regression,* Cambridge University Press, 2005.

[15] M. J. Lai and J. Wang, An unconstrained l_q minimization with $0 < q \leq 1$ for sparse solution of underdetermined linear systems, *SIAM J. Optim.*, **21** (2011), 82–101.

[16] L. C. Kong, J. Sun and N. H. Xiu, A regularized smoothing newton method for symmetric cone complimentarity problem, *SIAM J. Optim.*, **19** (2008), 1028–1047.

[17] Z. S. Lu, Iterative reweighted minimization methods for ℓ_p regularized unconstrained nonlinear programming, *Mathematical Programming*, **147** (2014), 227–307.

[18] L. Wang, Y. Wu and R. Li, Quantile regression for analyzing heterogeneity in ultra highdemension, *Journal of the American Statistical Association*, **107** (2012), 214–222.

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

[20] Y. B. Zhao and D. Li, Reweighted ℓ_1-minimization for sparse solutions to underdetermined linear systems, *SIAM J. Optim.*, **22** (2012), 1065–1088.

[21] S. L. Zhou, N. H. Xiu, Y. N. Wang and L. C. Kong, Exact recovery for sparse signal via weighted ℓ_1 minimization, preprint, arXiv:1312.2358.

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

Received June 2014; 1st revision May 2015; 2nd revision December 2015.

E-mail address: zhanglianjun0709@163.com
E-mail address: konglchen@126.com
E-mail address: liyan2010@uibe.edu.cn
E-mail address: longnan_zsl@163.com