Yokohama Publishers
ISSN 1349-8169 ONLINE JOURNAL

Pacific Journal of Optimization
Vol. 16, No. 3, pp. 369–393, 2020

SMOOTHING QUANTILE REGRESSION WITH ELASTIC NET PENALTY*

Bingzhen Chen†, Lingchen Kong and Nana Xu

Abstract: High-dimensional data are commonly encountered in various scientific fields, such as information technology, biology, economics and so on. This poses great challenges to modern statistical analysis and optimal computation. First, the high dimensionality often induces the collinearity of variables. Second, the error of high-dimensional data may be heavy-tailed. In order to deal with these two issues, we introduce the penalized quantile regression with the elastic net penalty that combines the strengths of the quadratic regularization and the lasso shrinkage. By smoothing quantile loss function with the Huber smooth function, we give the smoothing quantile regression with elastic net penalty (SQEN). In this model, the regularizer which leads to a grouping effect can treat collinearity well and the Huber smooth loss is suitable for heavy-tailed data. In high-dimensional setting, we derive the statistical consistent property of the SQEN estimator. To make the SQEN practically feasible, we propose an efficient iterative SQEN-MM method and establish its global convergence. From numerical results, we can see our method can solve SQEN model efficiently and effectively.

Key words: *high-dimensional, smoothing quantile, elastic net penalty, statistical consistent property, majorize-minimize algorithm, heavy-tailed noise*

Mathematics Subject Classification: *47N10, 62J05, 49M15, 62-07, 65L20*

1 Introduction

High-dimensional data sets are easier to be collected because of the advent of modern technology, which are commonly encountered in various scientific fields, such as information technology, biology, economics and so on. Such data make great unprecedented challenges and opportunities for statistical analysis and optimal computation. 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. It is impossible to tackle such kind of data without additional assumptions. One natural assumption is the sparsity of the true coefficients. That is to say, only a few unknown variables can affect the value of the response.

Regularization methods are popular ways to select sparse variables since the ordinary least squares method is not consistent in the high-dimensional setting. Methods based on l_1 penalization or constrained l_1 minimization have been extensively studied, see [5,6,15,17,19, 20]. To be more specific, Tibshirani [17] proposed a technique for high-dimensional linear

*The work was supported in part by National Natural Science Foundation of China (11671029), the Key Program of Haibin College (HB202001002)

†Corresponding author

regression model, which is the so-called LASSO and is a penalized least squares method imposed an l_1-penalty on the regression coefficients. Owing to the nature of the l_1-penalty, the LASSO does both continuous shrinkage and automatic variables selection simultaneously. Zou et al [20] proposed a new method for variables selection which penalized least squares with elastic net penalty. The elastic net has a grouping effect, where strongly correlated predictors tend to be in or out the model together. The above LASSO type methods have nice properties under the Gaussian assumption. However, the Gaussian assumption may not hold in practice, especially in the high-dimensional setting.

Quantile regression introduced by Koenker et al. [10] performs well in the situations where the noises are heavy-tailed or heterogeneous, which includes the least absolute deviation (LAD) regression as one special case. Quantile regression has been widely studied in many different areas such as economics, survival analysis, technology, sociology and biology and so on, see [7, 11–14] for more details. Recently, regularized quantile regression was studied for high-dimensional sparse model, see, [1, 2, 4, 8, 9, 18, 21]. For instance, Wang [18] studied the l_1-penalized LAD regression and showed the estimator achieves near oracle risk performance with a nearly universal penalty parameter. Aravkin et al. [1] considered quantile regression with l_0 and l_1 penalties and substituted the quantile loss function with Huber smooth function. They proposed a generalized orthogonal matching pursuit (OMP) method for variable selection. Fan et al. [4] constructed a penalized quantile regression with the weighted l_1-penalty (WR-LASSO) and established the oracle properties of the estimator. Mkhadri et al. [16] discussed a coordinate descent algorithm for computing the penalized smooth quantile regression (cdaSQR) with convex and nonconvex penalties.

Motivated by the above arguments to handle the high-dimensional sparse model, we consider quantile regression with elastic net penalty (QEN) to regress the heavy-tailed and collinear high-dimensional data. Although the quantile loss function is convex, it isn't smooth. In order to get an efficient algorithm to solve the optimization problem, we will take advantages of the Huber smooth function. By replacing the quantile loss function with the Huber smooth function, we obtain the smoothing quantile regression with elastic net penalty (SQEN). We show that such model not only can produce a sparse solution, but also has grouping effect property. Meanwhile, our analysis shows that the SQEN estimator has statistical consistent properties. In order to get the SQEN estimator, we propose an iterative method based on the majorize minimize (MM) technique, which is called SQEN-MM. We then establish its global convergence. Finally, we illustrate the efficiency of the SQEN-MM algorithm by some numerical experiments.

The remainder of the paper is organized as follows. In section 2, we introduce the elastic-net penalized smoothing quantile regression model and show its grouping effect property. Then, we derive the statistical properties of the estimator in Section 3. In Section 4, we establish the algorithm of the SQEN-MM and show its convergence. In Section 5, numerical experiments are reported to show the efficiency of the proposed method. Finally, we make some conclusions in Section 6.

2 Elastic-Net Penalized Smoothing Quantile Regression

In this section, we first introduce the quantile regression with elastic net penalty (QEN). Then, we construct the smoothing quantile regression with elastic net penalty (SQEN) via smoothing the general quantile loss function by the Huber smooth function. Finally, we show that the SQEN model has a grouping effect property which is important to collinearity.

Consider the high-dimensional linear regression model

$$\mathbf{y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \tag{2.1}$$

where $X = (\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n)^{\mathrm{T}} = (\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_p)$ is a $n \times p$ design matrix with $p >> n$. Here the elements of $\mathbf{x}_i$ denote the values of predictors for i-th sample, the elements of $\mathbf{X}_j$ represent the values of j-th predictor for n samples. Throughout the paper, we use bold letters to represent column vectors. $\mathbf{y} = (y_1, y_2, \ldots, y_n)^{\mathrm{T}}$ is a n-dimensional response vector, $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_p)^{\mathrm{T}}$ is a p-dimensional regression coefficient vector, and $\boldsymbol{\varepsilon} = (\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n)^{\mathrm{T}}$ is the n-dimensional measurement error/noise vector with all components $\varepsilon_i (i = 1, 2, \ldots, n)$ being independently distributed and satisfying $\Pr(\varepsilon_i \leq 0) = \tau$ for some known constant $\tau \in (0, 1)$. In our model, there is no intercept, so we assume $X = (\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_p)$ with each vector $\mathbf{X}_i$ being normalized such that $\|\mathbf{X}_i\|_2 = \sqrt{n}$ for $i = 1, 2, \ldots, p$. Under this model, for a given $\mathbf{x}_i$, $\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}$ is the conditional τth-quantile of y_i. For the high-dimensional linear regression problem, a key assumption is the sparsity of the true coefficient vector $\boldsymbol{\beta}^*$, which means that the proportion of nonzero coefficients is small. Sparsity guarantees the model identifiability and enhances the model fitting accuracy and interpretability. Here, we denote the number of nonzero coefficients by s. Without loss of generality, we assume $\boldsymbol{\beta}^* = ((\boldsymbol{\beta}_1^*)^{\mathrm{T}}, \mathbf{0}^{\mathrm{T}})^{\mathrm{T}}$ with $\boldsymbol{\beta}_1^* \in \mathbb{R}^s, \mathbf{0} \in \mathbb{R}^{p-s}$, that is, only the first s entries are nonzero. We rewrite the design matrix X as $X = (S, Q)$ where $S = (\mathbf{X}_1, \ldots, \mathbf{X}_s)$ the submatrix of X corresponds to the covariates whose coefficient are nonzero. We view those covariates as signal covariates and the rest as noise covariates. Here, the submatrix $Q = (\mathbf{X}_{s+1}, \ldots, \mathbf{X}_n)$ corresponds to the noise covariates whose coefficients are zero.

In order to estimate sparse regression coefficient vector $\boldsymbol{\beta}^*$, we propose the quantile regression with elastic net penalty (QEN)

$$\min_{\boldsymbol{\beta} \in \mathbb{R}^p} \left\{ \sum_{i=1}^{n} \rho_\tau(y_i - \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}) + n\lambda_1 \|\boldsymbol{\beta}\|_1 + n\lambda_2 \|\boldsymbol{\beta}\|_2^2 \right\}. \tag{2.2}$$

The quantile loss function is defined as $\rho_\tau(x) = (\tau - \mathbf{1}(x \leq 0))x$, $\tau \in (0, 1)$, with $\mathbf{1}(\cdot)$ being the indicator function, $\|\boldsymbol{\beta}\|_1 = \sum_{i=1}^p |\beta_i|$ is the l_1-norm of $\boldsymbol{\beta}$, $\|\boldsymbol{\beta}\|_2 = \sqrt{\sum_{i=1}^p \beta_i^2}$ is the l_2-norm of $\boldsymbol{\beta}$, and $\lambda_1 > 0$, $\lambda_2 > 0$, are the penalized/regularization parameters. We call $n\lambda_1 \|\boldsymbol{\beta}\|_1 + n\lambda_2 \|\boldsymbol{\beta}\|_2^2$ the elastic net penalty term.

The quantile loss function in model (2.2) is strongly convex but nonsmooth. It is difficult to calculate the sub-differential of the objective function. One way to study nonsmooth model is to adopt a smooth loss function to substitute the nonsmooth loss function. In this paper, we will use the quantile Huber smooth function to take the place of the quantile loss function.

For any given $\tau \in (0, 1)$ and $\delta > 0$, the quantile Huber smooth function is defined as

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

It is obvious that this function is convex. By the definition of $\rho_\tau(\theta, \delta)$, we can obtain the following differential proposition. The proof of the proposition is simple, which is omitted for brevity.

Proposition 2.1. *For any fixed $\delta > 0$ and $\tau \in (0, 1)$, the function $\rho_\tau(y_i - \mathbf{x}_i^T\boldsymbol{\beta}, \delta)$ is continuously differentiable with respect to $\boldsymbol{\beta} \in \mathbb{R}^p$. Moreover, its first order partial derivative*

with respect to $\boldsymbol{\beta}$ *is*

$$\nabla_{\boldsymbol{\beta}}\rho_\tau(y_i-\mathbf{x}_i^T\boldsymbol{\beta},\delta)=-\mathbf{x}_i\rho'_\tau(y_i-\mathbf{x}_i^T\boldsymbol{\beta},\delta)=\begin{cases}-(\tau-1)\mathbf{x}_i, & \theta_i\in(-\infty,(\tau-1)\delta),\\ -\frac{\theta_i}{\delta}\mathbf{x}_i, & \theta_i\in[(\tau-1)\delta,\tau\delta],\\ -\tau\mathbf{x}_i, & \theta_i\in(\tau\delta,+\infty).\end{cases}$$

where $\theta_i=y_i-\mathbf{x}_i^T\boldsymbol{\beta}$ *for all* $i=1,2,\ldots,n$.

Now the model (2.2) can be smoothed as smoothing quantile regression with elastic net penalty (SQEN)

$$\min_{\boldsymbol{\beta}\in\mathbb{R}^p}\Phi_\tau(\boldsymbol{\beta},\delta):=\left\{S_\tau(\boldsymbol{\beta},\delta)+n\lambda_1\|\boldsymbol{\beta}\|_1+n\lambda_2\|\boldsymbol{\beta}\|_2^2\right\}, \tag{2.3}$$

where $S_\tau(\boldsymbol{\beta},\delta)=\sum_{i=1}^n\rho_\tau(y_i-\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta},\delta)$. Let $\widehat{\boldsymbol{\beta}},\widehat{\boldsymbol{\beta}}^\delta$ stand for the solution of (2.2), (2.3), respectively. It is obvious that $\widehat{\boldsymbol{\beta}}_\delta\to\widehat{\boldsymbol{\beta}}$ when $\delta\to0$. Therefore, in order to obtain the optimal solution of (2.2), we can take advantage of $\rho_\tau(\cdot,\delta)$. So we will mainly discuss the properties about the problem (2.3). Based on this model, we will propose an effective method to purse the sparse solution of high-dimensional linear regression in Section 4. In the following, we will show model (2.3) possesses an important property–grouping effect. First, let's consider an extreme situation.

Proposition 2.2. *Let* $\widehat{\boldsymbol{\beta}}$ *be the minimizer of* (2.3). *If we assume* $\mathbf{X}_i=\mathbf{X}_j, i,j\in\{1,2,\ldots,p\}$, *then* $\widehat{\beta}_i=\widehat{\beta}_j$ *for any* $\lambda_1,\lambda_2>0$.

Proof. Assume $\widehat{\beta}_i\neq\widehat{\beta}_j$. We construct $\widehat{\boldsymbol{\beta}}^*$ as follows

$$\widehat{\beta}_k^*=\begin{cases}\widehat{\beta}_k, & \text{if } k\neq i \text{ and } k\neq j,\\ \frac{1}{2}(\widehat{\beta}_i+\widehat{\beta}_j), & \text{if } k=i \text{ or } k=j.\end{cases}$$

Since $\mathbf{X}_i=\mathbf{X}_j$, it is obvious that $X\widehat{\boldsymbol{\beta}}^*=X\widehat{\boldsymbol{\beta}}$, which means that $\mathbf{x}_i^{\mathrm{T}}\widehat{\boldsymbol{\beta}}^*=\mathbf{x}_i^{\mathrm{T}}\widehat{\boldsymbol{\beta}}$, for $i=1,2,\ldots,n$. Thus $\sum_{i=1}^n\rho_\tau(y_i-\mathbf{x}_i^{\mathrm{T}}\hat{\boldsymbol{\beta}}^*,\delta)=\sum_{i=1}^n\rho_\tau(y_i-\mathbf{x}_i^{\mathrm{T}}\widehat{\boldsymbol{\beta}},\delta)$. Considering two simple inequalities $|\frac{1}{2}(\widehat{\beta}_i+\widehat{\beta}_j)|\le\frac{1}{2}\big(|\widehat{\beta}_i|+|\widehat{\beta}_j|\big)$, and $\big(\frac{\widehat{\beta}_i+\widehat{\beta}_j}{2}\big)^2<\frac{1}{2}\big(\widehat{\beta}_i^2+\widehat{\beta}_j^2\big)$, we have $\Phi_\tau(\widehat{\boldsymbol{\beta}}^*,\delta)<\Phi_\tau(\widehat{\boldsymbol{\beta}},\delta)$. Therefore, $\widehat{\boldsymbol{\beta}}$ can't be the minimizer of model (2.3), which is a contradiction. So we obtain the desired result $\widehat{\beta}_i=\widehat{\beta}_j$. □

This proposition exhibits that when two predictors $\mathbf{X}_i,\mathbf{X}_j$ are equal, then their coefficients $\widehat{\beta}_i$ and $\widehat{\beta}_j$ will also be equal. In the next theorem, we will generalize this result and establish an upper bound for the difference $|\widehat{\beta}_i-\widehat{\beta}_j|$ by using the sample correlation between $\mathbf{X}_i$ and $\mathbf{X}_j$.

Theorem 2.3. *Given data* $(\mathbf{y},X)$ *and parameters* $\lambda_1,\lambda_2\ge0$, *the predictors are standardized, i.e.,* $||\mathbf{X}_i||_2=\sqrt{n}, i=1,2,\ldots,p$. *Let* $\widehat{\boldsymbol{\beta}}$ *be the SQEN estimator. Suppose that* $\widehat{\beta}_i\widehat{\beta}_j>0$. *Then*

$$|\widehat{\beta}_i-\widehat{\beta}_j|\le\frac{1}{2\lambda_2}\sqrt{2(1-\rho)}, \tag{2.4}$$

where $\rho=\frac{1}{n}\mathbf{X}_i^T\mathbf{X}_j$ *is the sample correlation.*

Proof. If $\widehat{\beta}_i\widehat{\beta}_j>0$, then we have $\mathrm{sign}(\widehat{\beta}_i)=\mathrm{sign}(\widehat{\beta}_j)$. From the first order optimal condition of the elastic-net penalized smoothing quantile regression problem (2.3), we obtain that

$$\mathbf{0}\in\partial_{\boldsymbol{\beta}}\left(S_\tau(\boldsymbol{\beta},\delta)+n\lambda_1||\boldsymbol{\beta}||_1+n\lambda_2||\boldsymbol{\beta}||_2^2\right)|_{\boldsymbol{\beta}=\widehat{\boldsymbol{\beta}}} \tag{2.5}$$

Let $\boldsymbol{\rho}_\tau = (\rho'_\tau(y_1 - \mathbf{x}_1^{\mathrm{T}}\boldsymbol{\beta}, \delta), \rho'_\tau(y_2 - \mathbf{x}_2^{\mathrm{T}}\boldsymbol{\beta}, \delta), \ldots, \rho'_\tau(y_n - \mathbf{x}_n^{\mathrm{T}}\boldsymbol{\beta}, \delta))^{\mathrm{T}}$, then we can obtain $\partial_{\boldsymbol{\beta}} S_\tau(\boldsymbol{\beta}, \delta) = -X^{\mathrm{T}}\boldsymbol{\rho}_\tau$. Therefore, for nonzero $\widehat{\beta}_i$ and $\widehat{\beta}_j$, we have

$$\begin{aligned} 0 &= -\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\rho}_\tau + n\lambda_1 \mathrm{sign}(\widehat{\beta}_i) + 2n\lambda_2\widehat{\beta}_i, \\ 0 &= -\mathbf{X}_j^{\mathrm{T}}\boldsymbol{\rho}_\tau + n\lambda_1 \mathrm{sign}(\widehat{\beta}_j) + 2n\lambda_2\widehat{\beta}_j. \end{aligned}$$

Note that $\mathrm{sign}(\widehat{\beta}_i) = \mathrm{sign}(\widehat{\beta}_j)$, we obtain

$$0 = -(\mathbf{X}_i - \mathbf{X}_j)^{\mathrm{T}}\boldsymbol{\rho}_\tau + 2n\lambda_2(\widehat{\beta}_i - \widehat{\beta}_j).$$

Then, we have $|\widehat{\beta}_i - \widehat{\beta}_j| = \frac{1}{2n\lambda_2}|(\mathbf{X}_i - \mathbf{X}_j)^{\mathrm{T}}\rho_\tau|$. Clearly, $||\boldsymbol{\rho}_\tau||_2 \leq \sqrt{n}$ and hence

$$\begin{aligned} |\widehat{\beta}_i - \widehat{\beta}_j| &= \frac{1}{2n\lambda_2}|(\mathbf{X}_i - \mathbf{X}_j)^{\mathrm{T}}\boldsymbol{\rho}_\tau| \\ &\leq \frac{1}{2n\lambda_2}||\mathbf{X}_i - \mathbf{X}_j||_2||\boldsymbol{\rho}_\tau||_2 \\ &\leq \frac{1}{2\sqrt{n}\lambda_2}\sqrt{||\mathbf{X}_i||_2^2 + ||\mathbf{X}_j||_2^2 - 2\mathbf{X}_i^{\mathrm{T}}\mathbf{X}_j} \\ &\leq \frac{1}{2\lambda_2}\sqrt{2(1-\rho)}. \end{aligned}$$

□

Note that, when $\rho \to 1$, $\widehat{\beta}_i$ and $\widehat{\beta}_j$ tend to be equal which means that strong correlated predictors tend to have the same coefficients. Thus, in the estimation model, they will be in or out of the model together. It has been pointed out that LASSO doesn't possess this property in [20]. This is why we use elastic net penalty in this paper.

3 Statistical Property

In this section, we will establish statistical consistent properties of SQEN. We start with some notations. Following the terminology in [3], we define the oracle regularized estimator (ORE) as $\boldsymbol{\beta}^0 = (\boldsymbol{\beta}_1^{0^{\mathrm{T}}}, \mathbf{0}^{\mathrm{T}})^{\mathrm{T}}$ with $\boldsymbol{\beta}_1^0 \in \mathbb{R}^s$ and $\mathbf{0} \in \mathbb{R}^{p-s}$ being the vector of all 0, which minimizes $\Phi_\tau(\boldsymbol{\beta}, \delta)$ over the subspace $\{\boldsymbol{\beta} = (\boldsymbol{\beta}_1^{\mathrm{T}}, \boldsymbol{\beta}_2^{\mathrm{T}})^{\mathrm{T}} \in \mathbb{R}^p : \boldsymbol{\beta}_2 = \mathbf{0} \in \mathbb{R}^{p-s}\}$. We claim that the ORE is consistent with the true coefficient vector. We then show that the SQEN estimator enjoys the same property as the ORE when some conditions are met. In order to do so, we need introduce some conditions.

Condition 3.1. If $f_i(x)$ and $F_i(x)$ are the density function and distribution function of the error ε_i, respectively. Then we assume that there exists universal constants $c_1 > 0$ and $c_2 > 0$ such that for any x satisfying $|x| \leq c_1$, $f_i(x)$ is uniformly bounded in $(0, \infty)$ and

$$|F_i(x) - F_i(0) - xf_i(0)| \leq c_2 x^2. \tag{3.1}$$

Condition 3.1 is commonly used in noise distribution, which says the Lipschitz property of $f_i(x)$ around the origin. For instance, the Laplace distribution and stable distributions including the normal distribution, Cauchy distribution all satisfy this condition.

Condition 3.2. Let $H = \mathrm{diag}(f_1(0), \ldots, f_n(0)) \in \mathbb{R}^{n\times n}$ be a diagonal matrix. The eigenvalues of $\frac{1}{n}S^{\mathrm{T}}HS$ are sandwiched as

$$0 < \lambda_l \leq \lambda\left(\frac{1}{n}S^{\mathrm{T}}HS\right) \leq \lambda_u \leq \infty.$$

Furthermore,

$$\kappa_n \equiv \max_{i,j}|x_{ij}| = o(\sqrt{n}s^{-1}). \tag{3.2}$$

Condition 3.2 is on the submatrix S and the magnitude of the entries of X. It is worth to note that the above condition on κ_n is satisfied with asymptotic probability one when the design matrix is generated from some distributions. If the entries of X are independent copies from a subexponential distribution, the bound on κ_n is satisfied with asymptotic probability one as long as $s = (\sqrt{n}/\log p)$; if the components are generated from sub-Gaussian distribution, then the condition on κ_n is satisfied with probability tending to one when $s = o(\sqrt{n/\log p})$

Condition 3.3. Let $\gamma_n = C_1(\sqrt{s(\log n)/n} + \sqrt{s}\lambda_1 + 2\lambda_2\|\boldsymbol{\beta}_1^*\|_2)$ with $C_1 > 0$ a constant. It holds that

$$\left\|\frac{1}{n}Q^{\mathrm{T}}HS\right\|_{2,\infty} < \frac{\lambda_1}{2\gamma_n},$$

where $\|A\|_{2,\infty} = \sup_{x\neq 0}\|A\mathbf{x}\|_\infty/\|\mathbf{x}\|_2$ for a matrix A and a vector $\mathbf{x}$. Furthermore, $\log(p) = o(n^b)$ for a constant $b \in (0, 1)$.

Condition 3.3 is on the correlation of columns in the design submatrix Q and S of X. Note that the optimal λ_1 should be larger than $\sqrt{(\log p)/n}$ when the data has heavy-tailed errors.

The following theorem states that the ORE can estimate the correct sign of the true coefficient vector with probability tending to one, which tells that how well one can do with the assistance of the oracle information on the location of signal covariates.

Theorem 3.1. *Let $\gamma_n = C_1(2\sqrt{s(\log n)/n} + \sqrt{s}\lambda_1 + 2\lambda_2\|\boldsymbol{\beta}_1^*\|_2)$ with a constant $C_1 > 0$. Assume Conditions 3.1 and 3.2 hold if $(\sqrt{s}\lambda_1 + 2\lambda_2\|\boldsymbol{\beta}_1^*\|_2)\sqrt{s}\kappa_n \to 0$, and δ is sufficiently small, then there exists some constant $c > 0$ such that*

$$P(\|\boldsymbol{\beta}_1^0 - \boldsymbol{\beta}_1^*\|_2 \le \gamma_n) \ge 1 - n^{-cs}. \tag{3.3}$$

If in addition $\gamma_n^{-1}\min_{1\le j\le s}|\boldsymbol{\beta}_{1j}^*| \to \infty$, then with probability at least $1 - n^{-cs}$,

$$\operatorname{sign}(\boldsymbol{\beta}_1^0) = \operatorname{sign}(\boldsymbol{\beta}_1^*), \tag{3.4}$$

where the above should be understood componentwisely.

Proof. The proof is motivated by Theorem 1 in [4], where they deal with l_1-norm penalized quantile function. Here we consider a different case, elastic-net penalized smoothing quantile regression model. For a given deterministic $M > 0$, we define the set

$$\mathcal{B}_0(M) = \{\boldsymbol{\beta} \in \mathbb{R}^p : \|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_2 \le M, \operatorname{supp}(\boldsymbol{\beta}) \subseteq \operatorname{supp}(\boldsymbol{\beta}^*)\}. \tag{3.5}$$

Then, define the function

$$Z_n(M) = \sup_{\boldsymbol{\beta}\in\mathcal{B}_0(M)} \frac{1}{n}\left|(S_\tau(\boldsymbol{\beta}, \delta) - S_\tau(\boldsymbol{\beta}^*, \delta)) - \mathbb{E}\left[S_\tau(\boldsymbol{\beta}, \delta) - S_\tau(\boldsymbol{\beta}^*, \delta)\right]\right|. \tag{3.6}$$

Observing that $\boldsymbol{\beta}^*$ is the minimizer of the function $\mathbb{E}(S_\tau(\boldsymbol{\beta}, \delta))$ for any $\boldsymbol{\beta} = (\boldsymbol{\beta}_1^{\mathrm{T}}, 0^{\mathrm{T}})^{\mathrm{T}} \in \mathcal{B}_0(M)$ with $M = o(\kappa_n^{-1}s^{-1/2})$, we immediately obtain that the first order derivation of $\mathbb{E}[S_\tau(\boldsymbol{\beta}, \delta) - S_\tau(\boldsymbol{\beta}^*, \delta)]$ is zero at the point $\boldsymbol{\beta} = \boldsymbol{\beta}^*$. Then, we can control $\mathbb{E}[S_\tau(\boldsymbol{\beta}, \delta) -$

$S_\tau(\boldsymbol{\beta}^*,\delta)]$ in terms of $Z_n(M)$. In addition, if we can show that for any $\boldsymbol{\beta}=(\boldsymbol{\beta}_1^{\mathrm{T}},0^{\mathrm{T}})^{\mathrm{T}}\in\mathcal{B}_0(M)$ with $M=o(\kappa_n^{-1}s^{-1/2})$,

$$\mathbb{E}[S_\tau(\boldsymbol{\beta},\delta)-S_\tau(\boldsymbol{\beta}^*,\delta)]\geq\frac{1}{3}\lambda_l n\|\boldsymbol{\beta}_1-\boldsymbol{\beta}_1^*\|_2^2, \tag{3.7}$$

we will easily establish the convergence in Theorem 3.1.

We start to prove (3.7). Taking $a_i=|\mathbf{S}_i^{\mathrm{T}}(\boldsymbol{\beta}_1-\boldsymbol{\beta}_1^*)|$, we have for any $\boldsymbol{\beta}\in\mathcal{B}_0(M)$,

$$a_i\leq\|\mathbf{S}_i\|_2\|\boldsymbol{\beta}_1-\boldsymbol{\beta}_1^*\|_2\leq\sqrt{s}\kappa_n M\to 0. \tag{3.8}$$

We consider two cases of $a_i=|\mathbf{S}_i^{\mathrm{T}}(\boldsymbol{\beta}_1-\boldsymbol{\beta}_1^*)|$. When $a_i\leq\delta$, we can obtain that

$$\begin{aligned}
&\rho_\tau(\varepsilon_i-a_i,\delta)-\rho_\tau(\varepsilon_i,\delta)\\
&=\begin{cases}
(1-\tau)a_i, & \varepsilon_i<(\tau-1)\delta,\\
-\frac{1}{2\delta}(\varepsilon_i-(\tau-1)\delta)^2+(1-\tau)a_i, & (\tau-1)\delta\leq\varepsilon_i<(\tau-1)\delta+a_i,\\
\frac{1}{2\delta}(-2\varepsilon_i a_i+a_i^2), & (\tau-1)\delta+a_i\leq\varepsilon_i\leq\tau\delta,\\
\frac{1}{2\delta}(\varepsilon_i-(\tau\delta+a_i))^2-\tau a_i, & \tau\delta<\varepsilon_i\leq\tau\delta+a_i,\\
-\tau a_i, & \varepsilon_i>\tau\delta+a_i.
\end{cases}\\
&=\begin{cases}
(1-\tau)a_i, & \varepsilon_i<(\tau-1)\delta,\\
-\frac{1}{2\delta}(\varepsilon_i-(\tau-1)\delta)^2+(1-\tau)a_i, & (\tau-1)\delta\leq\varepsilon_i<(\tau-1)\delta+a_i,\\
\frac{1}{2\delta}(-2\varepsilon_i a_i+a_i^2)-(1-\tau)a_i+(1-\tau)a_i, & (\tau-1)\delta+a_i\leq\varepsilon_i<0,\\
\frac{1}{2\delta}(-2\varepsilon_i a_i+a_i^2)+\tau a_i-\tau a_i, & 0\leq\varepsilon_i\leq\tau\delta,\\
\frac{1}{2\delta}(\varepsilon_i-(\tau\delta+a_i)^2-\tau a_i, & \tau\delta<\varepsilon_i\leq\tau\delta+a_i,\\
-\tau a_i, & \varepsilon_i>\tau\delta+a_i.
\end{cases}
\end{aligned}$$

Using the indicator function, we have

$$\rho_\tau(\varepsilon_i-a_i,\delta)-\rho_\tau(\varepsilon_i,\delta)=I_1-I_2+I_3+I_4+I_5, \tag{3.9}$$

where

$$\begin{aligned}
I_1&=(1-\tau)a_i\cdot\mathbf{1}\{\varepsilon_i<0\}-\tau a_i\cdot\mathbf{1}\{\varepsilon_i\geq 0\},\\
I_2&=\frac{1}{2\delta}(\varepsilon_i-(\tau-1)\delta)^2\cdot\mathbf{1}\{(\tau-1)\delta\leq\varepsilon_i<(\tau-1)\delta+a_i\},\\
I_3&=\left(\frac{1}{2\delta}(-2\varepsilon_i a_i+a_i^2)-(1-\tau)a_i\right)\cdot\mathbf{1}\{(\tau-1)\delta+a_i\leq\varepsilon_i<0\},\\
I_4&=\left(\frac{1}{2\delta}(-2\varepsilon_i a_i+a_i^2)+\tau a_i\right)\cdot\mathbf{1}\{0\leq\varepsilon_i\leq\tau\delta\},\\
I_5&=\frac{1}{2\delta}\left(\varepsilon_i-(\tau\delta+a_i)\right)^2\cdot\mathbf{1}\{\tau\delta<\varepsilon_i\leq\tau\delta+a_i\}.
\end{aligned}$$

Hence $\mathbb{E}\left[\rho_\tau(\varepsilon_i-a_i,\delta)-\rho_\tau(\varepsilon_i,\delta)\right]=\mathbb{E}[I_1]-\mathbb{E}[I_2]+\mathbb{E}[I_3]+\mathbb{E}[I_4]+\mathbb{E}[I_5]$. Because $\Pr(\varepsilon_i\leq 0)=\tau,\mathbb{E}\left[\mathbf{1}\{\varepsilon_i\leq 0\}\right]=\tau$, we can directly calculate

$$\mathbb{E}[I_1]=\mathbb{E}\left[(1-\tau)a_i\cdot\mathbf{1}\{\varepsilon_i<0\}-\tau a_i\cdot\mathbf{1}\{\varepsilon_i\geq 0\}\right]=0,$$

and

$$\mathbb{E}(I_2)=\mathbb{E}\left[\frac{1}{2\delta}(\varepsilon_i-(\tau-1)\delta)^2\cdot\mathbf{1}\{(\tau-1)\delta\leq\varepsilon_i<(\tau-1)\delta+a_i\}\right]$$

$$
\begin{aligned}
&= \mathbb{E}\left[\frac{1}{\delta}\int_0^{a_i} s\cdot \mathbf{1}\{s+(\tau-1)\delta \le \varepsilon_i < (\tau-1)\delta + a_i\}\mathrm{d}s\right] \\
&= \frac{1}{\delta}\int_0^{a_i} s\cdot \mathbb{E}\left[\mathbf{1}\{s+(\tau-1)\delta \le \varepsilon_i < (\tau-1)\delta + a_i\}\right]\mathrm{d}s \quad \text{(by Fubini's theorem)} \\
&= \frac{1}{\delta}\int_0^{a_i} s\cdot (F_i((\tau-1)\delta + a_i) - F_i(s+(\tau-1)\delta))\mathrm{d}s \\
&= \frac{1}{\delta}\int_0^{a_i} s\cdot (f_i(0)(a_i - s) + o(1)(a_i - s))\mathrm{d}s \quad \text{(by Condition (3.1))} \\
&= \frac{a_i^3}{6\delta} f_i(0) + \frac{a_i^3}{6\delta} o(1).
\end{aligned}
$$

Similarly,

$$
\begin{aligned}
\mathbb{E}[I_3] &= (f_i(0) + o(1))\frac{-(\tau-1)^2\delta a_i - a_i^2(\tau-1)}{2}, \\
\mathbb{E}[I_4] &= (f_i(0) + o(1))\frac{\tau^2 a_i\delta + \tau a_i^2}{2}, \\
\mathbb{E}[I_5] &= \frac{a_i^3}{6\delta} f_i(0) + \frac{a_i^3}{6\delta} o(1).
\end{aligned}
$$

Hence, combining the above analysis on $\mathbb{E}(I_i), i = 1, \ldots, 5$, we obtain

$$
\begin{aligned}
\mathbb{E}\left[\rho_\tau(\varepsilon_i - a_i, \delta) - \rho_\tau(\varepsilon_i, \delta)\right] &= \mathbb{E}[I_1] - \mathbb{E}[I_2] + \mathbb{E}[I_3] + \mathbb{E}[I_4] + \mathbb{E}[I_5] \\
&= 0 - \left(\frac{a_i^3}{6\delta} f_i(0) + \frac{a_i^3}{6\delta} o(1)\right) \\
&\quad + \frac{-(\tau-1)^2\delta a_i - a_i^2(\tau-1)}{2}(f_i(0) + o(1)) \\
&\quad + \frac{\tau^2 a_i\delta + \tau a_i^2}{2}(f_i(0) + o(1)) + \frac{a_i^3}{6\delta} f_i(0) + \frac{a_i^3}{6\delta} o(1) \\
&= \frac{\tau^2 a_i\delta + \tau a_i^2 - (\tau-1)^2\delta a_i - a_i^2(\tau-1)}{2}(f_i(0) + o(1)) \\
&= \frac{a_i^2}{2} f_i(0) + \frac{(2\tau-1)a_i\delta}{2} f_i(0) + \frac{(2\tau-1)a_i\delta + a_i^2}{2} o(1).
\end{aligned}
$$

where the $o(1)$ is uniformly over all $i = 1, \ldots, n$. When $a_i \ge \delta$, we can obtain the same result. And when δ is sufficiently small and tends to 0, we easily obtain $\mathbb{E}\left[\rho_\tau(\varepsilon_i - a_i, \delta) - \rho_\tau(\varepsilon_i, \delta)\right] \ge \frac{a_i^2}{3} f_i(0)$. Furthermore, by Condition 3.2,

$$
\begin{aligned}
\mathbb{E}\left[S_\tau(\boldsymbol{\beta}, \delta) - S_\tau(\boldsymbol{\beta}^*, \delta)\right] &= \sum_{i=1}^n \mathbb{E}\left[\rho_\tau(\varepsilon_i - a_i, \delta) - \rho_\tau(\varepsilon_i, \delta)\right] \\
&\ge \frac{1}{3}\sum_{i=1}^n a_i^2 f_i(0) \\
&= \frac{1}{3}(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_1^*)^{\mathrm{T}} S^{\mathrm{T}} H S(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_1^*) \\
&\ge \frac{1}{3}\lambda_l n \|\boldsymbol{\beta}_1 - \boldsymbol{\beta}_1^*\|_2^2.
\end{aligned}
$$

So the inequality (3.7) holds for any $\boldsymbol{\beta} \in (\boldsymbol{\beta}_1^{\mathrm{T}}, \mathbf{0}^{\mathrm{T}})^{\mathrm{T}} \in \mathcal{B}_0(M)$, but $\boldsymbol{\beta}^0 = ((\boldsymbol{\beta}_1^0)^{\mathrm{T}}, \mathbf{0}^{\mathrm{T}})^{\mathrm{T}}$ may not be in the set. Thus, we let $\widetilde{\boldsymbol{\beta}} = ((\widetilde{\boldsymbol{\beta}}_1^0)^{\mathrm{T}}, \mathbf{0}^{\mathrm{T}})^{\mathrm{T}}$, where

$$\widetilde{\boldsymbol{\beta}}_1^0 = u\boldsymbol{\beta}_1^0 + (1-u)\boldsymbol{\beta}_1^*, \text{ with } u = M/(M + \|\boldsymbol{\beta}_1^0 - \boldsymbol{\beta}_1^*\|_2), \tag{3.10}$$

which falls in the set $\mathcal{B}_0(M)$. Then, by the convexity and the definition of $\boldsymbol{\beta}_1^0$,

$$\Phi_\tau(\widetilde{\boldsymbol{\beta}}, \delta) \le u\Phi_\tau(\boldsymbol{\beta}_1^0, 0) + (1-u)\Phi_\tau(\boldsymbol{\beta}_1^*, 0) \le \Phi_\tau(\boldsymbol{\beta}_1^*, 0) = \Phi_\tau(\boldsymbol{\beta}^*, \delta). \tag{3.11}$$

Using this and triangle inequality, we have

$$\begin{aligned}
&\mathbb{E}\left[S_\tau(\widetilde{\boldsymbol{\beta}}, \delta) - S_\tau(\boldsymbol{\beta}^*, \delta)\right] \\
&= \{S_\tau(\boldsymbol{\beta}^*, \delta) - \mathbb{E}\left[S_\tau(\boldsymbol{\beta}^*, \delta)\right]\} - \left\{S_\tau(\widetilde{\boldsymbol{\beta}}, \delta) - \mathbb{E}\left[S_\tau(\widetilde{\boldsymbol{\beta}}, \delta)\right]\right\} \\
&\quad + \Phi_\tau(\widetilde{\boldsymbol{\beta}}, \delta) - \Phi_\tau(\boldsymbol{\beta}^*, \delta) + n\lambda_1\|\boldsymbol{\beta}_1^*\|_1 - n\lambda_1\|\widetilde{\boldsymbol{\beta}}_1\|_1 + n\lambda_2\|\boldsymbol{\beta}_1^*\|_2^2 - n\lambda_2\|\widetilde{\boldsymbol{\beta}}_1\|_2^2 \\
&\le nZ_n(M) + n\lambda_1\|\boldsymbol{\beta}_1^* - \widetilde{\boldsymbol{\beta}}_1\|_1 + 2n\lambda_2\left\langle \boldsymbol{\beta}_1^*, \boldsymbol{\beta}_1^* - \widetilde{\boldsymbol{\beta}}_1 \right\rangle - n\lambda_2\|\boldsymbol{\beta}_1^* - \widetilde{\boldsymbol{\beta}}_1\|_2^2 \\
&\le nZ_n(M) + n\sqrt{s}\lambda_1\|\boldsymbol{\beta}_1^* - \widetilde{\boldsymbol{\beta}}_1\|_2 + 2n\lambda_2\|\boldsymbol{\beta}_1^*\|_2\|\boldsymbol{\beta}_1^* - \widetilde{\boldsymbol{\beta}}_1\|_2 - n\lambda_2\|\boldsymbol{\beta}_1^* - \widetilde{\boldsymbol{\beta}}_1\|_2^2.
\end{aligned}$$

Define the event $\Gamma_n = \left\{Z_n(M) \le 2Mn^{-1/2}\sqrt{s\log n}\right\}$. Then by Lemma 1 in [4], we have

$$\Pr(\Gamma_n) \ge 1 - \exp(-c_0 s(\log n)/8). \tag{3.12}$$

On the event Γ_n, by (3.12), we have

$$\mathbb{E}\left[S_\tau(\widetilde{\boldsymbol{\beta}}, \delta) - S_\tau(\boldsymbol{\beta}^*, \delta)\right] \le 2M\sqrt{sn(\log n)} + n(\sqrt{s}\lambda_1 + 2\lambda_2\|\boldsymbol{\beta}_1^*\|_2)M - n\lambda_2\|\boldsymbol{\beta}_1^* - \widetilde{\boldsymbol{\beta}}_1\|_2^2. \tag{3.13}$$

Taking $M = 2\sqrt{s/n} + \sqrt{s}\lambda_1 + 2\lambda_2\|\boldsymbol{\beta}_1^*\|_2$, by Condition 3.2 and the assumption $(\lambda_1\sqrt{s} + 2\lambda_2\|\boldsymbol{\beta}_1^*\|_2)\sqrt{s}\kappa_n \to 0$, we can check that $M = o(\kappa_n^{-1}s^{-1/2})$. Combining these two results with (3.7), we obtain that on the event Γ_n,

$$\begin{aligned}
&\frac{1}{3}\left(\lambda_l + 3\lambda_2\right) n\|\widetilde{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_1^*\|_2^2 \\
&\qquad \le \left(2\sqrt{sn(\log n)} + n\sqrt{s}\lambda_1 + 2n\lambda_2\|\boldsymbol{\beta}_1^*\|_2\right)\left(2\sqrt{s/n} + \sqrt{s}\lambda_1 + 2\lambda_2\|\boldsymbol{\beta}_1^*\|_2\right),
\end{aligned} \tag{3.14}$$

hence

$$\|\widetilde{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_1^*\|_2^2 \le \frac{3}{\lambda_l + 3\lambda_2}\left(2\sqrt{s(\log n)/n} + \sqrt{s}\lambda_1 + 2\lambda_2\|\boldsymbol{\beta}_1^*\|_2\right)\left(2\sqrt{s/n} + \sqrt{s}\lambda_1 + 2\lambda_2\|\boldsymbol{\beta}_1^*\|_2\right), \tag{3.15}$$

which entails that

$$\|\boldsymbol{\beta}_1^* - \widetilde{\boldsymbol{\beta}}_1\|_2 \le O\left(\sqrt{s}\lambda_1 + 2\lambda_2\|\boldsymbol{\beta}_1^*\|_2 + 2\sqrt{s(\log n)/n}\right). \tag{3.16}$$

Note that $\|\boldsymbol{\beta}_1^* - \widetilde{\boldsymbol{\beta}}_1\|_2 \le M$ implies $\|\boldsymbol{\beta}_1^0 - \boldsymbol{\beta}_1^*\|_2 \le 2M$. Thus, on the event Γ_n, we have

$$\|\boldsymbol{\beta}_1^0 - \boldsymbol{\beta}_1^*\|_2 \le O\left(\sqrt{s}\lambda_1 + 2\lambda_2\|\boldsymbol{\beta}_1^*\|_2 + 2\sqrt{s(\log n)/n}\right). \tag{3.17}$$

Thus $\Pr(\|\boldsymbol{\beta}_1^0 - \boldsymbol{\beta}_1^*\|_2 \le \gamma_n) \ge 1 - n^{-cs}$ holds. From above analysis the second result follows immediately. □

As shown in Theorem 3.1, the consistency rate of $\boldsymbol{\beta}_1^0$ in terms of the l_2-norm is given by γ_n. The first component of $\gamma_n, C_1\sqrt{s(\log n)/n}$, is the oracle rate within a factor of $\log n$, and the second component $C_1(\sqrt{s}\lambda_1 + 2\lambda_2\|\boldsymbol{\beta}_1^*\|_2)$ reflects the bias due to penalization.

Though the ORE has the consistent property which helps us to understand the utility of the oracle information on finding the locations of signal covariates. However locations of signals are unknown. An important question is that whether the SQEN estimator performs well when there is no oracle information. The following theorem will give an answer, which shows that the SQEN estimator enjoys the same property as ORE with probability tending to one when λ_1 and λ_2 are appropriately chosen.

Theorem 3.2. *Let $\gamma_n = C_1(2\sqrt{s(\log n)/n}+\sqrt{s}\lambda_1+2\lambda_2\|\boldsymbol{\beta}_1^*\|_2)$ with $C_1 > 0$ a constant. Suppose Conditions 3.1-3.3 hold. In addition, assume that $\gamma_n s^{3/2}\kappa_n^2(\log_2 n)^2 = o(n\lambda_1^2), \lambda_1 s\kappa_n \to 0, (\sqrt{s}\lambda_1 + 2\lambda_2\|\boldsymbol{\beta}_1^*\|_2)\sqrt{s}\kappa_n \to 0$ and $\lambda_1 > 2\sqrt{(1+c)(\log p)/n}$ where c is some positive constant. Then, there exists a global minimizer $\widehat{\boldsymbol{\beta}} = (\boldsymbol{\beta}_1^{0^T}, \widehat{\boldsymbol{\beta}}_2^T)^T$ of $\Phi_\tau(\boldsymbol{\beta}, \delta)$ which satisfies*

(1) $\widehat{\boldsymbol{\beta}}_2 = 0$,

(2) $\|\boldsymbol{\beta}_1^0 - \boldsymbol{\beta}_1^*\|_2 \le \gamma_n$,

with probability at least $1 - O(n^{-cs})$.

In order to prove Theorem 3.2, we need the following lemma.

Lemma 3.3. *Consider a ball in $\mathbb{R}^s$ around $\boldsymbol{\beta}^*$, $\mathcal{N} = \{\boldsymbol{\beta} = (\boldsymbol{\beta}_1^T, \boldsymbol{\beta}_2^T)^T \in \mathbb{R}^p : \boldsymbol{\beta}_2 = \mathbf{0}, \|\boldsymbol{\beta}_1 - \boldsymbol{\beta}_1^*\|_2 \le \gamma_n\}$ with some sequence $\gamma_n \to 0$. Assume that $\sqrt{1+\gamma_n s^{3/2}\kappa_n^2}\log_2 n = o(\sqrt{n}\lambda_1), \sqrt{n}\lambda_1(\log p)^{-1/2} \to \infty$, and $\kappa_n^2\gamma_n = o(\lambda_1)$. Then under Conditions 3.1-3.3, there exists a constant $c > 0$ such that*

$$\Pr\left(\sup_{\boldsymbol{\beta}\in\mathcal{B}_0(M)} \|Q^T\rho'_\tau(y - S\boldsymbol{\beta}_1, \delta)\|_\infty \ge n\lambda_1\right) \le O(p^{-c}), \tag{3.18}$$

where

$$\rho'_\tau(u,\delta) = \begin{cases} \tau - 1, & u \in (-\infty, (\tau-1)\delta), \\ \frac{u}{\delta}, & u \in [(\tau-1)\delta, \tau\delta], \\ \tau, & u \in (\tau\delta, +\infty). \end{cases} \tag{3.19}$$

Proof of Lemma 3.3. For a fixed $j \in \{s+1, \ldots, p\}$ and $\boldsymbol{\beta} = (\boldsymbol{\beta}_1^{\mathrm{T}}, \boldsymbol{\beta}_2^{\mathrm{T}})^{\mathrm{T}} \in \mathcal{N}$, define

$$\gamma_{\boldsymbol{\beta},j}(\mathbf{x}_i, y_i) = x_{ij}\left[\rho'_\tau(y_i - \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}, \delta) - \rho'_\tau(\varepsilon_i, \delta) - \mathbb{E}\left[\rho'_\tau(y_i - \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}, \delta) - \rho'_\tau(\varepsilon_i, \delta)\right]\right], \tag{3.20}$$

where $\mathbf{x}_i^{\mathrm{T}} = (x_{i1}, \ldots, x_{ip})$ is the i-th row of the design matrix X. In order to prove this lemma we will do the following decomposition:

$$\begin{aligned} \sup_{\boldsymbol{\beta}\in\mathcal{N}} \left\|\frac{1}{n}Q^{\mathrm{T}}\rho'_\tau(\mathbf{y} - S\boldsymbol{\beta}_1)\right\|_\infty &\le \sup_{\boldsymbol{\beta}\in\mathcal{N}} \left\|\frac{1}{n}Q^{\mathrm{T}}\mathbb{E}[\rho'_\tau(\mathbf{y} - S\boldsymbol{\beta}_1) - \rho'_\tau(\varepsilon_i, \delta)]\right\|_\infty \\ &\quad + \left\|\frac{1}{n}Q^{\mathrm{T}}\rho'_\tau(\varepsilon_i, \delta)\right\|_\infty + \max_{j>s}\sup_{\boldsymbol{\beta}\in\mathcal{N}}\frac{1}{n}\sum_{i=1}^n |\gamma_{\boldsymbol{\beta},j}(\mathbf{x}_i, y_i)| \\ &= I_1 + I_2 + I_3, \end{aligned}$$

where

$$I_1 \equiv \sup_{\boldsymbol{\beta}\in\mathcal{N}} \left\|\frac{1}{n}Q^{\mathrm{T}}\mathbb{E}[\rho'_\tau(\mathbf{y} - S\boldsymbol{\beta}_1) - \rho'_\tau(\varepsilon_i, \delta)]\right\|_\infty, \tag{3.21}$$

$$I_2 \equiv \left\| \frac{1}{n} Q^{\mathrm{T}} \rho'_\tau(\varepsilon_i, \delta) \right\|_\infty, \tag{3.22}$$

$$I_3 \equiv \max_{j>s} \sup_{\boldsymbol{\beta}\in\mathcal{N}} \frac{1}{n} \sum_{i=1}^{n} |\gamma_{\boldsymbol{\beta},j}(\mathbf{x}_i, y_i)|. \tag{3.23}$$

If we can prove that $I_1 < \lambda_1/2 + o(\lambda_1), I_2 = o(\lambda_1), I_3 = o_p(\lambda_1)$ with probability at least $1 - O(p^{-c})$, then the lemma will be proved. Now we proceed to prove (3.21). Note that I_1 can be rewritten as

$$I_1 = \max_{j>s} \sup_{\boldsymbol{\beta}\in\mathcal{N}} \left| \frac{1}{n} \sum_{i=1}^{n} x_{ij} \mathbb{E}\left[\rho'_\tau(\varepsilon_i, \delta) - \rho'_\tau(y_i - \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta})\right] \right|. \tag{3.24}$$

Note that

$$\begin{aligned}
\rho'_\tau(\varepsilon_i, \delta) &= \tau \cdot \mathbf{1}\{\varepsilon_i > \tau\delta\} + (\tau - 1) \cdot \mathbf{1}\{\varepsilon_i < (\tau - 1)\delta\} + \frac{\varepsilon_i}{\delta} \cdot \mathbf{1}\{(\tau - 1)\delta \le \varepsilon_i \le \tau\delta\} \\
&= \tau \cdot \mathbf{1}\{\varepsilon_i > \tau\delta\} + (\tau - 1) \cdot \mathbf{1}\{\varepsilon_i < (\tau - 1)\delta\} \\
&\quad + \int_{\tau-1}^{\tau} \mathbf{1}\{s\delta \le \varepsilon_i \le \tau\delta\} \mathrm{d}s + (\tau - 1) \cdot \mathbf{1}\{(\tau - 1)\delta \le \varepsilon_i \le \tau\delta\},
\end{aligned}$$

and

$$\begin{aligned}
\rho'_\tau(y_i - \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}) &= \tau \cdot \mathbf{1}\{y_i - \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta} > \tau\delta\} + (\tau - 1) \cdot \mathbf{1}\{y_i - \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta} < (\tau - 1)\delta\} \\
&\quad + \frac{y_i - \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}}{\delta} \cdot \mathbf{1}\{(\tau - 1)\delta \le y_i - \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta} \le \tau\delta\} \\
&= \tau \cdot \mathbf{1}\{\varepsilon_i > \tau\delta + \mathbf{S}_i^{\mathrm{T}}(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_1^*)\} + (\tau - 1) \cdot \mathbf{1}\{\varepsilon_i < (\tau - 1)\delta + \mathbf{S}_i^{\mathrm{T}}(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_1^*)\} \\
&\quad + \frac{\varepsilon_i - \mathbf{S}_i^{\mathrm{T}}(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_1^*)}{\delta} \cdot \mathbf{1}\{(\tau - 1)\delta \le \varepsilon_i - \mathbf{S}_i^{\mathrm{T}}(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_1^*) \le \tau\delta\} \\
&= \tau \cdot \mathbf{1}\{\varepsilon_i > \tau\delta + \mathbf{S}_i^{\mathrm{T}}(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_1^*)\} + (\tau - 1) \cdot \mathbf{1}\{\varepsilon_i < (\tau - 1)\delta + S_i^{\mathrm{T}}(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_1^*)\} \\
&\quad + \int_{\tau-1}^{\tau} \mathbf{1}\{s\delta + \mathbf{S}_i^{\mathrm{T}}(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_1^*) \le \varepsilon_i \le \tau\delta + \mathbf{S}_i^{\mathrm{T}}(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_1^*)\} \mathrm{d}s \\
&\quad + (\tau - 1) \cdot \mathbf{1}\{(\tau - 1)\delta + \mathbf{S}_i^{\mathrm{T}}(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_1^*) \le \varepsilon_i \le \tau\delta + \mathbf{S}_i^{\mathrm{T}}(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_1^*)\}.
\end{aligned}$$

Hence we have that

$$\begin{aligned}
\mathbb{E}[\rho'_\tau(\varepsilon_i, \delta)] &= \tau(1 - F_i(\tau\delta)) + (\tau - 1)F_i((\tau - 1)\delta) + \int_{\tau-1}^{\tau} [F_i(\tau\delta) - F_i(s\delta)] \, \mathrm{d}s \\
&\quad + (\tau - 1)(F_i(\tau\delta) - F_i((\tau - 1)\delta)) \\
&= \tau - \int_{\tau-1}^{\tau} F_i(s\delta) \mathrm{d}s.
\end{aligned}$$

where $F_i(t)$ is the cumulative distribution function of ε_i. Moreover,

$$\begin{aligned}
\mathbb{E}[\rho'_\tau(y_i - \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta})] &= \tau(1 - F_i(\tau\delta + \mathbf{S}_i^{\mathrm{T}}(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_1^*))) + (\tau - 1)F_i((\tau - 1)\delta + \mathbf{S}_i^{\mathrm{T}}(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_1^*)) \\
&\quad + \int_{\tau-1}^{\tau} [F_i(\tau\delta + \mathbf{S}_i^{\mathrm{T}}(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_1^*)) - F_i(s\delta + \mathbf{S}_i^{\mathrm{T}}(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_1^*))] \, \mathrm{d}s \\
&\quad + (\tau - 1)(F_i(\tau\delta + \mathbf{S}_i^{\mathrm{T}}(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_1^*)) - F_i((\tau - 1)\delta + \mathbf{S}_i^{\mathrm{T}}(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_1^*)))
\end{aligned}$$

$$= \tau - \int_{\tau-1}^{\tau} F_i(s\delta + \mathbf{S}_i^{\mathrm{T}}(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_1^*))\mathrm{d}s.$$

Thus

$$\begin{aligned}
\mathbb{E}\left[\rho'_\tau(\varepsilon_i, \delta) - \rho'_\tau(y_i - \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta})\right] &= \left(\tau - \int_{\tau-1}^{\tau} F_i(s\delta)\mathrm{d}s\right) - \left(\tau - \int_{\tau-1}^{\tau} F_i(s\delta + \mathbf{S}_i^{\mathrm{T}}(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_1^*))\mathrm{d}s\right) \\
&= \int_{\tau-1}^{\tau} F_i(s\delta + \mathbf{S}_i^{\mathrm{T}}(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_1^*)) - F_i(s\delta)\mathrm{d}s \\
&= \int_{\tau-1}^{\tau} \left[\mathbf{S}_i^{\mathrm{T}}(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_1^*) f_i(0) + \mathbf{S}_i^{\mathrm{T}}(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_1^*) o(1)\right] \mathrm{d}s \\
&= \mathbf{S}_i^{\mathrm{T}}(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_1^*) f_i(0) + \mathbf{S}_i^{\mathrm{T}}(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_1^*) o(1).
\end{aligned}$$

Thus for any $j > s$,

$$\sum_{i=1}^{n} x_{ij}\mathbb{E}\left[\rho'_\tau(\varepsilon_i, \delta) - \rho'_\tau(y_i - \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta})\right] = \sum_{i=1}^{n} x_{ij}\mathbf{S}_i^{\mathrm{T}}(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_1^*) f_i(0) + \sum_{i=1}^{n} x_{ij}\mathbf{S}_i^{\mathrm{T}}(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_1^*) o(1). \tag{3.25}$$

This together with (3.24) and Cauchy-Schwarz inequality entails that

$$I_1 \leq \left\| \frac{1}{n} Q^{\mathrm{T}} H S(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_1^*) \right\|_\infty + \max_{j>s} \left| \sum_{i=1}^{n} x_{ij} o(1) \mathbf{S}_i^{\mathrm{T}}(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_1^*) \right|, \tag{3.26}$$

where $H = \mathrm{diag}(f_1(0), \ldots, f_n(0))$. We consider the two terms on the right-hand side of (3.26) one by one. By Condition 3.3, the first term can be bounded as

$$\left\| \frac{1}{n} Q^{\mathrm{T}} H S(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_1^*) \right\|_\infty \leq \left\| \frac{1}{n} Q^{\mathrm{T}} H S \right\|_{2,\infty} \|\boldsymbol{\beta}_1 - \boldsymbol{\beta}_1^*\|_2 < \lambda_1/2. \tag{3.27}$$

By Condition 3.2 the second term of (3.26) can be bounded as

$$\max_{j>s} \left| \sum_{i=1}^{n} x_{ij} \mathbf{S}_i^{\mathrm{T}}(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_1^*) o(1) \right| \leq \kappa_n \sum_{i=1}^{n} |\mathbf{S}_i^{\mathrm{T}}(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_1^*) o(1)| \leq n\sqrt{p-s}\kappa_n^2 \|\boldsymbol{\beta}_1 - \boldsymbol{\beta}_1^*\|_2 o(1).$$

Since $\boldsymbol{\beta} \in \mathcal{N}$, it follows from the assumption $\kappa_n^2 \gamma_n = o(\lambda_1)$ that

$$\max_{j>s} \left| \sum_{i=1}^{n} x_{ij} \mathbf{S}_i^{\mathrm{T}}(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_1^*) o(1) \right| \leq C\kappa_n \gamma_n^2 o(1) = o(\lambda_1). \tag{3.28}$$

Plugging the above inequality and (3.27) into (3.26) completes the proof of (3.21).

By Hoeffiding's inequality, if $\lambda_1 > 2\sqrt{(1+c)(\log p)/n}$ with c is some positive constant, then

$$\begin{aligned}
\Pr\left(\|Q^{\mathrm{T}} \rho'_\tau(\varepsilon_i, \delta)\|_\infty \geq n\lambda_1\right) &\leq \sum_{j=s+1}^{p} 2\exp\left(-\frac{n^2\lambda_1^2}{4\sum_{i=1}^{n} x_{ij}^2}\right) \\
&= 2\exp\left(\log p - s - n\lambda_1^2/4\right) \\
&= O(p^{-c}).
\end{aligned}$$

Thus, (3.22) holds with probability at least $1 - O(p^{-c})$.

We now apply Corollary 14.4 in [3] to prove (3.23). To the end, we need to check three conditions of this corollary. For each fixed j, define the function space $\Gamma_j = \{\gamma_{\boldsymbol{\beta},j} : \boldsymbol{\beta} \in \mathcal{N}\}$. From the expression of $\gamma_{\boldsymbol{\beta},j}$, $\mathbb{E}[\gamma_{\boldsymbol{\beta},j}(\mathbf{x}_i, y_i)] = 0$ for any $\gamma_{\boldsymbol{\beta},j} \in \Gamma_j$. Since $\rho'_\tau(\cdot, \delta)$ is bounded, we have

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n}\gamma^2_{\boldsymbol{\beta},j}(\mathbf{x}_i, y_i) &= \frac{1}{n}\sum_{i=1}^{n} x_{ij}^2 \left(\rho'_\tau(y_i - \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}, \delta) - \rho'_\tau(\varepsilon_i, \delta) - \mathbb{E}\left[\rho'_\tau(y_i - \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}, \delta) - \rho'_\tau(\varepsilon_i, \delta)\right]\right)^2 \\ &\le 4.\end{aligned}$$

Thus $\|\gamma_{\boldsymbol{\beta},j}\|_n := \left(\frac{1}{n}\sum_{i=1}^{n}\gamma^2_{\boldsymbol{\beta},j}(x_i, y_i)\right)^{1/2} \le 2$.

We will calculate the covering number of the functional space Γ_j, i.e., $N(\cdot, \Gamma_j, \|\cdot\|_2)$. For any $\boldsymbol{\beta} = (\boldsymbol{\beta}_1^{\mathrm{T}}, \boldsymbol{\beta}_2^{\mathrm{T}})^{\mathrm{T}} \in \mathcal{N}$ and $\widetilde{\boldsymbol{\beta}} = (\widetilde{\boldsymbol{\beta}}_1^{\mathrm{T}}, \widetilde{\boldsymbol{\beta}}_2^{\mathrm{T}})^{\mathrm{T}} \in \mathcal{N}$, by Condition 3.1 and the mean value theorem,

$$\begin{aligned}&\mathbb{E}[\rho'_\tau(y_i - \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}, \delta) - \rho'_\tau(\varepsilon_i, \delta)] - \mathbb{E}[\rho'_\tau(y_i - \mathbf{x}_i^{\mathrm{T}}\widetilde{\boldsymbol{\beta}}, \delta) - \rho'_\tau(\varepsilon_i, \delta)] \\ &\quad= \int_{\tau-1}^{\tau}[F_i(s\delta) - F_i(s\delta + \mathbf{S}_i^{\mathrm{T}}(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_1^*))]\,\mathrm{d}s \\ &\qquad- \int_{\tau-1}^{\tau}\left[F_i(s\delta) - F_i(s\delta + \mathbf{S}_i^{\mathrm{T}}(\widetilde{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_1^*))\right]\mathrm{d}s \\ &\quad= \int_{\tau-1}^{\tau}\left[F_i(s\delta + \mathbf{S}_i^{\mathrm{T}}(\widetilde{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_1^*)) - F_i(s\delta + \mathbf{S}_i^{\mathrm{T}}(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_1^*))\right]\mathrm{d}s \\ &\quad= \int_{\tau-1}^{\tau} f_i(a_{1i})\mathbf{S}_i^{\mathrm{T}}(\widetilde{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_1)\mathrm{d}s \\ &\quad= f_i(a_{1i})\mathbf{S}_i^{\mathrm{T}}(\widetilde{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_1), \qquad (3.29)\end{aligned}$$

where a_{1i} lies on the segment connecting $\mathbf{S}_i^{\mathrm{T}}(\widetilde{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_1^*)$ and $\mathbf{S}_i^{\mathrm{T}}(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_1^*)$. Since $f_i(u)$s are uniformly bounded and $\kappa_n = \max_{ij}|x_{ij}|$, we have

$$\begin{aligned}&\left|x_{ij}\mathbb{E}\left[\rho'_\tau(y_i - \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}, \delta) - \rho'_\tau(\varepsilon_i, \delta)\right] - x_{ij}\mathbb{E}\left[\rho'_\tau(y_i - \mathbf{x}_i^{\mathrm{T}}\widetilde{\boldsymbol{\beta}}, \delta) - \rho'_\tau(\varepsilon_i, \delta)\right]\right| \\ &\qquad\le C\left|x_{ij}\mathbf{S}_i^{\mathrm{T}}(\boldsymbol{\beta}_1 - \widetilde{\boldsymbol{\beta}}_1)\right| \le C\|x_{ij}\mathbf{S}_i\|_2\|\boldsymbol{\beta}_1 - \widetilde{\boldsymbol{\beta}}_1\|_2 \\ &\qquad\le C\sqrt{s}\kappa_n^2\|\boldsymbol{\beta}_1 - \widetilde{\boldsymbol{\beta}}_1\|_2 \qquad (3.30)\end{aligned}$$

where C is a positive constant. Considering the definition of $\mathcal{N}$, it is easy to know that for any $\boldsymbol{\beta} = (\boldsymbol{\beta}_1^{\mathrm{T}}, \boldsymbol{\beta}_2^{\mathrm{T}})^{\mathrm{T}} \in \mathcal{N}$, we have $\|\boldsymbol{\beta}\|_2 \le 2\gamma_n$. Then from Lemma 14.27 in [3], $\mathcal{N}$ in $\mathbb{R}^s$ can be covered by $(1 + 4\gamma_n/\delta)^s$ balls with radius $0 < \delta \le 2\gamma_n$. It follows from (3.30) that the 2^{1-k}-covering number of Γ_j is $N\left(2^{2-k}, \Gamma_j, \|\cdot\|_2\right) \le 7\left(1 + C^{-1}2^k\gamma_n\kappa_n^2\sqrt{s}\right)^s$, where $0 \le k \le \frac{\log_2 n}{2}$. Thus, after some simple calculations, we obtain

$$\begin{aligned}&\log\left(1 + N\left(2^{2-k}, \Gamma_j, \|\cdot\|_2\right)\right) \\ &\qquad\le \log 14 + s\log\left(1 + C^{-1}2^k\gamma_n\kappa_n^2\sqrt{s}\right) \\ &\qquad\le \log 14 + C^{-1}2^k\gamma_n s^{3/2}\kappa_n^2 \\ &\qquad\le 4\left(1 + C^{-1}\gamma_n s^{3/2}\kappa_n^2\right)2^{2k}.\end{aligned}$$

Hence, conditions of Corollary 14.4 in [3] are checked and we obtain that for any $t > 0$,

$$\Pr\left(\sup_{\boldsymbol{\beta}\in\mathcal{N}}\left|\frac{1}{n}\sum_{i=1}^{n}\gamma_{\boldsymbol{\beta},j}(\mathbf{x}_i,y_i)\right| \geq \frac{8}{\sqrt{n}}\left(6\sqrt{1+C^{-1}\gamma_n s^{3/2}\kappa_n^2}\log_2 n + 4 + 4t\right)\right) \leq 4\exp\left(-\frac{nt^2}{8}\right).$$

Taking $t = \sqrt{C(\log p)/n}$ with $C > 0$ large enough constant, we can get that

$$\Pr\left(\max_{j>s}\sup_{\boldsymbol{\beta}\in\mathcal{N}}\left|\frac{1}{n}\sum_{i=1}^{n}\gamma_{\boldsymbol{\beta},j}(\mathbf{x}_i,y_i)\right| \geq \frac{48}{\sqrt{n}}\sqrt{1+C^{-1}\gamma_n s^{3/2}\kappa_n^2}\log_2 n\right) \leq 4(p-s)\exp\left(-\frac{C\log p}{8}\right) \to 0.$$

Hence, if $\sqrt{1+\gamma_n s^{3/2}\kappa_n^2}\log_2 n = o(\sqrt{n}\lambda_1)$, then (3.23) holds with probability at least $1 - o(p^{-c})$. Now, the proof of this lemma is completed.

Based on the above lemma, we can complete the proof of Theorem 3.2.

Proof of Theorem 3.2. Since $\widehat{\boldsymbol{\beta}} = (\boldsymbol{\beta}_1^{0\mathrm{T}}, \mathbf{0}^{\mathrm{T}})^{\mathrm{T}} \in \mathbb{R}^p$ satisfies the KKT condition, then $\widehat{\boldsymbol{\beta}}$ is a minimizer of $\Phi_\tau(\boldsymbol{\beta}, \delta)$. To prove Theorem 3.2, we only need to verify the following condition

$$\left\|Q^{\mathrm{T}}\rho_\tau'(\mathbf{y} - S\boldsymbol{\beta}_1^0, \delta)\right\|_\infty < n\lambda_1, \tag{3.31}$$

where $\rho_\tau'(\mathbf{u}, \delta) = (\rho_\tau'(u_1, \delta), \ldots, \rho_\tau'(u_n, \delta))^{\mathrm{T}}$ for any vector $\mathbf{u} = (u_1, \ldots, u_n)^{\mathrm{T}} \in \mathbb{R}^n$ with $\rho_\tau'(u_i, \delta)$ defined as before. The KKT condition and the strong convexity of $\Phi_\tau(\boldsymbol{\beta}, \delta)$ ensure that $\widehat{\boldsymbol{\beta}}$ is a global minimizer of $\Phi_\tau(\boldsymbol{\beta}, \delta)$. Define events

$$A_1 = \left\{\|\boldsymbol{\beta}_1^0 - \boldsymbol{\beta}_1^*\|_2 \leq \gamma_n\right\},\ A_2 = \left\{\sup_{\boldsymbol{\beta}\in\mathcal{N}}\|Q^{\mathrm{T}}\rho_\tau'(\mathbf{y} - S\boldsymbol{\beta}_1^0, \delta)\|_\infty < n\lambda_1\right\}, \tag{3.32}$$

where γ_n is defined in Theorem 3.1 and

$$\mathcal{N} = \left\{\boldsymbol{\beta} = (\boldsymbol{\beta}_1^{\mathrm{T}}, \boldsymbol{\beta}_2^{\mathrm{T}})^{\mathrm{T}} \in \mathbb{R}^p : \boldsymbol{\beta}_2 = 0, \|\boldsymbol{\beta}_1 - \boldsymbol{\beta}_1^*\|_2 \leq \gamma_n\right\}. \tag{3.33}$$

Then by Theorem 3.1 and Lemma 3.3, $\Pr(A_1 \cap A_2) \geq 1 - O(n^{-cs})$. Since $\widehat{\boldsymbol{\beta}} \in \mathcal{N}$ on the event A_1, the inequality (3.31) holds on the event $A_1 \cap A_2$. Now the proof of Theorem 3.2 is completed. □

Until now, we have established model selection consistent property of the SQEN estimator. It ensures a good statistical behavior. In the next section we will consider how to solve the optimization problem (2.3) which is another crucial issue.

4 Optimization Algorithm

In this section, we will construct an efficient algorithm based on the Majorize-Minimize (MM) technique. Moreover, we will prove the convergence of this algorithm. Finally, we will show that the solution sequence generated by this algorithm converges to the optimal minimizer of (2.3).

In order to obtain the optimal solution of the model(2.3), we will take advantage of majorized function of $S_\tau(\boldsymbol{\beta},\delta)$, which is defined as

$$M_{\tau,\xi}(\boldsymbol{\beta},\boldsymbol{\alpha},\delta)=S_\tau(\boldsymbol{\alpha},\delta)+\langle\nabla_{\boldsymbol{\beta}}S_\tau(\boldsymbol{\alpha},\delta),\boldsymbol{\beta}-\boldsymbol{\alpha}\rangle+\frac{L_\delta+\xi}{2}\|\boldsymbol{\beta}-\boldsymbol{\alpha}\|_2^2,$$

where $\xi>0$ is sufficiently small, $\nabla_{\boldsymbol{\beta}}S_\tau(\boldsymbol{\alpha},\delta)=\sum_{i=1}^n -x_i\rho'_\tau(y_i-\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\alpha},\delta)$ (if it is not ambiguous, we shortly write$\nabla S_\tau(\boldsymbol{\alpha},\delta)=\nabla_{\boldsymbol{\beta}}S_\tau(\boldsymbol{\alpha},\delta)$), $\boldsymbol{\alpha}\in\mathbb{R}^p$ and $L_\delta=\frac{1}{\delta}\lambda_{\max}(X^{\mathrm{T}}X)$. Thus, for any $\boldsymbol{\beta}\neq\boldsymbol{\alpha}\in\mathbb{R}^p$, we have

$$M_{\tau,\xi}(\boldsymbol{\beta},\boldsymbol{\alpha},\delta)>S_\tau(\boldsymbol{\beta},\delta)\ \text{ and }\ M_{\tau,\xi}(\boldsymbol{\beta},\boldsymbol{\beta},\delta)=S_\tau(\boldsymbol{\beta},\delta).\tag{4.1}$$

By virtue of $M_{\tau,\xi}(\boldsymbol{\beta},\boldsymbol{\alpha},\delta)$, we consider the following optimization problem

$$\boldsymbol{\beta}^{k+1}=\underset{\boldsymbol{\beta}\in\mathbb{R}^p}{\operatorname{argmin}}\left\{M_{\tau,\xi}(\boldsymbol{\beta},\boldsymbol{\beta}^k,\delta)+n\lambda_1\|\boldsymbol{\beta}\|_1+n\lambda_2\|\boldsymbol{\beta}\|_2^2\right\},\tag{4.2}$$

where $\boldsymbol{\beta}^0$ is a given initial point. The following, we will show that we can solve this optimization problem in stead of (2.3). Namely, we need to demonstrate $\boldsymbol{\beta}^{k+1}\to\widehat{\boldsymbol{\beta}}$ and $\widehat{\boldsymbol{\beta}}$ is the optimal minimizer of (2.3). Now, let's show this step by step. First lets give the analytic expression of $\boldsymbol{\beta}^{k+1}$. The following theorem will show this result.

Theorem 4.1. *If $\widehat{\boldsymbol{\beta}}$ is a global minimizer of $M_{\tau,\xi}(\boldsymbol{\beta},\boldsymbol{\alpha},\delta)+n\lambda_1\|\boldsymbol{\beta}\|_1+n\lambda_2\|\boldsymbol{\beta}\|_2^2$ for any fixed $\tau\in(0,1),\delta,\xi,\ \lambda_1,\lambda_2>0,\boldsymbol{\alpha}\in\mathbb{R}^p$ and $L_\delta=\frac{1}{\delta}\lambda_{max}(X^TX)$, then $\widehat{\boldsymbol{\beta}}$ can be analytically expressed by*

$$\widehat{\boldsymbol{\beta}}=\frac{L_\delta+\xi}{L_\delta+\xi+2n\lambda_2}\operatorname{sign}(\widetilde{\boldsymbol{\beta}})\circ\max\left\{|\widetilde{\boldsymbol{\beta}}|-\frac{n\lambda_1}{L_\delta+\xi},0\right\},\tag{4.3}$$

where $\widetilde{\boldsymbol{\beta}}=\boldsymbol{\alpha}-\nabla S_\tau(\boldsymbol{\alpha},\delta)/(L_\delta+\xi)$, "$\circ$" is the Harmard product of two vectors, and $|\widetilde{\boldsymbol{\beta}}|$ represents the vector as the same dimension of $\widetilde{\boldsymbol{\beta}}$ and its components are the absolute value of corresponding one of $\widetilde{\boldsymbol{\beta}}$.

Proof. For the definition of $\widehat{\boldsymbol{\beta}}$, we have

$$\begin{aligned}
\widehat{\boldsymbol{\beta}}&=\underset{\boldsymbol{\beta}\in\mathbb{R}^p}{\operatorname{argmin}}\left\{M_{\tau,\xi}(\boldsymbol{\beta},\boldsymbol{\alpha},\delta)+n\lambda_1\|\boldsymbol{\beta}\|_1+n\lambda_2\|\boldsymbol{\beta}\|_2^2\right\}\\
&=\underset{\boldsymbol{\beta}\in\mathbb{R}^p}{\operatorname{argmin}}\left\{S_\tau(\boldsymbol{\alpha},\delta)+\langle\nabla_{\boldsymbol{\beta}}S_\tau(\boldsymbol{\alpha},\delta),\boldsymbol{\beta}-\boldsymbol{\alpha}\rangle+\frac{L_\delta+\xi}{2}\|\boldsymbol{\beta}-\boldsymbol{\alpha}\|_2^2+n\lambda_1\|\boldsymbol{\beta}\|_1+n\lambda_2\|\boldsymbol{\beta}\|_2^2\right\}\\
&=\underset{\boldsymbol{\beta}\in\mathbb{R}^p}{\operatorname{argmin}}\left\{\langle\nabla_{\boldsymbol{\beta}}S_\tau(\boldsymbol{\alpha},\delta),\boldsymbol{\beta}-\boldsymbol{\alpha}\rangle+\frac{L_\delta+\xi}{2}\|\boldsymbol{\beta}-\boldsymbol{\alpha}\|_2^2+n\lambda_1\|\boldsymbol{\beta}\|_1+n\lambda_2\|\boldsymbol{\beta}\|_2^2\right\}\\
&=\underset{\boldsymbol{\beta}\in\mathbb{R}^p}{\operatorname{argmin}}\left\{\frac{L_\delta+\xi}{2}\|\boldsymbol{\beta}-\widetilde{\boldsymbol{\beta}}\|_2^2+n\lambda_1\|\boldsymbol{\beta}\|_1+n\lambda_2\|\boldsymbol{\beta}\|_2^2\right\},
\end{aligned}$$

here $\widetilde{\boldsymbol{\beta}}=\boldsymbol{\alpha}-\nabla S_\tau(\boldsymbol{\alpha},\delta)/(L_\delta+\xi)$. Hence, for any $i\in\{1,2,\ldots,p\}$,

$$\widehat{\beta}_i=\underset{\beta_i\in\mathbb{R}}{\operatorname{argmin}}\left\{\frac{L_\delta+\xi}{2}(\beta_i-\widetilde{\beta}_i)^2+n\lambda_1|\beta_i|+n\lambda_2\beta_i^2\right\}:=f(\beta_i).\tag{4.4}$$

Denoting $\Delta:=L_\delta+\xi$, then $f(\beta_i)$ will be

$$f(\beta_i)=\begin{cases}
\left(\frac{\Delta}{2}+n\lambda_2\right)\left(\beta_i-\frac{\Delta\widetilde{\beta}_i-n\lambda_1}{\Delta+2n\lambda_2}\right)^2+\frac{\Delta}{2}\widetilde{\beta}_i^2-\left(\frac{\Delta}{2}+n\lambda_2\right)\left(\frac{\Delta\widetilde{\beta}_i-n\lambda_1}{\Delta+2n\lambda_2}\right)^2, & \beta_i>0,\\
\frac{\Delta}{2}\widetilde{\beta}_i^2, & \beta_i=0,\\
\left(\frac{\Delta}{2}+n\lambda_2\right)\left(\beta_i-\frac{\Delta\widetilde{\beta}_i+n\lambda_1}{\Delta+2n\lambda_2}\right)^2+\frac{\Delta}{2}\widetilde{\beta}_i^2-\left(\frac{\Delta}{2}+n\lambda_2\right)\left(\frac{\Delta\widetilde{\beta}_i+n\lambda_1}{\Delta+2n\lambda_2}\right)^2, & \beta_i<0.
\end{cases}$$

In order to get $\widehat{\beta}_i$, we will discuss in two cases.

Case 1: If $\frac{\Delta\widetilde{\beta}_i - n\lambda_1}{\Delta + 2n\lambda_2} > 0$, then $\Delta\widetilde{\beta}_i - n\lambda_1 \geq 0$. It follows that $\widetilde{\beta}_i \geq 0$ and further more $\frac{\Delta\widetilde{\beta}_i + n\lambda_1}{\Delta + 2n\lambda_2} > 0$.

Therefore, when $\beta_i > 0$, the minimum point of $f(\beta_i)$ is $\frac{\Delta\widetilde{\beta}_i - n\lambda_1}{\Delta + 2n\lambda_2}$, and the objective function value is $\frac{\Delta}{2}\widetilde{\beta}_i^2 - \left(\frac{\Delta}{2} + n\lambda_2\right)\left(\frac{\Delta\widetilde{\beta}_i - n\lambda_1}{\Delta + 2n\lambda_2}\right)^2$. While when $\beta_i \leq 0$, the minimum point of $f(\beta_i)$ is 0, and the objective function value is $\frac{\Delta}{2}\Delta\widetilde{\beta}_i^2$. Considering

$$\frac{\Delta}{2}\widetilde{\beta}_i^2 > \frac{\Delta}{2}\widetilde{\beta}_i^2 - \left(\frac{\Delta}{2} + n\lambda_2\right)\left(\frac{\Delta\widetilde{\beta}_i - n\lambda_1}{\Delta + 2n\lambda_2}\right)^2, \tag{4.5}$$

hence the global optimal solution of (4.4) is $\widehat{\beta}_i = \frac{\Delta\widetilde{\beta}_i - n\lambda_1}{\Delta + 2n\lambda_2}$. We can see $\widehat{\beta}_i$ and $\widetilde{\beta}_i$ have the same sign.

Case 2: If $\frac{\Delta\widetilde{\beta}_i - n\lambda_1}{\Delta + 2n\lambda_2} \leq 0$, the sign of $\frac{\Delta\widetilde{\beta}_i + n\lambda_1}{\Delta + 2n\lambda_2}$ can't be determined, we neen to consider two cases $\frac{\Delta\widetilde{\beta}_i + n\lambda_1}{\Delta + 2n\lambda_2} < 0$ and $\frac{\Delta\widetilde{\beta}_i + n\lambda_1}{\Delta + 2n\lambda_2} \geq 0$.

First, when $\beta_i \geq 0$, the minimum point of $f(\beta_i)$ is 0, the objective function value is $\frac{\Delta}{2}\widetilde{\beta}_i^2$.

Second, when $\beta_i < 0$, if we assume $\frac{\Delta\widetilde{\beta}_i + n\lambda_1}{\Delta + 2n\lambda_2} < 0$ (which implies $\widetilde{\beta}_i \leq 0$), the minimum point of $f(\beta_i)$ is $\frac{\Delta\widetilde{\beta}_i + n\lambda_1}{\Delta + 2n\lambda_2}$, the objective function value is $\frac{\Delta}{2}\widetilde{\beta}_i^2 - \left(\frac{\Delta}{2} + n\lambda_2\right)\left(\frac{\Delta\widetilde{\beta}_i + n\lambda_1}{\Delta + 2n\lambda_2}\right)^2$. If we assume $\frac{\Delta\widetilde{\beta}_i + n\lambda_1}{\Delta + 2n\lambda_2} \geq 0$, the minimum point of (4.4) is $\widehat{\beta}_i = 0$. Therefore the global optimal solution of (4.4) is $\widehat{\beta}_i = \frac{\Delta\widetilde{\beta}_i + n\lambda_1}{\Delta + 2n\lambda_2} \leq 0$, which has the same sign as $\widetilde{\beta}_i$.

Combining the above discussion, we can obtain the optimal solution of (4.4)

$$\begin{aligned}\widehat{\beta}_i &= \operatorname{sign}(\widetilde{\beta}_i) \cdot \max\left\{\frac{\Delta|\widetilde{\beta}_i| - n\lambda_1}{\Delta + 2n\lambda_2}, 0\right\} \\ &= \frac{L_\delta + \varepsilon}{L_\delta + \xi + 2n\lambda_2}\operatorname{sign}(\widetilde{\beta}_i) \cdot \max\left\{|\widetilde{\beta}_i| - \frac{n\lambda_1}{L_\delta + \xi}, 0\right\}.\end{aligned}$$

Thus,

$$\widehat{\boldsymbol{\beta}} = \frac{L_\delta + \xi}{L_\delta + \xi + 2n\lambda_2}\operatorname{sign}(\widetilde{\boldsymbol{\beta}}) \circ \max\left\{|\widetilde{\boldsymbol{\beta}}| - \frac{n\lambda_1}{L_\delta + \xi}, 0\right\}.$$

□

We have calculated the form of $\widehat{\boldsymbol{\beta}}$, so we can update $\boldsymbol{\beta}^{k+1}$ as

$$\boldsymbol{\beta}^{k+1} = \frac{L_\delta + \xi}{L_\delta + \xi + 2n\lambda_2}\operatorname{sign}(\widetilde{\boldsymbol{\beta}}^k) \circ \max\left\{|\widetilde{\boldsymbol{\beta}}^k| - \frac{n\lambda_1}{L_\delta + \xi}, 0\right\}, \tag{4.6}$$

where $\widetilde{\boldsymbol{\beta}}^k = \boldsymbol{\beta}^k - \nabla S_\tau(\boldsymbol{\beta}^k, \delta)/(L_\delta + \xi)$.

In order to solve model (2.3), we are going to present the iterative scheme called SQEN-MM in the following Table 1.

Based on Theorem 4.1, the following theorem establishes the relationship of optimal solutions between (2.3) and the problem

$$\underset{\boldsymbol{\beta} \in \mathbb{R}^p}{\operatorname{argmin}}\left\{M_{\tau,\xi}(\boldsymbol{\beta}, \widehat{\boldsymbol{\beta}}, \delta) + n\lambda_1\|\boldsymbol{\beta}\|_1 + n\lambda_2\|\boldsymbol{\beta}\|_2^2\right\}. \tag{4.7}$$

Table 1: Iterative scheme of SQEN-MM for SQEN.

SQEN-MM Algorithm
Initialize: $X, y, \text{tol}, \lambda_1 > 0, \lambda_2 > 0, \delta > 0, \xi > 0, \text{maxiter}, L_\delta = \frac{1}{\delta}\lambda_{\max}(X^{\mathrm{T}}X)$ and $\boldsymbol{\beta}^0$
for k=0:maxiter
Compute $\widetilde{\boldsymbol{\beta}}^k = \boldsymbol{\beta}^k - \nabla S_\tau(\boldsymbol{\beta}^k, \delta)/(L_\delta + \xi)$;
Compute $\boldsymbol{\beta}^{k+1} = \frac{L_\delta+\xi}{L_\delta+\xi+2n\lambda_2}\text{sign}(\widetilde{\boldsymbol{\beta}}^k) \circ \max\left\{|\widetilde{\boldsymbol{\beta}}^k| - \frac{n\lambda_1}{L_\delta+\xi}, 0\right\}$
if $\|\boldsymbol{\beta}^{k+1} - \boldsymbol{\beta}^k\|_2 < \text{tol}$
break
end
End
Output: $\boldsymbol{\beta}^{k+1}$

Theorem 4.2. *Let $\tau \in (0,1), \delta, \xi, \lambda_1, \lambda_2 > 0$. If $\widehat{\boldsymbol{\beta}}$ is a global minimizer of (2.3), then $\widehat{\boldsymbol{\beta}}$ is also the global minimizer of (4.7).*

Proof. If $\widehat{\boldsymbol{\beta}}$ is a global minimizer of (2.3), since $M_{\tau,\xi}(\boldsymbol{\beta}, \widehat{\boldsymbol{\beta}}, \delta)$ is the majorized function of $S_\tau(\boldsymbol{\beta}, \delta)$, so we will have

$$\begin{aligned} M_{\tau,\xi}(\boldsymbol{\beta}, \widehat{\boldsymbol{\beta}}, \delta) + n\lambda_1\|\boldsymbol{\beta}\|_1 + n\lambda_2\|\boldsymbol{\beta}\|_2^2 &\geq S_\tau(\boldsymbol{\beta}, \delta) + n\lambda_1\|\boldsymbol{\beta}\|_1 + n\lambda_2\|\boldsymbol{\beta}\|_2^2 \\ &\geq S_\tau(\widehat{\boldsymbol{\beta}}, \delta) + n\lambda_1\|\widehat{\boldsymbol{\beta}}\|_1 + n\lambda_2\|\widehat{\boldsymbol{\beta}}\|_2^2 \\ &= M_{\tau,\xi}(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\beta}}, \delta) + n\lambda_1\|\widehat{\boldsymbol{\beta}}\|_1 + n\lambda_2\|\widehat{\boldsymbol{\beta}}\|_2^2. \end{aligned}$$

□

Then we can build the necessary and sufficient conditions for the solution of fixed point equation to be the optimal solution of (2.3).

Theorem 4.3. *Let $\tau \in (0,1), \delta, \xi, \lambda_1, \lambda_2 > 0, L_\delta = \frac{1}{\delta}\lambda_{max}(X^{\tau}X)$. Then $\widehat{\boldsymbol{\beta}}$ is a global minimizer of (2.3) if and only if $\widehat{\boldsymbol{\beta}}$ satisfies the following fixed point equation*

$$\widehat{\boldsymbol{\beta}} = \frac{L_\delta + \xi}{L_\delta + \xi + 2n\lambda_2}\text{sign}(\widetilde{\boldsymbol{\beta}}) \circ \max\left\{|\widetilde{\boldsymbol{\beta}}| - \frac{n\lambda_1}{L_\delta + \xi}, 0\right\}, \tag{4.8}$$

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

Proof. **Necessity**: Since $\widehat{\boldsymbol{\beta}}$ is the global minimizer of (2.3), it is also the global minimizer of $M_{\tau,\xi}(\boldsymbol{\beta}, \widehat{\boldsymbol{\beta}}, \delta) + \lambda_1\|\boldsymbol{\beta}\|_1 + \lambda_2\|\boldsymbol{\beta}\|_2^2$ from Theorem 4.2. Then by the Theorem 4.1 (with $\boldsymbol{\alpha} = \widehat{\boldsymbol{\beta}}$), $\widehat{\boldsymbol{\beta}}$ satisfies the fixed point equation.

Sufficiency: Due to the convexity of the objective function in (2.3), $\boldsymbol{\beta}^*$ is the global minimizer of model (2.3) if and only if $0 \in \partial(\Phi_\tau(\boldsymbol{\beta}^*))$. That is to say

$$\begin{aligned} 0 &\in \partial(S_\tau(\boldsymbol{\beta}, \delta) + n\lambda_1\|\boldsymbol{\beta}\|_1 + n\lambda_2\|\boldsymbol{\beta}\|_2^2)\,|_{\boldsymbol{\beta}=\boldsymbol{\beta}^*} \\ &= \nabla S_\tau(\boldsymbol{\beta}^*, \delta) + 2n\lambda_2\boldsymbol{\beta}^* + n\lambda_1\partial(\|\boldsymbol{\beta}^*\|_1). \end{aligned}$$

From the fixed point equation, we have

$$\widehat{\boldsymbol{\beta}} = \frac{L_\delta + \xi}{L_\delta + \xi + 2n\lambda_2}\text{sign}(\widetilde{\boldsymbol{\beta}}) \circ \max\left\{|\widetilde{\boldsymbol{\beta}}| - \frac{n\lambda_1}{L_\delta + \xi}, 0\right\}$$

$$= \underset{\boldsymbol{\beta}\in\mathbb{R}^p}{\operatorname{argmin}} \left\{ M_{\tau,\xi}(\boldsymbol{\beta}, \widehat{\boldsymbol{\beta}}, \delta) + n\lambda_1\|\boldsymbol{\beta}\|_1 + n\lambda_2\|\boldsymbol{\beta}\|_2^2 \right\}$$
$$= \underset{\boldsymbol{\beta}\in\mathbb{R}^p}{\operatorname{argmin}} \left\{ \frac{L_\delta + \xi}{2}\|\boldsymbol{\beta} - \widetilde{\boldsymbol{\beta}}\|_2^2 + n\lambda_1\|\boldsymbol{\beta}\|_1 + n\lambda_2\|\boldsymbol{\beta}\|_2^2 \right\}. \tag{4.9}$$

Considering the convexity of the objective function in (4.9), we have

$$\begin{aligned} 0 &\in \partial\left(\frac{L_\delta + \xi}{2}\|\boldsymbol{\beta} - \widetilde{\boldsymbol{\beta}}\|_2^2 + n\lambda_1\|\boldsymbol{\beta}\|_1 + n\lambda_2\|\boldsymbol{\beta}\|_2^2 \right)\Big|_{\boldsymbol{\beta}=\widehat{\boldsymbol{\beta}}} \\ &= (L_\delta + \xi)(\widehat{\boldsymbol{\beta}} - \widetilde{\boldsymbol{\beta}}) + 2n\lambda_2\widehat{\boldsymbol{\beta}} + n\lambda_1\partial(\|\widehat{\boldsymbol{\beta}}\|_1) \\ &= \nabla S_\tau(\widehat{\boldsymbol{\beta}}, \delta) + 2n\lambda_2\widehat{\boldsymbol{\beta}} + n\lambda_1\partial(\|\widehat{\boldsymbol{\beta}}\|_1), \end{aligned}$$

where the last equality derives from $\widetilde{\boldsymbol{\beta}} = \widehat{\boldsymbol{\beta}} - \nabla S_\tau(\widehat{\boldsymbol{\beta}}, \delta)/(L_\delta + \varepsilon)$. Hence, $\widehat{\boldsymbol{\beta}}$ is the global minimizer of model (2.3). Thus theorem 4.3 holds. □

By means of the above results, we are able to establish the convergence theorem of the proposed algorithm SQEN-MM.

Theorem 4.4. *For a given $\tau \in (0,1), \delta, \xi, \lambda_1, \lambda_2 > 0$, let $\{\boldsymbol{\beta}^k\}$ be the sequence generated by (4.6) and $\Phi_\tau(\boldsymbol{\beta}^k, \delta) = S_\tau(\boldsymbol{\beta}^k, \delta) + n\lambda_1\|\boldsymbol{\beta}^k\|_1 + n\lambda_2\|\boldsymbol{\beta}^k\|_2^2$, then*

(1) *$\Phi_\tau(\boldsymbol{\beta}^k)$ is monotonically non-increasing and converges to $\Phi_\tau(\widehat{\boldsymbol{\beta}})$, where $\widehat{\boldsymbol{\beta}}$ is any accumulation point of $\{\boldsymbol{\beta}^k\}$;*

(2) *$\{\boldsymbol{\beta}^k\}$ is asymptotically regular, namely, $\lim_{k\to\infty} \|\boldsymbol{\beta}^{k+1} - \boldsymbol{\beta}^k\|_2 = 0$;*

(3) *$\{\boldsymbol{\beta}^k\}$ converges to the unique global minimizer of problem (2.3)*

Proof. (1) From the expression of $\Phi_\tau(\boldsymbol{\beta}^{k+1}, \delta)$, we have

$$\begin{aligned} \Phi_\tau(\boldsymbol{\beta}^{k+1}, \delta) &= S_\tau(\boldsymbol{\beta}^{k+1}, \delta) + n\lambda_1\|\boldsymbol{\beta}^{k+1}\|_1 + n\lambda_2\|\boldsymbol{\beta}^{k+1}\|_2^2 \\ &\overset{(i)}{\leq} M_{\tau,\xi}(\boldsymbol{\beta}^{k+1}, \boldsymbol{\beta}^k, \delta) + n\lambda_1\|\boldsymbol{\beta}^{k+1}\|_1 + n\lambda_2\|\boldsymbol{\beta}^{k+1}\|_2^2 \\ &\overset{(ii)}{\leq} M_{\tau,\xi}(\boldsymbol{\beta}^k, \boldsymbol{\beta}^k, \delta) + n\lambda_1\|\boldsymbol{\beta}^k\|_1 + n\lambda_2\|\boldsymbol{\beta}^k\|_2^2 \\ &= S_\tau(\boldsymbol{\beta}^k, \delta) + n\lambda_1\|\boldsymbol{\beta}^k\|_1 + n\lambda_2\|\boldsymbol{\beta}^k\|_2^2 \\ &= \Phi_\tau(\boldsymbol{\beta}^k, \delta), \end{aligned}$$

where (i) follows from $M_{\tau,\xi}(\boldsymbol{\beta}^{k+1}, \boldsymbol{\alpha}, \delta)$ is the majorization of $S_\tau(\boldsymbol{\beta}^{k+1}, \delta)$, and (ii) derives from (4.2). This indicates that $\{\Phi_\tau(\boldsymbol{\beta}^k, \delta)\}$ is monotonically non-increasing. As $\{\Phi_\tau(\boldsymbol{\beta}^k, \delta)\}$ is bounded from below, $\{\Phi_\tau(\boldsymbol{\beta}^k, \delta)\}$ converges to a constant $\widehat{\Phi}$. For any $\boldsymbol{\beta} \in \{\boldsymbol{\beta} | \Phi_\tau(\boldsymbol{\beta}, \delta) \leq \Phi_\tau(\boldsymbol{\beta}^0, \delta)\}$, we have $n\lambda_2\|\boldsymbol{\beta}\|_2^2 \leq \Phi_\tau(\boldsymbol{\beta}, \delta) \leq \Phi_\tau(\boldsymbol{\beta}^0, \delta)$, hence $\|\boldsymbol{\beta}\|_2^2 \leq \frac{1}{n\lambda_2}\Phi_\tau(\boldsymbol{\beta}^0, \delta)$, so the set $\{\boldsymbol{\beta} | \Phi_\tau(\boldsymbol{\beta}, \delta) \leq \Phi_\tau(\boldsymbol{\beta}^0, \delta)\}$ is bounded. From $\{\boldsymbol{\beta}^k\} \subset \{\boldsymbol{\beta} | \Phi_\tau(\boldsymbol{\beta}, \delta) \leq \Phi_\tau(\boldsymbol{\beta}^0, \delta)\}$ which is bounded, it yields that $\{\boldsymbol{\beta}^k\}$ is also bounded, and thus $\{\boldsymbol{\beta}^k\}$ has at least one accumulation point. Let $\widehat{\boldsymbol{\beta}}$ be an accumulation point of $\{\boldsymbol{\beta}^k\}$. By the continuity of $\Phi_\tau(\boldsymbol{\beta})$ and the convergence of $\{\Phi_\tau(\boldsymbol{\beta}^k, \delta)\}$, we get $\{\Phi_\tau(\boldsymbol{\beta}^k, \delta)\} \to \widehat{\Phi} = \Phi_\tau(\widehat{\boldsymbol{\beta}}, \delta)$ as $k \to \infty$.

(2) By simple calculation, we have

$$\Phi_\tau(\boldsymbol{\beta}^k, \delta) - \Phi_\tau(\boldsymbol{\beta}^{k+1}, \delta) = S_\tau(\boldsymbol{\beta}^k, \delta) + n\lambda_1\|\boldsymbol{\beta}^k\|_1 + n\lambda_2\|\boldsymbol{\beta}^k\|_2^2$$

$$
\begin{aligned}
&\quad - \Big(S_\tau(\boldsymbol{\beta}^{k+1},\delta) + n\lambda_1\|\boldsymbol{\beta}^{k+1}\|_1 + n\lambda_2\|\boldsymbol{\beta}^{k+1}\|_2^2\Big) \\
&= M_{\tau,\xi}(\boldsymbol{\beta}^k,\boldsymbol{\beta}^k,\delta) + n\lambda_1\|\boldsymbol{\beta}^k\|_1 + n\lambda_2\|\boldsymbol{\beta}^k\|_2^2 \\
&\quad - \Big(S_\tau(\boldsymbol{\beta}^{k+1},\delta) + n\lambda_1\|\boldsymbol{\beta}^{k+1}\|_1 + n\lambda_2\|\boldsymbol{\beta}^{k+1}\|_2^2\Big) \\
&\overset{(i)}{\geq} M_{\tau,\xi}(\boldsymbol{\beta}^{k+1},\boldsymbol{\beta}^k,\delta) + n\lambda_1\|\boldsymbol{\beta}^{k+1}\|_1 + n\lambda_2\|\boldsymbol{\beta}^{k+1}\|_2^2 \\
&\quad - \Big(S_\tau(\boldsymbol{\beta}^{k+1},\delta) + n\lambda_1\|\boldsymbol{\beta}^{k+1}\|_1 + n\lambda_2\|\boldsymbol{\beta}^{k+1}\|_2^2\Big) \\
&= M_{\tau,\xi}(\boldsymbol{\beta}^{k+1},\boldsymbol{\beta}^k,\delta) - S_\tau(\boldsymbol{\beta}^{k+1},\delta) \\
&\overset{(ii)}{\leq} S_\tau(\boldsymbol{\beta}^k,\delta) + \langle\nabla S_\tau(\boldsymbol{\beta}^k,\delta),\boldsymbol{\beta}^{k+1}-\boldsymbol{\beta}^k\rangle \\
&\quad + \frac{L_\delta+\xi}{2}\|\boldsymbol{\beta}^{k+1}-\boldsymbol{\beta}^k\|_2^2 - S_\tau(\boldsymbol{\beta}^{k+1},\delta) \\
&\overset{(iii)}{\leq} S_\tau(\boldsymbol{\beta}^k,\delta) + \langle\nabla S_\tau(\boldsymbol{\beta}^k,\delta),\boldsymbol{\beta}^{k+1}-\boldsymbol{\beta}^k\rangle + \frac{L_\delta+\xi}{2}\|\boldsymbol{\beta}^{k+1}-\boldsymbol{\beta}^k\|_2^2 \\
&\quad - \Big(S_\tau(\boldsymbol{\beta}^k,\delta) + \langle\nabla S_\tau(\boldsymbol{\beta}^k,\delta),\boldsymbol{\beta}^{k+1}-\boldsymbol{\beta}^k\rangle + \frac{L_\delta}{2}\|\boldsymbol{\beta}^{k+1}-\boldsymbol{\beta}^k\|_2^2\Big) \\
&= \frac{\xi}{2}\|\boldsymbol{\beta}^{k+1}-\boldsymbol{\beta}^k\|_2^2,
\end{aligned}
$$

where (i) derives from $\boldsymbol{\beta}^{k+1} = \underset{\boldsymbol{\beta}\in\mathbb{R}^p}{\operatorname{argmin}}\Big\{M_{\tau,\xi}(\boldsymbol{\beta},\boldsymbol{\beta}^k,\delta) + n\lambda_1\|\boldsymbol{\beta}\|_1 + n\lambda_2\|\boldsymbol{\beta}\|_2^2\Big\}$, (ii) holds due to $S_\tau(\boldsymbol{\beta}^k,\delta) + \langle\nabla S_\tau(\boldsymbol{\beta}^k,\delta),\boldsymbol{\beta}^{k+1}-\boldsymbol{\beta}^k\rangle + \frac{L_\delta+\xi}{2}\|\boldsymbol{\beta}^{k+1}-\boldsymbol{\beta}^k\|_2^2$ being the majorization of $S_\tau(\boldsymbol{\beta},\delta)$, and (iii) follows from the convexity of $S_\tau(\boldsymbol{\beta},\delta)$. Hence, for a given $\xi>0$ and any $N>0$, considering $\Phi_\tau(\boldsymbol{\beta},\delta)\geq 0$, we have

$$
\begin{aligned}
\sum_{k=0}^{N}\|\boldsymbol{\beta}^{k+1}-\boldsymbol{\beta}^k\|_2^2 &\leq \frac{2}{\xi}\sum_{k=0}^{N}\Big(\Phi_\tau(\boldsymbol{\beta}^k) - \Phi_\tau(\boldsymbol{\beta}^{k+1})\Big) \\
&\leq \frac{2}{\xi}\Big(\Phi_\tau(\boldsymbol{\beta}^0,\delta) - \Phi_\tau(\boldsymbol{\beta}^N,\delta)\Big) \\
&\leq \frac{2}{\xi}\Phi_\tau(\boldsymbol{\beta}^0,\delta), \qquad (4.10)
\end{aligned}
$$

which implies that $\sum_{k=0}^{\infty}\|\boldsymbol{\beta}^{k+1}-\boldsymbol{\beta}^k\|_2^2<\infty$. Thus $\lim_{k\to\infty}\|\boldsymbol{\beta}^{k+1}-\boldsymbol{\beta}^k\|_2=0$.

(3) Let $\{\boldsymbol{\beta}^{k_j}\}$ be a convergent subsequence of $\{\boldsymbol{\beta}^k\}$ and $\widehat{\boldsymbol{\beta}}$ be its limit point, i.e., $\lim_{k_j\to\infty}\boldsymbol{\beta}^{k_j}=\widehat{\boldsymbol{\beta}}$. Since $S_\tau(\boldsymbol{\beta},\delta)$ is continuously differentiable, we can derive

$$
\lim_{k_j\to\infty}\widetilde{\boldsymbol{\beta}}^{k_j} = \lim_{k_j\to\infty}\boldsymbol{\beta}^{k_j} - \frac{S_\tau(\boldsymbol{\beta}^{k_j},\delta)}{L_\delta+\xi} = \widehat{\boldsymbol{\beta}} - \frac{S_\tau(\widehat{\boldsymbol{\beta}},\delta)}{L_\delta+\xi} = \widetilde{\boldsymbol{\beta}}.
$$

Then the convergency of $\{\boldsymbol{\beta}^{k_j}\}$ and the asymptoticy of $\{\boldsymbol{\beta}^k\}$ imply

$$
\|\boldsymbol{\beta}^{k_j+1}-\widehat{\boldsymbol{\beta}}\|_2 \leq \|\boldsymbol{\beta}^{k_j+1}-\boldsymbol{\beta}^{k_j}\|_2 + \|\boldsymbol{\beta}^{k_j}-\widehat{\boldsymbol{\beta}}\|_2 \to 0, \text{ as } k_j\to 0, \qquad (4.11)
$$

which guarantees that $\{\boldsymbol{\beta}^{k_j+1}\}$ also converges to $\widehat{\boldsymbol{\beta}}$. So we can obtain that

$$
\lim_{k_j\to\infty}\boldsymbol{\beta}^{k_j+1} = \lim_{k_j\to\infty}\frac{L_\delta+\xi}{L_\delta+\xi+2n\lambda_2}\operatorname{sign}(\widetilde{\boldsymbol{\beta}}^{k_j})\circ\max\left\{|\widetilde{\boldsymbol{\beta}}^{k_j}| - \frac{n\lambda_1}{L_\delta+\xi}, 0\right\}
$$

$$= \frac{L_\delta + \xi}{L_\delta + \xi + 2n\lambda_2} \text{sign}(\widetilde{\boldsymbol{\beta}}) \circ \max\left\{ |\widetilde{\boldsymbol{\beta}}| - \frac{n\lambda_1}{L_\delta + \xi}, 0 \right\},$$

which means $\widehat{\boldsymbol{\beta}} = \frac{L_\delta+\xi}{L_\delta+\xi+2n\lambda_2} \text{sign}(\widetilde{\boldsymbol{\beta}}) \circ \max\left\{ |\widetilde{\boldsymbol{\beta}}| - \frac{n\lambda_1}{L_\delta+\xi}, 0 \right\}$. Finally, $\widehat{\boldsymbol{\beta}}$ is the global minimizer of (2.3) from Theorem 4.3. Since (2.3) is strongly convex, $\widehat{\boldsymbol{\beta}}$ is unique. Hence, $\{\boldsymbol{\beta}^k\}$ converges to the global minimizer (2.3).

5 Numerical Study

In this section we focus on the numerical studies of our algorithm to show it's efficiency. We apply our method to simulation data with six different errors. And we also compare our method with Lasso and Elastic net.

To assess the performance of the proposed method, we simulate data from the high-dimensional linear regression model

$$\mathbf{y} = X\boldsymbol{\beta}^* + \boldsymbol{\varepsilon}, \tag{5.1}$$

where the data set has $n = 100$ observations and $p = 400$ parameters. In our experiments we chose the first 15 elements nonzero, i.e.,

$$\boldsymbol{\beta}^* = (\underbrace{3, \ldots, 3}_{15}, \underbrace{0, \ldots, 0}_{385}), \tag{5.2}$$

and the predictors X are generated as follows:

$$\begin{aligned}
&x_i = Z_1 + \varepsilon_i^x,\ Z_1 \sim N(0,1),\ i = 1, \ldots, 5,\\
&x_i = Z_2 + \varepsilon_i^x,\ Z_2 \sim N(0,1),\ i = 6, \ldots, 10,\\
&x_i = Z_3 + \varepsilon_i^x,\ Z_3 \sim N(0,1),\ i = 11, \ldots, 15,\\
&x_i \sim N(0,1),\ x_i \text{independent identically distributed}, i = 16, \ldots, 400,
\end{aligned}$$

where ε_i^x are independent identically distributed. For the distribution of the noise ε, we considered the following six scenarios:

(a) normal errors with mean 0 and variance 4 ($N(0, 2^2)$);

(b) two times the t-distribution with degrees of freedom 3($2t(3)$);

(c) mixture of normal distributions, MixN, $0.5N(-1, 2^2) + 0.5N(8, 1)$;

(d) log-normal distribution, LogNormal, $\varepsilon = \exp^{(1+1.2Z)}$, where Z is the standard normal distribution;

(e) the Cauchy distribution with density location parameter 0, scale parameter 2;

(f) the Laplace distribution with location parameter 0, scale parameter 2.

There are three equally important groups, and within each group there are five members. There are another 385 pure noise features. An ideal method would select only 15 true features and set the coefficients of the 385 noise features to 0.

We compared the SQEN method with two other methods in a high dimensional setting:

(a) LASSO, the penalized least squares estimator with l_1-penalty as in Tibshirani [17];

(b) Elastic net, the penalized least squares estimator with l_1-penalty and l_2-penalty proposed by Zou and Hastie in [19].

Note that, we use the code of these two methods provided by Matlab. And the tuning parameter grid is used as in our method. All runs are performed on an laptop with Intel Core(TM)i7-2640M CPU (2.80 GHz) and 8 GB RAM.

The following five performance measures were calculated:

(i) The mean-squared-errors (**MSE**), which is $\mathbf{MSE} := \|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2^2/400$;

(ii) The mean-squared-errors on the test data set (**Test-MSE**), which is $\mathbf{MSE} := \|\mathbf{y}_{\text{test}} - X_{\text{test}}\widehat{\boldsymbol{\beta}}\|_2^2/n_{\text{test}}$, where $(\mathbf{y}_{\text{test}}, X_{\text{test}})$ are the data from test data set, n_{test} is the number of observations in test data set;

(iii) the number of false positive results, **FP**, which is the number of noise covariates that are selected;

(iv) the number of false negative results, **FN**, which is the number of signal covariates that are not selected;

(v) the elapsed time per run, **CPU**.

For the SQEN, the tuning parameters λ_1, λ_2 were chosen optimally on the basis of an independent validation data set which has one thousand observations. We ran a two-dimensional grid search to find the best (λ_1, λ_2) pair that minimizes mean **Test-MSE**. Such an optimal pair was then used in the simulations. A similar method was applied in choosing the tuning parameters in the LASSO and the Elastic net. Throughout the experiments, we choose the initial iterate to be the initial point for this method is set to be $\beta^0 = O_p$, tolerance error is $\text{tol} = 10^{-4}$, the smoothing parameter $\delta = 2$ and the parameter $\xi = 0.1$. For the quantile parameter τ, we choose three values $\{0.25, 0.5, 0.75\}$ to demonstrate the performance of our method. The numerical results are displayed in Tables 2-4.

Table 2: Numerical results for $\tau = 0.25$ in terms of **MSE**, **Test-MSE**, **FP**, **FN** and **CPU**. All the measurements are the mean of the results after repeated 100 times. The numbers in parentheses are the corresponding standard errors.

MixN	LASSO	0.5830(7.91e-3)	0.1716(7.79e-2)	18(2.41)	0(0)	1.21
	Elastic net	0.5832(9.65e-3)	0.2272(9.79e-2)	43.5(6.36)	0(0)	1.06
	SQEN	0.3010(3.15e-3)	0.1753(5.73e-2)	17.5(2.12)	0(0)	9.70e-2
LogNormal	LASSO	0.5968(1.31e-3)	0.3525(4.02e-2)	10.5(2.78)	1(0)	2.19
	Elastic net	0.5954(7.30e-3)	0.4417(1.45e-1)	51.5(5.71)	0.5(0.71)	1.25
	SQEN	0.3052(7.62e-3)	0.3238(8.81e-2)	7.20(1.21)	0(0)	9.37e-2
Cauchy	LASSO	0.6104(3.03e-2)	1.0323(3.33e-1)	18.5(2.36)	3(4.24)	3.38
	Elastic net	0.6021(2.29e-2)	1.1240(3.32e-1)	46(4.24)	1(1.41)	1.93
	SQEN	0.6255(6.78e-4)	1.0751(4.78e-2)	20(4.38)	3.5(0.71)	1.27e-1
Laplace	LASSO	0.5973(8.75e-3)	0.3221(1.37e-1)	18.5(7.58)	0(0)	1.28
	Elastic net	0.5958(5.82e-3)	0.4079(2.65e-2)	48.5(10.02)	0(0)	1.46
	SQEN	0.3061(1.79e-3)	0.3499(3.30e-2)	13(2.41)	0(0)	8.81e-2

Tables 2-4 show the simulation results of **MSE**, **Test-MSE**, **FP** and **FN** for $\tau = 0.25, 0.5$ and 0.75. To be fair, we simulate 100 times for each model. Then we report the mean value and standard deviation (in parentheses) for all measurements. From these tables, we can see that the values of **FN** are all zeros except when the errors are from Cauchy distribution. This indicates that all significant variables are selected. In fact, when the data has Cauchy error, not only SQEN doesn't perform well, but also Lasso and Elastic net. For **MSE** and **CPU**, Lasso and Elastic net have almost the same performance, however, SQEN is much better than them.

In our numerical experiments, the first 15 predictors are collinear. We wish they are all in the model together. In fact, the simulation results show that these 15 predictors are included in the model together except for Cauchy distribution error. Hence, the SQEN has the grouping selection ability, and it also can produce a sparse estimator. In addition, our method works well not only for normal distribution, but also for heavy-tailed distribution.

Table 3: Numerical results for $\tau = 0.5$ in terms of **MSE**, **Test-MSE**, **FP**, **FN** and **CPU**. All the measurements are the mean of the results after repeated 100 times. The numbers in parentheses are the corresponding standard errors.

MixN	LASSO	0.5830(7.91e-3)	0.1716(7.79e-2)	18(2.41)	0(0)	1.21
	Elastic net	0.5832(9.65e-3)	0.2272(9.79e-2)	43.5(6.36)	0(0)	1.06
	SQEN	0.3010(3.15e-3)	0.1753(5.73e-2)	17.5(2.12)	0(0)	9.70e-2
LogNormal	LASSO	0.5968(1.31e-3)	0.3525(4.02e-2)	10.5(2.78)	1(0)	2.19
	Elastic net	0.5954(7.30e-3)	0.4417(1.45e-1)	51.5(5.71)	0.5(0.71)	1.25
	SQEN	0.3052(7.62e-3)	0.3238(8.81e-2)	7.20(1.21)	0(0)	9.37e-2
Cauchy	LASSO	0.6104(3.03e-2)	1.0323(3.33e-1)	18.5(2.36)	3(4.24)	3.38
	Elastic net	0.6021(2.29e-2)	1.1240(3.32e-1)	46(4.24)	1(1.41)	1.93
	SQEN	0.6255(6.78e-4)	1.0751(4.78e-2)	20(4.38)	3.5(0.71)	1.27e-1
Laplace	LASSO	0.5973(8.75e-3)	0.3221(1.37e-1)	18.5(7.58)	0(0)	1.28
	Elastic net	0.5958(5.82e-3)	0.4079(2.65e-2)	48.5(10.02)	0(0)	1.46
	SQEN	0.3061(1.79e-3)	0.3499(3.30e-2)	13(2.41)	0(0)	8.81e-2

Table 4: Numerical results for $\tau = 0.75$ in terms of **MSE**, **Test-MSE**, **FP**, **FN** and **CPU**. All the measurements are the mean of the results after repeated 100 times. The numbers in parentheses are the corresponding standard errors.

Error	Methods	**MSE**	**Test-MSE**	**FP**	**FN**	**CPU**
$N(0,4)$	LASSO	0.5931(1.19e-2)	0.2573(1.45e-1)	16.5(3.07)	0(0)	1.78
	Elastic net	0.5879(8.06e-3)	0.2701(1.55e-1)	20.5(5.07)	0(0)	1.33
	SQEN	0.3037(3.05e-3)	0.2561(7.46e-2)	19(4.89)	0(0)	9.68e-2
$2t(3)$	LASSO	0.5892(6.59e-3)	0.2076(8.61e-3)	10(1.66)	0(0)	1.66
	Elastic net	0.5835(2.09e-3)	0.2169(1.09e-2)	39.5(9.19)	0(0)	1.18
	SQEN	0.3068(7.81e-3)	0.3330(2.01e-1)	9.5(1.48)	0(0)	7.23e-2
MixN	LASSO	0.5811(5.18e-3)	0.1748(1.52e-2)	21(2.89)	0(0)	1.16
	Elastic net	0.5801(4.63e-3)	0.2021(2.24e-2)	49.5(3.54)	0(0)	0.64
	SQEN	0.2995(3.79E-3)	0.1458(7.99e-2)	23.5(1.62)	0(0)	9.23e-2
LogNormal	LASSO	0.5955(1.92e-2)	0.3081(2.55e-1)	13.5(2.85)	0(0)	1.53
	Elastic net	0.5873(1.44e-2)	0.2814(2.12e-1)	34.5(4.02)	0(0)	1.26
	SQEN	0.3012(1.42e-3)	0.2232(9.24e-3)	10.5(2.25)	0(0)	7.57e-2
Cauchy	LASSO	0.6024(1.11e-2)	0.9614(6.78e-1)	20.5(3.54)	0.5(5.07e-1)	1.64
	Elastic net	0.5974(1.05e-2)	1.0259(6.45e-1)	51.5(7.78)	0.5(7.07e-1)	1.53
	SQEN	0.3217(1.86e-2)	1.1587(2.05e-1)	15.85(1.68)	2.5(0.55)	6.08e-2

As we can see, all the methods seem to choose a bunch of variables. In fact, all the three methods can select few variables as you want. What we need to do is only increasing the value of λ_1. However, in our method, the criterion we used to select tuning parameters is the **Test-MSE**, so in the final model, there are more variables.

In our simulation studies, we choose fixed values for δ. When it comes to real data, we need to choose it from some more candidates. From the definition of smoothing quantile function, as $\delta \to \infty$, it tends to least squares, as $\delta \to 0$, it approximates to quantile. So for a proper choice, our smoothing quantile regression can show a good behavior. The parameter ξ in majorized function $M_{\tau,\xi}(\boldsymbol{\beta}, \boldsymbol{\alpha}, \delta)$ who must make $M_{\tau,\xi}(\boldsymbol{\beta}, \boldsymbol{\alpha}, \delta)$ greater than $S_\tau(\boldsymbol{\beta}, \delta)$. From the convexity of $S_\tau(\boldsymbol{\beta}, \delta)$, it seems that if ξ is large enough, $M_{\tau,\xi}(\boldsymbol{\beta}, \boldsymbol{\alpha}, \delta) \geq S_\tau(\boldsymbol{\beta}, \delta)$ holds definitely. However, in our algorithm, $1/(L_\delta + \xi)$ is the step size. If it is too small, the algorithm will consume more CPU. So, ξ must be not too large. Above all, this two

parameters must be chosen carefully.

6 Concluding Remarks

In this paper, we focus on how to analyse the high-dimensional data with the collinearity and heavy-tailed noise. In order to select significant variables, we introduced the penalized quantile regression with the elastic-net. In view of the quantile loss function is nonsmooth and nondifferential, we took advantage of its Huber smooth function. This developed the smoothing model SQEN. This model can work efficiently on the data with heavy-tailed noise for its robust loss function. Meanwhile, it can also deal with collinearity for the regularizer elastic-net term. For this new model, we obtained statistical consistent properties of its estimator. Further more, we proposed an optimization method using Majorize-Minimize (MM) technique to solve our SQEN model. Before applying it to solve problems, we prove the global convergence theoretically. Finally, we conduct some numerical experiments to show the efficiency of the proposed method. Our model is designed for single response, i.e., only one character is taken as dependent variable. However, as the development of the society, there are much more multi-responses data and matrix data. In the future, the extension of this model for these data sets is worth of study.

Acknowledgments

We would like to thank the anonymous referees and the associate editor for their insightful comments and suggestions, which improved the presentation of the paper.

References

[1] A. Aravkin, A. Kambadur, A. Lozano and R. Luss, Sparse Quantile Huber Regression for Efficient and Robust Estimation, Technical report, IBM T.J. Watson Research Center, http://arxiv.org/abs/1402.4624, 2014.

[2] A. Belloni and V. Chernozhukov, l_1 penalized quantile regression in high-dimensional sparse models, *Ann. Statist.* 39 (2011) 1012–1030.

[3] P. Bühlmann and D. Van, *Statistics for high-demensional data: Methods, Theorey and Applications*, Springer, Berlin Heidelberg, 2011.

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

[5] J. Fan and R. Li, Sure independence screening for ultrahigh dimensional feature space, *J. R. Stat. Soc. Ser. B Stat. Methodol.* 70 (2008) 849–911.

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

[7] A. Galvao and K. Kato, Smoothed quantile regression for panel data, *J. Econometrics* 193 (2016) 92–112.

[8] A. Giloni, J. Simonoff and B. Sengupta, Robust weighted LAD regression, *Comput. Statist. Data Anal.* 50 (2005) 3124–3140.

[9] R. Koenker, Quantile regression for longitudinal data, *J. Multivariate Anal.* 91 (2004) 74–89.

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

[11] R. Koenker and K. Hallock, Quantile regression, *J. Econ. Perspect.* 101 (2001) 143–156.

[12] R. Koenker and R. Geling, Reappraising medfly longevity: a quantile regression survival analysis, *J. Amer. Statist. Assoc.* 96 (2001) 458-468.

[13] R. Li and L. Peng, Quantile regression adjusting for dependent censoring from semi-competing risks, *J. R. Stat. Soc. Ser. B Stat. Methodol.* 77 (2015) 107–130.

[14] S. Magzamen, M. Amato, P. Imm, J. Havlena and M. Coons, Quantile regression in environmental health: Early life lead exposure and end-of-grade exams, *Environ. Res.* 137 (2015) 108–119.

[15] M. Meinshausen and P. Buhlmann, Stability selection, *J. R. Stat. Soc. Ser. B Stat. Methodol.* 72 (2010) 417–473.

[16] A. Mkhadri, M. Ouhourane and K. Oualkacha, A coordinate descent algorithm for computing penalized smooth quantile regression, *Stat. Comput.* 27 (2017) 1–19.

[17] R. Tibshirani, Regression shrinkage and selection via the Lasso, *J. R. Stat. Soc. Ser. B Stat. Methodol.* 58 (1996) 267–288.

[18] L. Wang, The l_1 penalized LAD estimator for high dimensional linear regression, *J. Multivariate Anal.* 120 (2013) 135–151.

[19] H. Zou, The adaptive lasso and its oracle properties, *J. Amer. Statist. Assoc.* 101 (2006) 1418–1429.

[20] H. Zou and T. Hastie, Regularization and variable selection via the elastic net, *J. R. Stat. Soc. Ser. B Stat. Methodol.* 67 (2005) 301–320.

[21] H. Zou and M. Yuan, Regularized simultaneous model selection in multiple quantiles regression, *Comput. Statist. Data Anal.* 52 (2008) 5296–5304.

Manuscript received 7 January 2018
revised 13 September 2018
accepted for publication 6 December 2018

Bingzhen Chen
Institute of Statistics and Big Data
Renmin University of China
Beijing 100872, China
E-mail address: chenbingzhen6026@163.com

Lingchen Kong
School of Science, Beijing Jiaotong University
Beijing 100044, P. R. China
E-mail address: lchkong@bjtu.edu.cn

Nana Xu
Zhidu Technology Ltd co., Beijing 100031, P. R. China
E-mail address: nanaxujiayou@163.com