Yokohama Publishers
ISSN 1349-8169 ONLINE JOURNAL

Pacific Journal of Optimization
Vol. 13, No. 2, pp. 279–300, April 2017

THE L_1-PENALIZED QUANTILE REGRESSION FOR TRADITIONAL CHINESE MEDICINE SYNDROME MANIFESTATION*

Yanqing Liu, Guokai Liu, Xianchao Xiu and Shenglong Zhou

Abstract: Traditional Chinese medicine syndrome manifestation is a nonlinear complex system, which has attracted much attention on its role in clinical study. With the help of the modern technique of the big data analysis, in this paper we provide a high-dimensional quantile regression model for Traditional Chinese medicine syndrome manifestation, where the term "high-dimensional" means that the number of observations is much less than the number of covariates. Moreover, we assume that the unknown vector is sparse, so we propose the use of an L_1-penalized quantile regression estimator to solve the model. Our estimator does not need any knowledge of of standard deviation of the noises or any moment assumptions of the noises. We show that the L_1-penalized quantile regression estimator (QRE) possesses near oracle performance, i.e. with large probability, the L_2 norm of the estimation error is of order $O(\sqrt{s(\log p)/n})$. The result is true for a wide range of noise distributions, even for the Cauchy distribution. In addition, we apply an alternating direction method to find the L_1-penalized QRE, which possesses the global convergence. Numerical results are reported to demonstrate the efficacy of our proposed method.

Key words: *high-dimensional linear regression, traditional chinese medicine syndrome manifestation, L_1-penalty, quantile regression, variable selection, alternating direction method*

Mathematics Subject Classification: *65K05, 90C90, 97K80*

1 Introduction

Traditional Chinese medicine syndrome manifestation is a nonlinear complex system, and its role in many diseases such as type 1 diabetes mellitus is still not clear, see, e.g., [1,29,30]. In particular, Lien et al [29] recently showed that integrative traditional Chinese medicine may reduce the risk of diabetic ketoacidosis in patients with type 1 diabetes mellitus. Based on the clinical study, they stated that traditional Chinese medicine may have a substantial positive impact on the management of type 1 diabetes mellitus. With the help of modern technique in big data analysis, large gene expression studies, such as those conducted using DNA arrays, often provide millions of different pieces of data. For instance, DNA arrays have been used to study variation in gene expression across collections of related samples from patients with some disease.

On the other hand, the high dimensional linear regression model says that the number of observations is much less than the number of unknown coefficients. High-dimensional data

*The work was supported in part by the National Natural Science Foundation of China (11431002, 11671029) and the Fundamental Research Funds for the Central Universities (KSJB16013536, 2016JBM081).

are frequently collected in a large variety of research areas such as genomics, functional magnetic resonance imaging, tomography, economics, and finance. Analysis of high-dimensional data poses many challenges and has attracted much recent interests in a number of fields such as applied math, electronic engineering, and statistics. Since we are considering a high dimensional linear regression problem, a key assumption is the sparsity of the true coefficient. This is the so-called the high-dimensional sparse linear model. The ordinary least squares method is not consistent in this setting since using least squares will not lead to a sparse solution in this context. In recent years, many new methods have been proposed to solve this problem. Methods based on L_1 penalization or constrained L_1 minimization have been extensively studied, see, e.g., [4, 14, 32, 35], where they demonstrated the fundamental result that L_1-penalized least squares estimators achieve the rate $O(\sqrt{s(\log p)/n})$, which is very close to the oracle rate $O(\sqrt{s/n})$. The lasso method has nice properties under the Gaussian assumption and a known variance. However, the Gaussian assumption may not hold in practice and the estimation of the standard deviation is not easy.

Quantile regression introduced by Koenker and Bassett [26] has become a popular and important tool in statistical analysis, which includes the well-known median regression or least absolute deviation (LAD) as a special case. A comprehensive review can be found in Koenker [25] and a general overview of many interesting recent developments in He [22]. More recently, LAD regression and quantile regression have been used for dealing with the cases where the error distribution is unknown or may have a heavy tail, see, e.g., [2, 6, 20, 35, 36]. Belloni and Chernozhukov [2] studied the L_1-penalized quantile regression in high-dimensional sparse models where the dimensionality could be larger than the sample size. We refer to their method as robust Lasso (R-Lasso). They showed that the R-Lasso estimate is consistent at the near-oracle rate, and gave conditions under which the selected model includes the true model, and derived bounds on the size of the selected model, uniformly in a compact set of quantile indices. Bradic et al. [6] introduced the penalized composite likelihood method for robust estimation in ultra-high dimensions with focus on the efficiency of the method. They still assumed sub-Gaussian tails. Wang et al. [36] 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. Wang [35] studied the L_1-penalized LAD regression and showed that the estimate achieves near oracle risk performance with a nearly universal penalty parameter and established a sure screening property for such an estimator. Fan et al. [20] studied the penalized quantile regression with the weighted L_1-penalty (WR-Lasso) and proposed a two-step procedure, called adaptive robust Lasso (AR-Lasso). Our theoretical results also reveal that adaptive choice of the weight vector is justified theoretically to possess the oracle property and the asymptotic normality. Note that properties of the estimator in [2] were presented under restricted eigenvalue type conditions and smooth assumptions on the density function of the noise, while in [20] different conditions on the model error distribution and adaptive choice of the weight vector are needed. However, in [35] they allowed for new interesting general noise structure and also the noiseless case.

Motivated by the above arguments and the uncertainty of the role of traditional Chinese medicine, we want to find out some valuable information which may support the important role of integrative traditional Chinese medicine in clinic study. In order to do so, in this paper we provide a high-dimensional L_1-penalized Quantile Regression for traditional Chinese medicine syndrome manifestation. In mathematics and statistics, an interesting question is whether the results in [35] can be extended from L_1-penalized LAD to L_1-penalized quantile regression for the fixed design case.

This paper will focus on this issue and we give an affirmative answer. We consider the selection of the penalty level for the L_1-penalized quantile regression, which does not depend on any unknown parameters or the noise distribution. Our analysis shows that the L_1-penalized quantile regression has near oracle properties as given in [35] for L_1-penalized LAD. Similarly, we do not have any assumptions on the moments of the noise and we only need a scale parameter to control the tail probability of the noise. The result is true for a wide range of noise distributions, even for Cauchy distributed noise, where the first order moment does not exist.

Moreover, the L_1-penalized quantile regression is a convex optimization problem, and it looks computationally efficient. However, the practical computational methods for quantile regression estimation now mainly cover three algorithms: simplex, interior point, and smoothing, for details, see, e.g., [15,25]. In view of the high-dimensional data analysis, this paper will apply another kind of algorithms to find the L_1-penalized QRE. We introduce an alternating direction method to find the estimator of the L_1-penalized QR model, and the global convergence result is maintained for the proposed method.

This paper is organized as follows. We introduce the L_1-penalized quantile regression for high-dimensional linear models and discuss the choice of the penalty level in Section 2. Then, we present the main results about the L_1-penalized quantile regression estimation error and several critical lemmas in Section 3. We propose an alternating direction method to find the estimator and report numerical results to demonstrate the efficacy of our method in Section 4. Technical lemmas and proofs of theorems are given in Appendix.

$\boxed{2}$ L_1-penalized Quantile Regression

We will introduce the L_1-penalized quantile regression for high-dimensional linear models and then discuss the choice of the penalty level. We begin with the following high-dimensional linear regression model

$$Y = X\beta + \varepsilon \tag{2.1}$$

where $X = (x_1, x_2, \cdots, x_n)^T = (X_1, X_2, \cdots, X_p)$ is an $n \times p$ fixed design matrix, $Y = (y_1, y_2, \cdots, y_n)^T$ is an n-dimensional response/observation vector, $\beta = (\beta_1, \beta_2, \cdots, \beta_p)^T$ is a p-dimensional regression coefficient vector, and $\varepsilon = (\varepsilon_1, \varepsilon_2, \cdots, \varepsilon_n)^T$ is an n-dimensional measurement error/noise vector. Throughout the paper, we assume $X = (X_1, X_2, \cdots, X_p)$ with each vector X_i being normalized such that $\|X_i\|_2 = \sqrt{n}$ for $i = 1, 2, \cdots, p$, and $\varepsilon = (\varepsilon_1, \varepsilon_2, \cdots, \varepsilon_n)^T$ with all components $\varepsilon_i (i = 1, 2, \cdots, n)$ are independently distributed and satisfy $P(\varepsilon_i \leq 0) = \tau$ for some known constant $\tau \in (0, 1)$. Under this model, $x_i^T\beta$ is the conditional τth-quantile of y_i given x_i. We will focus on the high dimensional case where the number of observations n is much less than the number of unknown coefficients p. For the high-dimensional linear regression problem, a key assumption is the sparsity of the true coefficient β^*, which guarantees the model identifiability and enhances the model fitting accuracy and interpretability, see, e.g, [21, 34]. We assume $T = \text{supp}(\beta^*)$ with the cardinality of T, $|T| = s < n$. The set T of nonzero coefficients or significant variables of β^* is unknown. In what follows, the parameters β, p and s are implicitly dependent of the sample size n, but we omit the index n in our notation whenever there is no confusion.

To reconstruct the high-dimensional linear model and estimate the sparse vector β^*, we consider the following L_1-penalized quantile regression problem

$$\min_{\beta} Q_\tau(Y - X\beta) + \lambda\|\beta\|_1 \tag{2.2}$$

where $Q_\tau(Y - X\beta) = \sum_{i=1}^n \rho_\tau(y_i - x_i^T\beta)$ with $\rho_\tau(u) = u(\tau - 1\{u \leq 0\})$ is the *quantile loss function*, $\|\beta\|_1 = \sum_{i=1}^p |\beta_i|$ is the L_1 norm of β, and $\lambda \geq 0$ is the penalty/regularization parameter. When $\tau = \frac{1}{2}$, it is the median regression and the so-called L_1-penalized least absolute deviation (L_1-penalized LAD). The use of quantile loss function in (2.2) is to deal with the cases where the error distribution is unknown or may have a heavy tail. Since $P(\varepsilon \leq 0) = \tau$, the solution of (2.2) can be interpreted as the sparse estimation of the conditional τth quantile. Then the L_1-penalized quantile regression estimator can be defined as

$$\hat{\beta} \in \text{argmin}_\beta Q_\tau(Y - X\beta) + \lambda\|\beta\|_1. \tag{2.3}$$

When $\tau = \frac{1}{2}$, the above becomes the L_1-penalized LAD estimator. Recently, the L_1-penalized LAD was studied in several papers, where the variable selection and estimation properties were discussed, see, e.g., [2, 27, 35]. Especially, Wang [35] proved that the L_1-penalized LAD estimator achieves near oracle risk performance with a nearly universal penalty parameter, while they allowed for new interesting general noise structure and also the noiseless case. We will extend these results from the L_1-penalized LAD to the L_1-penalized quantile regression estimator.

We below consider the choice of the penalty level for the L_1-penalized quantile regression estimator. We will choose a penalty λ that dominates the estimation error with large probability. Following the general principle of choosing the penalty introduced in [2–4], we know that the subdifferential of Q_τ evaluated at the point of true coefficient β^* measures the estimation error in the linear regression model. For the L_1-penalized quantile regression, we choose a penalty λ such that it is greater than the maximum absolute value of subdifferential of Q_τ at β^* with high probability. That is, we need to find a penalty level λ for a given constant $c > 1$ and a given small probability α such that

$$P(\lambda \geq c\|S\|_\infty) \geq 1 - \alpha \tag{2.4}$$

where the subdifferential of $Q_\tau(Y - X\beta)$ at the point of true coefficient $\beta = \beta^*$ is specified by $S = X^T w = X^T(w_1, w_2, \cdots, w_n)^T$ with $w_i = 1\{y_i - x_i^T\beta^* \leq 0\} - \tau$. Note that when $\varepsilon_i = y_i - x_i^T\beta^* = 0$ we can take the ith partial subdifferential of Q_τ as $w_i = 1 - \tau$, see, e.g., [33]. Since all ε_i are independently distributed and satisfy $P(\varepsilon_i \leq 0) = \tau$, we have $P(w_i = 1-\tau) = P(\varepsilon_i \leq 0) = \tau$ and $P(w_i = -\tau) = P(\varepsilon_i > 0) = 1-\tau$. Then the distribution of w is known and hence the distribution of $\|S\|_\infty$ is easy to know for any given X, which does not depend on any unknown parameters. Thus, the $1 - \alpha$ quantile of $\|S\|_\infty$ is known for any given X and then the inequality (2.4) holds when we take this $1 - \alpha$ quantile of $\|S\|_\infty$ as λ/c. Note that this penalty is considered in [2]. As in [35], to approximate this value, we employ the following choice of penalty.

$$\lambda = c\sqrt{2A(\alpha)n\log p}, \tag{2.5}$$

where $A(\alpha) > 0$ is a constant with $2p^{1-A(\alpha)} \leq \alpha$.

The following proposition states that the inequality (2.4) holds for the above choice of penalty.

Proposition 2.1. *The choice of penalty $\lambda = c\sqrt{2A(\alpha)n\log p}$ with $A(\alpha) > 0$ and $2p^{1-A(\alpha)} \leq \alpha$ satisfies the inequality (2.4).*

The proof of the above proposition is similar to that in the L_1-penalized LAD case in [35]. The main idea is to bound the tail probability of $X_j^T w$ for $j = 1, 2, \cdots, p$ with the help of

Hoeffding inequality and union bounds. From its proof and taking $A(\alpha) = 2$, we can easily obtain that the special choice of penalty $\lambda = 2c\sqrt{n \log p}$ satisfies the inequality

$$P(\lambda \geq c\|S\|_\infty) \geq 1 - \frac{2}{p}. \tag{2.6}$$

Note that the above result on the choice of penalty for the L_1-penalized quantile regression is independent of τ. Clearly, the above penalties are simple and they do not require any assumptions on matrix X or value of p and n, and the distribution of measurement errors ε_i. As long as all ε_i are independent random variables, the choices satisfy our requirement. This is a big advantage over the traditional Lasso method, which significantly relies on the Gaussian assumption and the variance of the errors. As mentioned in [35], these choices are from union bound and concentration inequalities, and they are not very tight when the sample size n is relatively small. Hence in practice, these penalty levels tend to be relatively large and can cause additional bias to the estimator. From practical point of view, we will need a smaller penalty level if the sample size is not large. For the L_1-penalized quantile regression, we can choose the following refined asymptotic penalty level as in the L_1-penalized LAD, which relies on moment conditions of X and relative size of p and n. This choice could be smaller than the previous ones and it will cause less bias in practice.

Proposition 2.2. *Suppose $\Phi^{-1}(1 - \alpha/(2p)) \leq (q-2)\sqrt{\log n}$, and for some constant $q > 2$*

$$B = \sup_n \sup_{1 \leq j \leq p} \frac{1}{n}\|X_j\|_q^q < \infty.$$

Then the choice of penalty $\lambda = c\sqrt{n}\Phi^{-1}(1 - \alpha/(2p))$ satisfies the inequality

$$P(\lambda \geq c\|S\|_\infty) \geq 1 - \alpha(1 + z_n)$$

where $z_n \to 0 (n \to \infty)$.

3 Near Oracle Property

This section deals with the properties of the L_1-penalized quantile regression estimator $\hat{\beta}$ in (2.3). We will establish the upper bound for estimation error and consider the variable selection properties for both noisy and noiseless cases. For simplicity, in this section we will use $\lambda = 2c\sqrt{n \log p}$ as the default choice of penalty, and we assume that this penalty satisfies $\lambda \geq c\|S\|_\infty$ for some fixed constant $c > 1$. This implies the following important property of the L_1-penalized quantile regression estimator $\hat{\beta}$. Setting $h = \beta^* - \hat{\beta}$ and $\bar{c} = \frac{c-1}{c+1}$, we have $h \in \Delta_{\bar{c}}$ where the restricted set $\Delta_{\bar{c}}$ is specified as

$$\Delta_{\bar{c}} = \{\delta \in \mathbb{R}^p : \|\delta_T\|_1 \geq \bar{c}\|\delta_{T^C}\|_1, \text{where } \mathrm{T} \subset \{1, 2, \cdots, \mathrm{p}\} \text{ and } |\mathrm{T}| \leq \mathrm{s}\}.$$

Here, δ_T denotes the p-dimensional vector such that $(\delta_T)_i = (\delta)_i$ if $i \in T$ and otherwise $(\delta_T)_i = 0$. In fact, it holds by (2.3) and $h = \beta^* - \hat{\beta}$ that

$$Q_\tau(Y - X\hat{\beta}) + \lambda\|\hat{\beta}\|_1 = Q_\tau(Xh + \varepsilon) + \lambda\|\hat{\beta}\|_1 \leq Q_\tau(\varepsilon) + \lambda\|\beta^*\|_1.$$

Let $T = \operatorname{supp}(\beta^*)$. Clearly, $h = h_T + h_{T^C}$ with $h_T = \beta^*_T - \hat{\beta}_T$ and $h_{T^C} = -\hat{\beta}_{T^C}$. Then

$$Q_\tau(Xh + \varepsilon) - Q_\tau(\varepsilon) \leq \lambda(\|h_T\|_1 - \|h_{T^C}\|_1). \tag{3.1}$$

From the convex property of $Q_\tau(Y - X\beta)$ and $X^T w$ is in its subdifferential at the point of $\beta = \beta^*$, we obtain that

$$Q_\tau(Xh + \varepsilon) - Q_\tau(\varepsilon) \geq \langle X^T w, -h\rangle \geq -\|h\|_1 \|X^T w\|_\infty \geq -\frac{\lambda}{c}(\|h_T\|_1 + \|h_{T^C}\|_1).$$

It is easy to derive that $\|h_T\|_1 \geq \bar{c}\|h_{T^C}\|_1$ with $\bar{c} = \frac{c-1}{c+1}$.

It is interesting to mention that the fact $h \in \Delta_{\bar{c}}$ is not only important for L_1-penalized quantile regression but also for the arguments of the classical lasso and the square-root lasso analysis, see, e.g., [2–4].

$\boxed{3.1}$ Conditions on design matrix

In order to study the near oracle property of the L_1-penalized quantile regression estimator, we first introduce some concepts on the design matrix X, which is related to the sparse recovery conditions in the compressed sensing (CS). CS is an interesting and novel area of research with many applications, see the papers by Donoho [18] and Candès, Romberg and Tao [12, 13]. In particular, Candès and Tao [13] introduced a restricted isometry property (RIP) of a sensing/design matrix which guarantees to recover a sparse solution of sparse signal recovery via L_1-norm relaxation. After that, several other sparse recovery conditions were introduced, such as null space properties [16] and s-goodness [23, 24]. For more details, see the recent papers [7, 8, 10] and a new monograph [19].

For simplicity, we follow the description as in [35]. First, we define two important constants λ_s^u and λ_s^l as

$$\lambda_s^u = \sup_{d \in \mathbb{R}^p, 0 < \|d\|_0 \leq s} \frac{\|Xd\|_2^2}{n\|d\|_2^2}, \quad \lambda_s^l = \inf_{d \in \mathbb{R}^p, 0 < \|d\|_0 \leq s} \frac{\|Xd\|_2^2}{n\|d\|_2^2}.$$

Here $\|d\|_0 \leq s$ says that the vector d has at most s nonzero entries, which is also called s-sparse vector. The definition of the above constants is related to the restricted isometry constant (RIC), which is just the maximum value of $\lambda_s^u - 1$ and $1 - \lambda_s^l$. We also need another important concept called the restricted orthogonal constant θ_{s_1, s_2}, which is the smallest number such that for any s_1 and s_2 sparse vectors d_1 and d_2 with disjoint supports

$$|\langle Xd_1, Xd_2\rangle| \leq n\theta_{s_1, s_2}\|d_1\|_2\|d_2\|_2.$$

It is well-known from [12, 18] that for i.i.d. Gaussian random design matrix, i.e. $X_{ij} \sim N(0, 1)$, for any $0 < c < 1$, there exist constants $C_1, C_2 > 0$ such that when $s \leq C_1 n/(\log p)$,

$$P(\max\{\lambda_s^u - 1, 1 - \lambda_s^l\} \leq c) \geq 1 - O(e^{-C_2 p}). \tag{3.2}$$

This means that if $s \log p = o(n)$ for large enough n, λ_s^l will be bounded away from zero by any given constant $c \in (0, 1)$, and λ_s^u be bounded above by any given constant greater than 1 with high probability. Moreover, it holds from the proof of Lemma 12 in [31] that the normalizing constant in our setting will not affect the above results in the case of Gaussian random design.

We below recall the following restricted eigenvalue of design matrix X, see [2, 4, 35] for details. That is,

$$\kappa_s^l(\bar{c}) = \min_{h \in \Delta_{\bar{c}}} \frac{\|Xh\|_1}{n\|h_T\|_2}.$$

To show the near oracle property of the L_1-penalized quantile regression estimator, we should ensure that $\kappa_s^l(\bar{c})$ is bounded away from 0 or approaches 0 slowly enough as in the L_1 penalized LAD case. For simplicity, we will write $\kappa_s^l(\bar{c})$ as κ_s^l whenever there is no confusion.

3.2 Conditions on measurement error

Before giving our main results, we first present some useful lemmas. From (3.1), we obtain that

$$Q_\tau(Xh+\varepsilon) - Q_\tau(\varepsilon) \le \lambda \|h_T\|_1. \tag{3.3}$$

Then we try to bound the estimation error via investigating the random variable $\frac{1}{\sqrt{n}}(Q_\tau(Xh+\varepsilon) - Q_\tau(\varepsilon))$. To do so, we define a random variable $B(d)$ for any vector $d \in \mathbb{R}^p$ as

$$B(d) = \frac{1}{\sqrt{n}} |(Q_\tau(Xd+\varepsilon) - Q_\tau(\varepsilon)) - E(Q_\tau(Xd+\varepsilon) - Q_\tau(\varepsilon))|.$$

Then we have the following important property of $B(d)$, which is very useful in the proof of our main result.

Lemma 3.1. *Suppose all ε_i are independent random variables. For $p > n$ and $p > 3\sqrt{s}$, we have*

$$P\left(\sup_{\|d\|_0 = s, \|d\|_2 = 1} B(d) \ge (1 + 2C_2\sqrt{\lambda_s^u})\sqrt{2s\log p} \right) \le 2p^{-4s(C_2^2-1)} \tag{3.4}$$

where $C_2 > 1$ is a constant.

This lemma tells us that with probability at least $1 - 2p^{-4s(C_2^2-1)}$, for any s sparse vector $d \in \mathbb{R}^p$,

$$\frac{1}{\sqrt{n}} |(Q_\tau(Xd+\varepsilon) - Q_\tau(\varepsilon)) - E(Q_\tau(Xd+\varepsilon) - Q_\tau(\varepsilon))| \le (1 + 2C_2\sqrt{\lambda_s^u})\sqrt{2s\log p}\|d\|_2. \tag{3.5}$$

This means that the value of the random variable $\frac{1}{\sqrt{n}}(Q_\tau(Xh+\varepsilon) - Q_\tau(\varepsilon))$ is very close to its expectation with high probability. Clearly, the expectation is not random and much easier to analyze than the random variable itself. We now consider the properties of $E(Q_\tau(Xh+\varepsilon) - Q_\tau(\varepsilon))$, or its component $E(\rho_\tau(\varepsilon_i + x) - \rho_\tau(\varepsilon_i))$.

Lemma 3.2. *For any continuous random variable ε_i, the following equation holds*

$$\frac{dE(\rho_\tau(\varepsilon_i + x) - \rho_\tau(\varepsilon_i))}{dx} = \tau - P(\varepsilon_i \le -x). \tag{3.6}$$

In order to give the lower bound of $E(\rho_\tau(\varepsilon_i + x) - \rho_\tau(\varepsilon_i))$, we need some conditions on the measurement errors ε_i. For the error distribution of ε_i, we assume there is a positive constant $a > 0$ such that

$$\begin{aligned} P(\varepsilon_i \ge x) \quad &\le \quad (1-\tau)\frac{1}{1+ax} \quad \text{for all x} \ge 0 \\ P(\varepsilon_i \le x) \quad &\le \quad \tau\frac{1}{1+a|x|} \quad \text{for all x} < 0. \end{aligned} \tag{3.7}$$

The above is the scale assumption on the measurement errors with a scale parameter a, which is from Wang [35]. This is a very weak condition and even Cauchy distribution satisfies it. Based on this assumption, we give the following lemma. Here $a \wedge b := \min\{a, b\}$ for $a, b \in \mathbb{R}$.

Lemma 3.3. *Assume the random variable ε_i satisfies the scale assumption (3.7). For any given real number $x \in \mathbb{R}$, the following inequality holds*

$$E(\rho_\tau(\varepsilon_i + x) - \rho_\tau(\varepsilon_i)) \ge \frac{a}{8}(\tau \wedge (1-\tau))|x|(|x| \wedge \frac{3}{a}). \tag{3.8}$$

For simplicity in subsequent analysis, we just employ the above weak bound, which clearly can be improved.

3.3 Main results

We will present our main result based on the following conditions.

$$(C1): \qquad \lambda_s^l > \theta_{s,s}(\frac{1}{\bar{c}}+\frac{1}{4}),$$

$$(C2): \qquad \frac{3\sqrt{n}}{16}(\tau\wedge(1-\tau))\kappa_s^l > \lambda\sqrt{s/n}+C_1\sqrt{2s\log p}(1.25+1/\bar{c}).$$

Theorem 3.4. *Consider the high-dimensional linear regression model (2.1). Assume random variables $\varepsilon_i (i=1,2,\cdots,n)$ are independent and distribution satisfying (3.7), and conditions (C1) and (C2) hold. Then with probability at least $1-2p^{-4s(C_2^2-1)+1}$, the L_1-penalized quantile regression estimator $\hat{\beta}$ satisfies*

$$\|\hat{\beta}-\beta^*\|_2 \le C\sqrt{\frac{2s\log p}{n}},$$

where $C = \frac{8\lambda_s^u(c\sqrt{2}+1.25C_1+C_1/\bar{c})}{a(\tau\wedge(1-\tau))\big(\lambda_s^l-\theta_{s,s}\big(\frac{1}{\bar{c}}+\frac{1}{4}\big)\big)^2}\sqrt{1+\frac{1}{\bar{c}}}$ *with constants* $C_1 = 1+2C_2\sqrt{\lambda_s^u}$ *and* $C_2 > 1$.

From the above theorem we easily obtain that with high probability,

$$\|\hat{\beta}-\beta^*\|_2 = O(\sqrt{\frac{s\log p}{n}}). \tag{3.9}$$

This claims that asymptotically, the L_1-penalized quantile regression estimator has near oracle performance and hence it matches the asymptotic performance of the Lasso method with a known variance.

Remark Actually, $\theta_{s,s}$ can be bounded by λ_s^l and λ_s^u and the condition (C1) can be replaced by some similar RIP conditions; see for example [8, 10]. We keep it here just to simplify the arguments. Condition (C2) states that the columns of X cannot be too sparse. Otherwise, if the columns of X are sparse, then the L_1 norm of columns of X will be small. Thus, the value κ_s^l will be small.

It is not hard to know from the above theorem that the L_1-penalized quantile regression estimator will select most of the significant variables with high probability. This implies that the L_1-penalized quantile regression method will select a model that contains all the variables with large coefficients of β^*. Thus, if all the nonzero coefficients are large enough in terms of absolute value, then the L_1-penalized quantile regression method can select all of them into the model. We summarize it as the following theorem.

Theorem 3.5. *Consider the high-dimensional linear regression model (2.1). Let $\hat{\beta}$ be the L_1-penalized quantile regression estimator and $\bar{T} = \mathrm{supp}(\hat{\beta})$. Then under the same conditions as in Theorem 3.4, with probability at least $1-2p^{-4s(C_2^2-1)+1}$, the support of the L_1-penalized quantile regression estimator $\hat{\beta}$ satisfies*

$$\left\{ i : |\beta_i^*| \ge C\sqrt{\frac{2s\log p}{n}} \right\} \subset \bar{T},$$

where C is specified in Theorem 3.4.

When there is no noise in the high dimensional linear regression model, i.e., $Y = X\beta$, we can show that the L_1-penalized quantile regression estimator still has a nice variable selection property.

Theorem 3.6. *In the noiseless case for the high dimensional linear regression model (2.1), if we take a penalty level λ such that $\lambda < (\tau \wedge (1-\tau))\frac{n\kappa_s^l(1)}{s}$, then L_1-penalized quantile regression estimator satisfies $\hat{\beta} = \beta^*$.*

Suppose $\kappa_s^l(1)$ are bounded away from 0 for all n and the penalty level λ is specified by $2\sqrt{n \log p}$. From the above theorem, if $\sqrt{\log p} = o(n)$ and n is large enough, then the L_1-penalized quantile regression estimator satisfies $\hat{\beta} = \beta^*$.

4 Numerical Study

In this section, we will use an alternating direction method of multipliers (ADMM) to solve (2.2), and present some numerical experiments to demonstrate its efficacy. In the simulation we consider the linear models $Y = X\beta + \varepsilon$ with sample matrix $X = (x_1, x_2, \cdots, x_n)^T \in \mathbb{R}^{n\times p}$ being from Gaussian matrices, and the noise contents $\mathbb{E}(\varepsilon) = 0$, $\mathrm{Var}(\varepsilon) = \sigma^2$.

4.1 Alternating direction method of Multipliers

We investigate the proximal mapping of function ρ_τ first. Note that $\rho_\tau(u) = u(\tau - 1\{u \le 0\})$. It is easy to show that for given $\alpha > 0$ and $\xi \in \mathbb{R}$, the optimal solution of the following problem

$$\min_u \quad \rho_\tau(u) + \frac{\alpha}{2}\|u - \xi\|^2, \tag{4.1}$$

i.e., the proximal mapping of ρ_τ, is given by

$$Prox_{\rho_\tau}(\xi, \alpha) = \begin{cases} \xi - \frac{\tau}{\alpha}, & \text{if } \xi > \frac{\tau}{\alpha} \\ \xi + \frac{1-\tau}{\alpha}, & \text{if } \xi \le -\frac{1-\tau}{\alpha} \\ 0, & \text{if } -\frac{1-\tau}{\alpha} < \xi \le \frac{\tau}{\alpha}. \end{cases} \tag{4.2}$$

By introducing a new variable $\gamma := Y - X\beta$, (2.2) can be rewritten as

$$\begin{array}{ll} \min_{\beta,\gamma} & Q_\tau(\gamma) + \lambda\|\beta\|_1 \\ s.t. & X\beta + \gamma - Y = 0. \end{array} \tag{4.3}$$

The augmented Lagrangian function for (4.3) is

$$\mathcal{L}_\mu(\beta, \gamma; \theta) = Q_\tau(\gamma) + \lambda\|\beta\|_1 - \langle \theta, X\beta + \gamma - Y\rangle + \frac{1}{2\mu}\|X\beta + \gamma - Y\|^2. \tag{4.4}$$

A typical iteration of ADMM for solving (4.3) is

$$\begin{cases} \beta^{k+1} & := & \mathrm{argmin}_\beta \quad \mathcal{L}_\mu(\beta, \gamma^k; \theta^k) \\ \gamma^{k+1} & := & \mathrm{argmin}_\gamma \quad \mathcal{L}_\mu(\beta^{k+1}, \gamma; \theta^k) \\ \theta^{k+1} & := & \theta_k - (X\beta^{k+1} + \gamma^{k+1} - Y)/\mu. \end{cases} \tag{4.5}$$

The solution to the second subproblem of (4.5) corresponds to the proximal mapping of function ρ_τ. More specifically, the solution γ^{k+1} is given by

$$\gamma_i^{k+1} = Prox_{\rho_\tau}((Y + \mu\theta^k - X\beta^{k+1})_i, 1/\mu), i = 1, \ldots, n. \tag{4.6}$$

The solution to the first subproblem of (4.5) actually is not in closed-form. We thus modify the first subproblem by a proximal gradient step, which leads to the following subproblem instead:

$$\beta^{k+1} := \mathrm{argmin}_\beta \quad \lambda\|\beta\|_1 + \frac{1}{2\nu}\|\beta - (\beta^k - \frac{\nu}{\mu}X^\top(X\beta^k + \gamma^k - Y - \mu\theta^k))\|^2, \tag{4.7}$$

where $\nu < 1/\lambda_{\max}(X^\top X)$ is the step size for the proximal gradient step. The optimal solution of (4.7) is given by the ℓ_1 shrinkage operation:

$$\beta^{k+1} := \text{Shrink}(\beta^k - \frac{\nu}{\mu}X^\top(X\beta^k + \gamma^k - Y - \mu\theta^k), \lambda\nu). \tag{4.8}$$

To summarize, the (linearized) ADMM for solving (4.3) is given by

$$\left\{\begin{array}{lll} \beta^{k+1} & := & \text{Shrink}(\beta^k - \frac{\nu}{\mu}X^\top(X\beta^k + \gamma^k - Y - \mu\theta^k), \lambda\nu) \\ \gamma_i^{k+1} & := & Prox_{\rho_\tau}((Y + \mu\theta^k - X\beta^{k+1})_i, 1/\mu), i = 1, \ldots, n \\ \theta^{k+1} & := & \theta_k - (X\beta^{k+1} + \gamma^{k+1} - Y)/\mu. \end{array}\right. \tag{4.9}$$

Theorem 4.1. *For given $\mu > 0$, and $\nu < 1/\lambda_{\max}(X^\top X)$, the sequence $\{(\beta^k, \gamma^k, \theta^k)\}$ generated by the above ADMM from any starting point converges to $(\beta^*, \gamma^*, \theta^*)$, where (β^*, γ^*) is a solution of* (4.3).

It is a typical alternating direction method of multipliers for minimizing the sum of two convex functions, so we omit the proof for simplicity. See, e.g., [5, 17, 28, 37, 38] for more details.

$\boxed{4.2}$ Example I: Simulated Experiment

Firstly, by considering synthetic data sets, we randomly generate 100 stimulations with noise $\xi \sim N(0, \sigma^2)$ (the Normal distribution) and $\xi \sim LN(0, \sigma^2)$ (Log-Normal distribution which is one of heavy-tailed distributions) respectively. We design the true regression quantiles estimator β^* from the generated measurements X. By exploiting ADMM to compute the approximately optimal solution, we denote it as β. The comparison of the estimation error $\|\beta - \beta^*\|_2^2$, the prediction error $\frac{1}{n}\|X\beta - X\beta^*\|_2^2$ and the average CPU time of the two different noise distributions will show the performance of our approach.

For each data set, the random matrix X, observations vector Y and the true regression quantiles estimator β^* are generated by the following MATLAB codes:

$$\begin{array}{ll} s = round(p * 0.05), & b = randperm(p), \\ \beta^* = zeros(p, 1), & \beta^*(b(1:s)) = randn(s, 1), \\ X = randn(m, n), & \\ Y1 = X\beta^* + \sigma * randn(0, 1), & \\ Y2 = X\beta^* + \sigma * lognrnd(0, 1, n, 1), & \end{array}$$

where the sparsity s of the true regression quantiles estimator β^* is always settled as $s = 5\% \times p$. The parameter σ will be taken as $\sigma = 0.01$ or 0.25. Other related parameters are given as $\mu = 1, \nu = 1/\lambda_{\max}\left(X^T X\right)$ and the stopping criteria $\|X\beta^k + \gamma^k - Y\|_2^2 \le 10^{-6}$.

From Table 1, one can check that $\lambda_1 \approx 0.05\sqrt{n \log p}$ and $\lambda_2 \approx 0.1\sqrt{n \log p}$. Compared with the error and CPU time from the noise of Log-Normal distribution, those of Normal distributed noise are relatively lower regardless of the parameters λ and τ. Evidently for any of these parameters, the estimation error and prediction error from Normal noise are identical, with 2×10^{-4} and 1×10^{-4} respectively, which means the ADMM approach to estimate the regression quantiles performs quite stably under such noise circumstance. Moreover, in terms of the data of Log-Normal noise, there is a slight upward trend of the estimation error and prediction error when the τ is increasing.

Next, by altering the dimension p with $n = p/2$ and $\lambda = 0.1\sqrt{n \log p}$, we implement 100 replicators for each data set to generate the estimation error, prediction error and CPU

Table 1: The average of estimation error $\|\beta - \beta^*\|_2^2$, the prediction error $\frac{1}{n}\|X\beta - X\beta^*\|_2^2$ and CPU time over 100 simulations under different penalty levels λ, quantiles τ and two error distributions. Here, the dimension $n = 500, p = 1000, \lambda = O(\sqrt{n \log p})$ and $\sigma = 0.01$, the values outside of '$(\cdot)$' stand for the data from noise $N(0, \sigma^2)$ and inside values stand for results from noise $LN(0, \sigma^2)$.

	τ	$\|\beta - \beta^*\|_2^2$	$\frac{1}{n}\|X\beta - X\beta^*\|_2^2$	CPU time
$\lambda_1 = 2$	0.1	2e-04 (3e-04)	1e-04 (2e-04)	15.15 (14.69)
	0.3	2e-04 (15e-04)	1e-04 (6e-04)	15.21 (15.92)
	0.5	2e-04 (13e-04)	1e-04 (6e-04)	15.08 (16.38)
	0.7	2e-04 (11e-04)	1e-04 (7e-04)	18.00 (18.80)
	0.9	2e-04 (15e-04)	1e-04 (8e-04)	14.06 (16.76)
$\lambda_2 = 5$	0.1	2e-04 (1e-04)	1e-04 (1e-04)	14.09 (12.63)
	0.3	2e-04 (4e-04)	1e-04 (3e-04)	12.89 (11.86)
	0.5	2e-04 (7e-04)	1e-04 (4e-04)	14.29 (12.80)
	0.7	2e-04 (10e-04)	1e-04 (5e-04)	13.57 (15.25)
	0.9	2e-04 (18e-04)	1e-04 (12e-04)	14.38 (16.28)

time from two noise patterns. From Figure 1, the two types of the error stemmed from Normal noise are basically smaller than those from Log-Normal noise, although there is not an apparent distinction of the CPU time between these two noise, see Figure 2. To be more exact, for the information of Normal noise, when the τ centering at the proximity of 0.5 the errors tend to be lower than $\tau = 0.1$ and $\tau = 0.9$. The phenomenon probably can be illustrated by the fact that the linear least absolutely deviation, where $\tau = 0.5$, has the advantage of estimating the linear regression. By contrast, referring to the data from the Log-Normal noise, both of the estimation error and prediction error are ascending with the increasing of τ, which is mainly from the fact that the Log-Normal distribution is non-centrally symmetrical and has a heavy tail. In addition, one can see that the error has not been enlarged with the p becoming larger, which manifests ADMM approach performs robustly indeed.

$\boxed{4.3}$ Example II: Toeplitz Correlation Matrix

The second modified example is from literature [35] which aims at considering the estimator $\beta^* = (3, 3, \cdots, 3,$
$3, 0, \cdots, 0)^T$ with s (being taken $s = 2.5\%p$ in this example) 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., $x_i \sim \Sigma^{1/2}N(0, 1), \ \ i = 1, 2, \cdots, n$; and then normalize the columns of X such that each column has L_2 norm $\sqrt{n}$. We use three noise patterns: (a) $N(0, \sigma^2)$ noise, (b) $LN(0, \sigma^2)$ noise, and (c) Student distribution $T(2)$ noise which belongs to the heavy two-tailed distribution. Corresponding MATLAB codes are as below:

$$
\begin{aligned}
Y1 &= X\beta^* + \sigma * randn(0, 1), \\
Y2 &= X\beta^* + \sigma * lognrnd(0, 1, n, 1), \\
Y3 &= X\beta^* + \sigma * trnd(2, n, 1).
\end{aligned}
$$

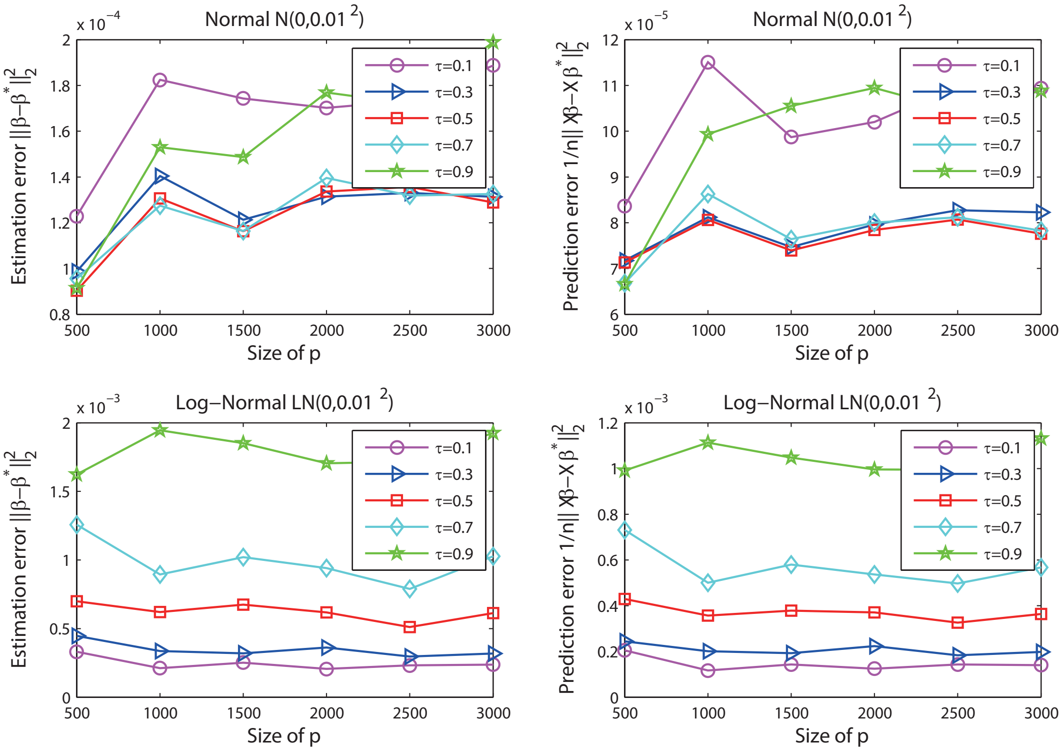

Figure 1: Average error from Normal and Log-Normal noise yielded by ADMM with $\lambda = 0.1\sqrt{n\log p}$.

Apart from the average L_2 norm square of the estimation errors $\|\beta-\beta^*\|_2^2$ and the prediction errors $\frac{1}{n}\|X\beta - X\beta^*\|_2^2$, we add to consider another two types of errors which are

$$\text{FPR} := \frac{\text{Card}\ \{\ j:\ \beta_j^* \neq 0\ \&\ \beta_j = 0\ \}}{\text{Card}\ \{\ j: \beta_j = 0\ \}}, \qquad \text{TPR} := \frac{\text{Card}\ \{\ j:\ \beta_j^* \neq 0\ \&\ \beta_j \neq 0\ \}}{\text{Card}\ \{\ j: \beta_j \neq 0\ \}},$$

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 ture 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 are, the better our approach would perform.

It can be seen from Table 2, the most obvious property is all the FPRs are equal to zero, which signifies all the significant variables are selected by the method. Since the noise is enhanced (from $\sigma = 0.01$ to $\sigma = 0.25$), the prediction errors increase dramatically compared with the corresponding data from Table 1. Corresponding properties can be shown in the following figures.

Moreover, one can discern from Figure 3, data from Normal distributed noise and Student $T(2)$ distributed noise have the symmetrical patterns primarily due to the symmetry of these two distributions, and because of this, the estimation and prediction errors reach the bottom while TRPs hit the peak when $\tau = 0.5$. By contrast, the errors stemmed from Log-Normal distribution have a monotonicity, which results in the estimation and prediction errors arriving the lowest points, and TPRs reaching the highest point when $\tau = 0.1$.

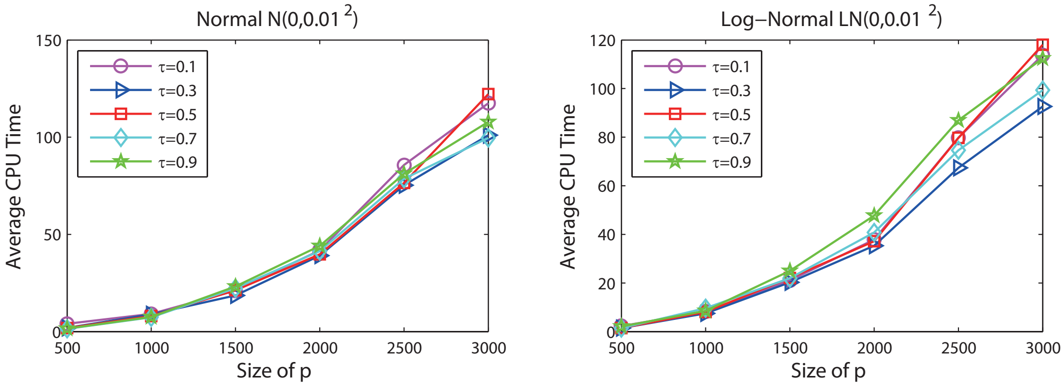

Figure 2: Average CPU Time from Normal noise and Log-Normal noise yielded by ADMM with $\lambda = 0.1\sqrt{n \log p}$.

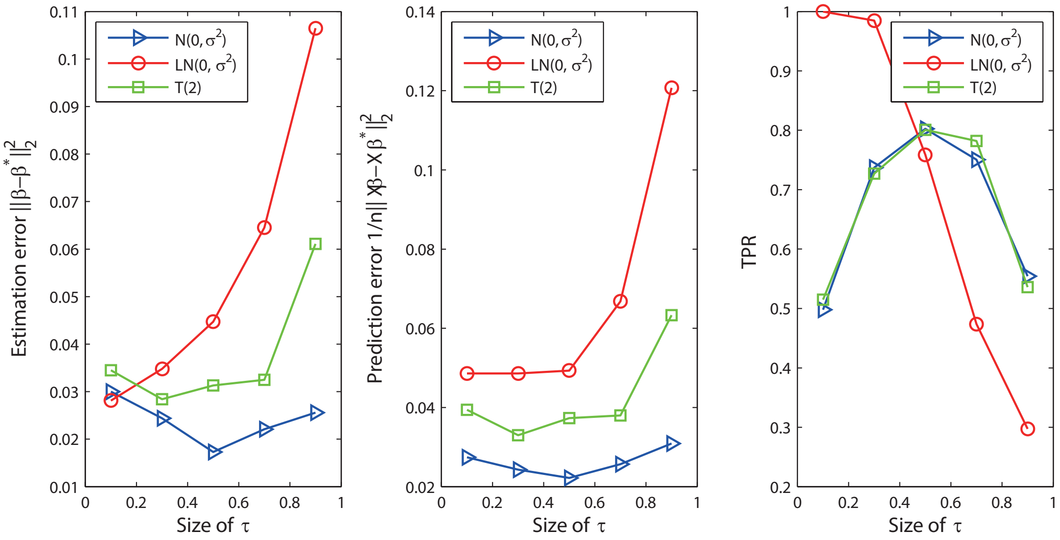

Figure 3: Average errors from three types of noise over 100 simulations yielded by ADMM with $n = 500, p = 1000, s = 25, \lambda = 0.5\sqrt{n \log(p)}$ and $\sigma = 0.25$.

$\boxed{5}$ Concluding remarks

This paper focuses on the L_1-penalized quantile regression estimator for the high-dimensional linear regression model for Traditional Chinese medicine syndrome manifestation. We have shown the near oracle properties of the L_1-penalized quantile regression estimator and established the upper bound for estimation error and considered the variable selection properties for both noisy and noiseless cases under weak assumptions, where we do not have any assumptions on the moments of the noise and we only need a scale parameter to control the tail probability of the noise. These results are true for a wide range of noise distributions, even for Cauchy distributed noise.

Moreover, we have proposed an alternating direction method to find the L_1-penalized

Table 2: The average of estimation error $\|\beta-\beta^*\|_2^2$, the prediction error $\frac{1}{n}\|X\beta-X\beta^*\|_2^2$, FPR and TPR over 100 simulations under several τ and three error distributions with $\lambda=0.5\sqrt{n\log(p)}$, where $n=500, p=1000, s=25$ and $\sigma=0.25$.

τ	Noise	$\|\beta-\beta^*\|_2^2$	$\frac{1}{n}\|X\beta-X\beta^*\|_2^2$	FPR	TPR
$\tau_1=0.1$	$N(0,\sigma^2)$	0.0300	0.0274	0	49.78%
	$LN(0,\sigma^2)$	0.0281	0.0486	0	100.0%
	$T(2)$	0.0345	0.0394	0	51.50%
$\tau_1=0.3$	$N(0,\sigma^2)$	0.0244	0.0243	0	73.71%
	$LN(0,\sigma^2)$	0.0348	0.0486	0	98.46%
	$T(2)$	0.0284	0.0330	0	72.70%
$\tau_1=0.5$	$N(0,\sigma^2)$	0.0173	0.0222	0	80.24%
	$LN(0,\sigma^2)$	0.0447	0.0493	0	75.88%
	$T(2)$	0.0313	0.0373	0	80.07%
$\tau_1=0.7$	$N(0,\sigma^2)$	0.0221	0.0257	0	75.06%
	$LN(0,\sigma^2)$	0.0645	0.0668	0	47.37%
	$T(2)$	0.0325	0.0380	0	78.19%
$\tau_1=0.9$	$N(0,\sigma^2)$	0.0256	0.0309	0	55.43%
	$LN(0,\sigma^2)$	0.1065	0.1207	0	29.72%
	$T(2)$	0.0611	0.0633	0	53.61%

quantile regression estimator and reported the numerical results to demonstrate the efficacy of our method. The proposed method is different from the common computational methods for quantile regression estimation [15, 25], and our method can deal with the big data and more efficient as shown in numerical experiments.

6 Appendix

Proof of Proposition 2.1. We will prove Proposition 2.1 by applying union bound and Hoeffding's inequality. It holds by the union bound

$$P(c\sqrt{2A(\alpha)n\log p}\leq c\|S\|_\infty)\leq\sum_{i=1}^{p}P(\sqrt{2A(\alpha)n\log p}\leq|X_i^Tw|).$$

It is easy to see and the value of X_i^Tw in the $[-\max\{\tau,1-\tau\}\|X_i\|_2,\max\{\tau,1-\tau\}\|X_i\|_2]$. For each i, it holds by Hoeffding's inequality

$$P(\sqrt{2A(\alpha)n\log p}\leq|X_i^Tw|)\leq 2\exp(-\frac{4A(\alpha)n\log p}{4(\max\{\tau,1-\tau\}\|X_i\|_2)^2}).$$

Since $\max\{\tau,1-\tau\}\leq 1$ and $\|X_i\|_2^2=n$, we have $\exp(-\frac{4A(\alpha)n\log p}{4(\max\{\tau,1-\tau\}\|X_i\|_2)^2})\leq\exp(-\frac{4A(\alpha)n\log p}{4n})=p^{-A(\alpha)}$. Thus, we obtain

$$P(c\sqrt{2A(\alpha)n\log p}\leq c\|S\|_\infty)\leq\sum_{i=1}^{p}2p^{-A(\alpha)}\leq\alpha.$$

The desired result follows immediately. □

Proof of Proposition 2.2. As in the proof of Proposition 2.1, it holds by the union bound

$$P(c\sqrt{2A(\alpha)n\log p} \leq c\|S\|_\infty) \leq \sum_{i=1}^{p} P(\sqrt{2A(\alpha)n\log p} \leq |X_i^T w|).$$

For each i, from the assumption and Corollary 1 in [35], we obtain

$$P(c\sqrt{n}\Phi^{-1}(1-\frac{\alpha}{2p})) \leq 2(1-\Phi(\Phi^{-1}(1-\frac{\alpha}{2p})))(1+z_n) = \frac{\alpha}{p}(1+z_n),$$

where $z_n \to 0$ as $n \to \infty$. Therefore, we get

$$P(c\sqrt{2A(\alpha)n\log p} \leq c\|S\|_\infty) \leq \alpha(1+z_n).$$

□

fProof of Lemma 3.1. The idea of this proof is similar as that of Lemma 3 in [35]. First, it is easy to show that for real numbers $a, b \in \mathbb{R}$,

$$|\rho_\tau(a) - \rho_\tau(b)| \leq |\max\{\tau, \tau-1\}(a-b)| \leq |a-b|. \tag{6.1}$$

Thus, $|\rho_\tau(x_i^T d+\varepsilon_i) - \rho_\tau(\varepsilon_i)| \leq |x_i^T d| \leq \|Xd\|_2$ for any $i = 1, 2, \cdots, n$. This means $|\rho_\tau(x_i^T d+\varepsilon_i) - \rho_\tau(\varepsilon_i)|$ is a bounded random variable for any given $d \in \mathbb{R}^p$. Hence for any fixed s sparse vector d, by Hoeffding's inequality, we have for any $t > 0$,

$$P(B(d) \geq t) \leq 2\exp\{-\frac{nt^2}{2\|Xd\|_2^2}\}.$$

From the definition of λ_s^u, we easily get

$$P(B(d) \geq t) \leq 2\exp\{-\frac{t^2}{2\lambda_s^u\|d\|_2^2}\}.$$

Taking $t = C\sqrt{2s\log p}\|d\|_2$, we obtain that for all $C > 0$,

$$P(B(d) \geq C\sqrt{2s\log p}\|d\|_2) \leq 2p^{-sC^2/\lambda_s^u}.$$

Below we will establish an upper bound for $\sup_{\|d\|_0=s,\|d\|_2=1} B(d)$. We will apply the ϵ-net method and covering number result. Consider the ϵ-net of the set $\{d \in \mathbb{R}^p : \|d\|_0 = s, \|d\|_2 = 1\}$. From the standard results of covering number, we know that the covering number of $\{d \in \mathbb{R}^p : \|d\|_2 = 1\}$ by ϵ-balls (i.e. $\{y \in \mathbb{R}^s : \|y-x\|_2 \leq \epsilon\}$) is at most $(3/\varepsilon)^s$ for $\varepsilon < 1$. Then the covering number of $\{d \in \mathbb{R}^p : \|d\|_0 = s, \|d\|_2 = 1\}$ by ϵ-balls is at most $(3p/\epsilon)^s$ for $\epsilon < 1$. Suppose N is such a ϵ-net of $\{d \in \mathbb{R}^p : \|d\|_0 = s, \|d\|_2 = 1\}$. It holds by union bound,

$$P(B(d)_{d\in N} \geq C\sqrt{2s\log p}) \leq 2(3p/\epsilon)^s p^{-sC^2/\lambda_s^u}. \tag{6.2}$$

By (6.1), it holds

$$\begin{aligned} & \sup_{\|d_1-d_2\|_0\leq s, \|d_1-d_2\|_2\leq\epsilon} |B(d_1) - B(d_2)| \\ \leq\ & \frac{2}{\sqrt{n}}\|X(d_1-d_2)\|_1 \\ \leq\ & \frac{2}{\sqrt{n}}\max_i\{\|X_i\|_2\}\|d_1-d_2\|_1 \\ \leq\ & 2\sqrt{ns}\epsilon. \end{aligned}$$

Then we get

$$\sup_{\|d\|_0=s,\|d\|_2=1} B(d) \le \sup_{d\in N} B(d) + 2\sqrt{ns}\epsilon.$$

Setting $\epsilon = \frac{\sqrt{2s\log p}}{2\sqrt{ns}}$, we obtain from (6.2) that

$$\begin{aligned} & P\left(\sup_{\|d\|_0=s,\|d\|_2=1} B(d) \ge C\sqrt{2s\log p}\right) \\ \le\; & P\left(\sup_{d\in N} B(d) \ge (C-1)\sqrt{2s\log p}\right) \\ \le\; & 2(3p/\epsilon)^s p^{-s(C-1)^2/\lambda_s^u} \le 2(\frac{3p\sqrt{ns}}{p^{(C-1)^2/\lambda_s^u}})^s. \end{aligned}$$

Choosing $C = 1 + 2C_2\sqrt{\lambda_s^u}$ with $C_2 > 1$, from the assumption, we obtain that

$$P\left(\sup_{\|d\|_0=s,\|d\|_2=1} B(d) \ge (1+2C_2\sqrt{\lambda_s^u})\sqrt{2s\log p}\right) \le 2p^{-4s(C_2^2-1)}.$$

The proof is completed. □

Proof of Lemma 3.2. As in the proof of Lemma 3.1, $|\rho_\tau(x+\varepsilon_i) - \rho_\tau(\varepsilon_i)| \le |x|$. Then the random variable $|\rho_\tau(x+\varepsilon_i) - \rho_\tau(\varepsilon_i)|$ is bounded, and hence the expectation always exists. Suppose $f(t)$ is the density function of ε_i. Note that $E(\rho_\tau(\varepsilon_i))$ is a constant with respect to x. We easily derive

$$\begin{aligned} E(\rho_\tau(\varepsilon_i+x) - \rho_\tau(\varepsilon_i)) &= E(\rho_\tau(\varepsilon_i+x)) - E(\rho_\tau(\varepsilon_i)) \\ &= \int_{-\infty}^{-x} (\tau-1)(t+x)f(t)dt + \int_{-x}^{\infty} \tau(t+x)f(t)dt - E(\rho_\tau(\varepsilon_i)) \\ &= \tau x - xP(\varepsilon_i \le -x) - \int_{-\infty}^{-x} tf(t)dt + \tau - E(\rho_\tau(\varepsilon_i)). \end{aligned}$$

Thus, direct calculation yields

$$\frac{dE(\rho_\tau(\varepsilon_i+x) - \rho_\tau(\varepsilon_i))}{dx} = \tau - P(\varepsilon_i \le -x).$$

□

fProof of Lemma 3.3. From the scale assumption and Lemma 3.2, we obtain that for $x > 0$,

$$\begin{aligned} E(\rho_\tau(\varepsilon_i+x) - \rho_\tau(\varepsilon_i)) &= \int_0^x (\tau - P(\varepsilon_i \le -t))dt \\ &\ge \tau x - \int_0^x \tau\frac{1}{1+at}dt \\ &= \tau(x - \frac{1}{a}\log(1+ax)). \end{aligned}$$

Similarly, for $x < 0$, we have

$$\begin{aligned} E(\rho_\tau(\varepsilon_i + x) - \rho_\tau(\varepsilon_i)) &= \int_0^x (\tau - P(\varepsilon_i \leq -t))dt \\ &= (\tau - 1)x + \int_0^x P(\varepsilon_i \geq -t)dt \\ &\geq \int_0^x \tau \frac{1}{1+at}dt \\ &= (\tau - 1)(x + \frac{1}{a}\log(1 - ax)) \\ &= (1 - \tau)(-x - \frac{1}{a}\log(1 + a(-x))). \end{aligned}$$

Combining the above arguments, we obtain that

$$E(\rho_\tau(\varepsilon_i + x) - \rho_\tau(\varepsilon_i)) \geq (\tau \wedge (1 - \tau))(|x| - \frac{1}{a}\log(1 + a|x|)). \tag{6.3}$$

We can easily obtain that when $|x| \geq 3/a$,

$$|x| - \frac{1}{a}\log(1 + a|x|) \geq |x| - \frac{1}{a}\frac{a|x|}{2} = \frac{|x|}{2},$$

and when $|x| \geq 3/a$,

$$|x| - \frac{1}{a}\log(1 + a|x|) \geq |x| - \frac{1}{a}\left(\frac{a|x|}{2} - \frac{1}{8}(a|x|)^2\right) = \frac{ax^2}{8}.$$

This together with (6.3) completes the proof. □

Proof of Theorem 3.4. Consider the high-dimensional linear regression model (2.1), and let $\hat{\beta}$ be its L_1-penalized quantile regression estimator and β^* the true s-sparse coefficient. Note that $h = \beta^* - \hat{\beta}$ and $h \in \Delta_{\bar{c}}$. Without loss of generality, let the entries of h satisfy $|h_1| \geq |h_2| \geq \cdots \geq |h_p|$. We partition the index set $\{1, 2, \cdots, p\}$ into the following subsets

$$\begin{aligned} &S_0 = \{1, 2, \cdots, s\}, S_1 = \{s + 1, s + 2, \cdots, 2s\}, \cdots, S_{\lfloor (p-1)/s \rfloor} \\ &\quad = \{\lfloor (p - 1)/s \rfloor s + 1, \lfloor (p - 1)/s \rfloor s + 2, \cdots, p\}. \end{aligned}$$

For brevity, let $S_0^C = \bigcup_{i \geq 1} S_i$. Then it follows from inequality (6) in [9] (or Lemma 8 in [35]) that

$$\begin{aligned} \sum_{i \geq 1} \|h_{S_i}\|_2 &\leq \sum_{i \geq 1} \frac{\|h_{S_i}\|_1}{\sqrt{s}} + \frac{\sqrt{s}}{4}|h_{s+1}| \\ &\leq \frac{\|h_{S_0^C}\|_1}{\sqrt{s}} + \frac{1}{4\sqrt{s}}\|h_{S_0}\|_1 \\ &\leq \left(\frac{1}{\sqrt{s}\bar{c}} + \frac{1}{4\sqrt{s}}\right)\|h_{S_0}\|_1 \leq \left(\frac{1}{\bar{c}} + \frac{1}{4}\right)\|h_{S_0}\|_2. \end{aligned} \tag{6.4}$$

Moreover, it is easy to derive that

$$\begin{aligned} \frac{1}{\sqrt{n}}(Q_\tau(Xh + \varepsilon) - Q_\tau(\varepsilon)) \geq\ & \frac{1}{\sqrt{n}}(Q_\tau(Xh_{S_0} + \varepsilon) - Q_\tau(\varepsilon)) \\ & + \sum_{i \geq 1} \frac{1}{\sqrt{n}}\left(Q_\tau(X\sum_{j=0}^{i} h_{S_j} + \varepsilon) - Q_\tau(X\sum_{j=0}^{i-1} h_{S_j} + \varepsilon)\right). \end{aligned} \tag{6.5}$$

For simplicity, we define

$$M(d) = \frac{1}{\sqrt{n}} E(Q_\tau(Xd + \varepsilon) - Q_\tau(\varepsilon))$$

for any given $d \in \mathbb{R}^p$. It holds by Lemma 3.1 that with probability at least $1 - 2p^{-4s(C_2^2-1)}$,

$$\frac{1}{\sqrt{n}}(Q_\tau(Xh_{S_0} + \varepsilon) - Q_\tau(\varepsilon)) \geq M(h_{S_0}) - C_1\sqrt{2s\log p}\|h_{S_0}\|_2,$$

where $C_1 = 1+2C_2\sqrt{\lambda_s^u}$ with $C_2 > 1$. Likewise, we obtain that for any $i \geq 1$ with probability at least $1 - 2p^{-4s(C_2^2-1)}$,

$$\frac{1}{\sqrt{n}}(Q_\tau(X\sum_{j=0}^{i} h_{S_j} + \varepsilon) - Q_\tau(X\sum_{j=0}^{i-1} h_{S_j} + \varepsilon)) \geq M(h_{S_i}) - C_1\sqrt{2s\log p}\|h_{S_i}\|_2.$$

Combining the above arguments, we obtain from (6.5) that with probability at least $1 - 2p^{-4s(C_2^2-1)+1}$,

$$\frac{1}{\sqrt{n}}(Q_\tau(Xh + \varepsilon) - Q_\tau(\varepsilon)) \geq M(h) - C_1\sqrt{2s\log p}\sum_{i\geq 0}\|h_{S_i}\|_2. \tag{6.6}$$

This together with (3.1) and (6.4) yields that with probability at least $1 - 2p^{-4s(C_2^2-1)+1}$,

$$M(h) \leq \frac{\lambda\sqrt{s}}{\sqrt{n}}\|h_{S_0}\|_2 + C_1\sqrt{2s\log p}\left(\frac{1}{\bar{c}} + \frac{5}{4}\right)\|h_{S_0}\|_2. \tag{6.7}$$

We below discuss two cases. On one hand, if $\|Xh\|_1 \geq 3n/(2a)$, from Lemma 7 in [35] and Lemma 3.3, we obtain that

$$\begin{aligned} M(h) &= \frac{1}{\sqrt{n}} E(Q_\tau(Xh + \varepsilon) - Q_\tau(\varepsilon)) \\ &\geq \frac{3}{16\sqrt{n}}(\tau \wedge (1-\tau))\|Xh\|_1 \geq \frac{3\sqrt{n}}{16}(\tau \wedge (1-\tau))\kappa_s^l\|h_{S_0}\|_2. \end{aligned} \tag{6.8}$$

This combining with assumption (C2) yields that $\|h_{S_0}\|_2 = 0$. So, $h = 0$ and $\hat{\beta} = \beta^*$.

On the other hand, if $\|Xh\|_1 < 3n/(2a)$, we similarly obtain that

$$M(h) = \frac{1}{\sqrt{n}} E(Q_\tau(Xh + \varepsilon) - Q_\tau(\varepsilon)) \geq \frac{a}{8\sqrt{n}}(\tau \wedge (1-\tau))\|Xh\|_2^2. \tag{6.9}$$

Moreover, by applying the common arguments in CS area (see, e.g., [8,11,19], we get

$$\begin{aligned} |\langle Xh_{S_0}, Xh\rangle| &= |\langle Xh_{S_0}, X(h_{S_0} + \sum_{i\geq 1} h_{S_i})\rangle| \\ &\geq n\lambda_s^l\|h_{S_0}\|_2^2 - n\theta_{s,s}\|h_{S_0}\|_2\sum_{i\geq 1}\|h_{S_i}\|_2 \\ &\geq n\left(\lambda_s^l - \theta_{s,s}\left(\frac{1}{\bar{c}} + \frac{1}{4}\right)\right)\|h_{S_0}\|_2^2, \end{aligned}$$

and

$$|\langle Xh_{S_0}, Xh\rangle| \le \|Xh_{S_0}\|_2\|Xh\|_2 \le \sqrt{n\lambda_s^u}\|h_{S_0}\|_2\|Xh\|_2.$$

It immediately follows

$$\|Xh\|_2^2 \ge \frac{n}{\lambda_s^u}\left(\lambda_s^l - \theta_{s,s}\left(\frac{1}{\bar{c}} + \frac{1}{4}\right)\right)^2 \|h_{S_0}\|_2^2.$$

Thus, it holds by (6.7) and (6.9) that with probability at least $1 - 2p^{-4s(C_2^2-1)+1}$,

$$\|h_{S_0}\|_2 \le \frac{8\lambda\sqrt{s}}{nab} + \frac{8C_1}{\sqrt{n}ab}\sqrt{2s\log p}\left(\frac{1}{\bar{c}} + \frac{1}{4}\right), \tag{6.10}$$

where $b = (\tau \wedge (1-\tau))\left(\lambda_s^l - \theta_{s,s}\left(\frac{1}{\bar{c}} + \frac{1}{4}\right)\right)^2/\lambda_s^u$. Since we can take $\lambda = 2c\sqrt{n\log p}$, then we yield

$$\|h_{S_0}\|_2 \le \frac{8(c\sqrt{2} + 1.25C_1 + C_1/\bar{c})}{ab}\sqrt{\frac{2s\log p}{n}}. \tag{6.11}$$

Noting that

$$\sum_{i\ge 1}\|h_{S_i}\|_2^2 \le |h_{s+1}|\sum_{i\ge 1}\|h_{S_i}\|_1 \le \frac{1}{\bar{c}}\|h_{S_0}\|_2^2,$$

we obtain that with probability at least $1 - 2p^{-4s(C_2^2-1)+1}$,

$$\|\hat{\beta} - \beta^*\|_2 \le \frac{8(c\sqrt{2} + 1.25C_1 + C_1/\bar{c})}{ab}\sqrt{1 + \frac{1}{\bar{c}}}\sqrt{\frac{2s\log p}{n}}.$$

We complete the proof. □

Proof of Theorem 3.6. In the noiseless case, from $\varepsilon = 0$ and (3.1), we easily obtain that

$$Q_\tau(Xh) = Q_\tau(Xh) - Q_\tau(0) \le \lambda(\|h_T\|_1 - \|h_{T^C}\|_1).$$

This shows $\|h_T\|_1 \ge \|h_{T^C}\|_1$ and then $h \in \Delta_1$. Therefore $\|Xh\|_1 \ge n\kappa_s^l(1)\|h_T\|_2 \ge n\kappa_s^l(1)\|h_T\|_1/s$. Observe that

$$Q_\tau(Xh) \ge (\tau \wedge (1-\tau))\|Xh\|_1.$$

From the assumption, it follows $\|h_T\|_1 = 0$ and hence $\hat{\beta} = \beta^*$ in the noiseless case. □

Acknowledgments

The authors would like to thank two anonymous referees for their numerous insightful comments and suggestions, which have greatly improved the paper.

References

[1] A. Barker, A. Lauria, N. Schloot, N. Hosszufalusi, J. Ludvigsson, C. Mathieu, D. Mauricio, M. Nordwall, B. Van der Schueren, T. Mandrup-Poulsen, W.A. Scherbaum, I. Weets, F.K. Gorus, N. Wareham, R. D. Leslie and P. Pozzilli, Age-dependent decline of β-cell function in type 1 diabetes after diagnosis: a multi-centre longitudinal study, *Diabetes Obes Metab.* 3 (2014) 262–267.

[2] A. Belloni and V. Chernozhukov, L1-penalized quantile regression in high-dimensional sparse models, *Ann. Statist.* 39 (2011) 82–130.

[3] A. Belloni, V. Chernozhukov and L. Wang, Square-root lasso: pivotal recovery of sparse signals via conic programming, *Biometrika* 98 (2011) 791–806.

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

[5] S. Boyd, N. Parikh, E. Chu, B. Peleato and J. Eckstein, Distributed optimization and statistical learning via the alternating direction method of multipliers, *Found. Trends Mach. Learn.* 3 (2011) 1–122.

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

[7] A.M. Bruckstein, D.L. Donoho and M. Elad, From sparse solutions of systems of equations to sparse modeling of signals and images, *SIAM Rev.* 51 (2009) 34–81.

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

[9] T. Cai, L. Wang and G. Xu, New Bounds for restricted isometry constants, *IEEE Trans. Inform. Theory* 56 (2010) 4388–4394.

[10] T. Cai, and A. Zhang, Sparse representation of a polytope and recovery of sparse signals and low–rank matrices, *IEEE Trans. Inform. Theory* 60 (2014) 122–132.

[11] E.J. Candès, The restricted isometry property and its implications for compressed sensing, *C. R. Acad. Sci. Ser. I* 346 (2008) 589–592.

[12] E.J. Candès, J. Romberg and T. Tao, Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency information, *IEEE Trans. Inform. Theory* 52 (2006) 489–509.

[13] E. J. Candès and T. Tao, Decoding by linear programming, *IEEE Trans. Inform. Theory* 51 (2005) 4203–4215.

[14] E. J. Candès and T. Tao, The Dantzig selector: statistical estimation when p is much larger than n, *Ann. Statist.* 35 (2007) 2313–2351.

[15] C. Chen and Y. Wei, Computational issues for quantile regression, *Sankhy*鐃緒申 67 (2005) 399–417.

[16] A. Cohen, W. Dahmen and R. DeVore, Compressed sensing and best k–term approximation, *J. Amer. Math. Soc.* 22 (2009) 211–231.

[17] G. Donald, S.Q. Ma and K. Scheinberg, Fast alternating linearization methods for minimizing the sum of two convex functions, *Math. Program.* 141 (2013) 349–382.

[18] D.L. Donoho, Compressed sensing, *IEEE Trans. Inform. Theory* 52 (2006) 1289–1306.

[19] Y.C. Eldar and G. Kutyniok, *Compressed sensing: theory and applications*, Cambridge University Press, 2012.

[20] Y. Fan, J. Fan and E. Barut, Adaptive robust variable selection, *Ann. Statist.* 42 (2014) 324–351.

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

[22] X. He, Modeling and Inference by Quantile Regression, Technical Report, University of Illinois at Urbana–Champaign, Department of Statistics, 2009.

[23] A. Juditsky, F. Karzan and A. S. Nemirovski, Verifiable conditions of ℓ_1–recovery of sparse signals with sign restrictions, *Math. Program.* 127 (2011) 89–122.

[24] A. Juditsky and A. S. Nemirovski, On verifiable sufficient conditions for sparse signal recovery via ℓ_1 minimization, *Math. Program.* 127 (2011) 57–88.

[25] R. Koenker, *Quantile regression*, Econometric Society Monographs 38, Cambridge University Press, 2005.

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

[27] S. Lambert–Lacroix and L. Zwald, Robust regression through the Huber's criterion and adaptive lasso penalty, *Electron. J. Stat.* 5 (2011) 1015–1053.

[28] X.X. Li, L.L. Mo, X.M. Yuan and J.Z. Zhang, Linearized alternating direction method of multipliers for sparse group and fused LASSO models, *Comput. Stat. Data Anal.* 79 (2014) 203–221.

[29] A.S. Lien, Y.D. Jiang, C.H. Mou, M.F. Sun, B.S. Gau and H.R. Yen, Integrative traditional Chinese medicine therapy reduces the risk of diabetic ketoacidosis in patients with type 1 diabetes mellitus, *J. Ethnopharmacology*, 191 (2016) 324–330.

[30] W.H. Lin, M.C. Wang, W.M. Wang, D.C. Yang, C.F. Lam, J.N. Roan and C.Y. Li, Incidence of and Mortality from Type I Diabetes in Taiwan From 1999 through 2010: A Nationwide Cohort Study, *PloS One*, e86172 (2014).

[31] H. Liu and L. Wang, Tiger: a tuning–insensitive approach for optimally estimating large undirected graphs, *submiited to Ann. Statist.* 2012.

[32] N. Meinshausen and B. Yu, Lasso–type recovery of sparse representations for high–dimensional data, *Ann. Statist.* 37 (2009) 246–270.

[33] R.T. Rockafellar and R.J.B. Wets, *Variational Analysis*, Second Edition, Springer, New York, 2004.

[34] R. Tibshirani, Regression shrinkage and selection via the Lasso, *J. R. Stat. Soc. Ser. B Stat. Methodol.* (1996) 267–288.

[35] L. Wang, L1 penalized LAD estimator for high dimensional linear regression, *J. Multivariate Anal.* 120 (2013) 135–151.

[36] L. Wang, Y. Wu and R. Li, Quantile regression for analyzing heterogeneity in ultrahigh dimension, *J. Amer. Statist. Assoc.* 107 (2012) 214–222.

[37] X. Xiu and L. Kong, Rank–min–one and sparse tensor decomposition for surveillance video, *Pac. J. Optim.* 11 (2015) 403–418.

[38] J.F. Yang and Y. Zhang, Alternating direction algorithms for ℓ_1–problems in compressive sensing, *SIAM J. Sci. Comput.* 33 (2011) 250–278.

Manuscript received 30 September 2016
revised 8 January 2016
accepted for publication 15 January 2017

Yanqing Liu
Department of Applied Mathematics
Beijing Jiaotong University
Beijing 100044, P. R. China
E–mail address: liuyanqing678@126.com

Guokai Liu
Department of Anesthesia, Dongzhimen Hospital Beijing
University of Chinese Medicine
No.5 Haiyuncang, Dongcheng District, Beijing 100700, P. R. China
E–mail address: guokai.liu@yahoo.com

Xianchao Xiu
Department of Applied Mathematics, School of Science
Beijing Jiaotong University, Beijing 100044, P. R. China
E–mail address: xianchaoxiu@163.com

Shenglong Zhou
Mathematical Sciences
University of Southampton, Southampton SO17 1BJ, UK
E–mail address: longnan_zsl@163.com