Statistica Sinica **26** (2016), 69-95
doi:http://dx.doi.org/10.5705/ss.2014.049

REGULARIZED QUANTILE REGRESSION AND ROBUST FEATURE SCREENING FOR SINGLE INDEX MODELS

Wei Zhong, Liping Zhu, Runze Li and Hengjian Cui

*Xiamen University, Renmin University of China,
Pennsylvania State University and Capital Normal University*

Abstract: We propose a penalized quantile regression and an independence screening procedure to identify important covariates and to exclude unimportant ones for a general class of ultrahigh dimensional single index models, in which the conditional distribution of the response depends on the covariates via a single index structure. We observe that linear quantile regression yields a consistent estimator of the direction of the index parameter in the single index model. Such an observation dramatically reduces computational complexity in selecting important covariates in the single index model. We establish an oracle property for the penalized quantile regression estimator when the covariate dimension increases at an exponential rate of the sample size. From a practical perspective, however, when the covariate dimension is extremely large, the penalized quantile regression may suffer from at least two drawbacks: computational expediency and algorithmic stability. To address these issues, we propose an independence screening procedure which is robust to model misspecification, and has reliable performance when the distribution of the response variable is heavily tailed or response realizations contain extreme values. The new independence screening procedure offers a useful complement to penalized quantile regression since it helps to reduce the covariate dimension from ultrahigh dimensionality to a moderate scale. Based on the reduced model, penalized linear quantile regression further refines selection of important covariates at different quantile levels. We examine the finite sample performance of the newly proposed procedure by Monte Carlo simulations and demonstrate the proposed methodology by an empirical analysis of a data set.

Key words and phrases: Distance correlation, penalized quantile regression, single index models, sure screening property, ultrahigh dimensionality.

1. Introduction

Single index regression models are widely assumed to avoid the "curse of dimensionality". Let Y be a response variable and $\mathbf{x}$ be the associated covariate vector. The traditional single index regression model is

$$Y = m(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}_0) + \varepsilon, \tag{1.1}$$

where $m(\cdot)$ is an unknown regression function, $\boldsymbol{\beta}_0$ consists of unknown index parameters, and ε is a random error with $E(\varepsilon \mid \mathbf{x}) = 0$ and $\mathrm{var}(\varepsilon \mid \mathbf{x}) = \sigma^2$. This model has been well studied in the literature, for example, Powell, Stock, and Stoker (1989) and Härdle, Hall, and Ichimura (1993). Zhu, Huang, and Li (2012) studied the heteroscedastic single index regression model

$$Y = m(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}_0) + \sigma(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}_0)\varepsilon, \tag{1.2}$$

for unknown functions $m(\cdot)$ and $\sigma(\cdot)$, where ε has mean zero and is assumed independent of $\mathbf{x}$. Zhu, Huang, and Li (2012) developed an estimation procedure for $\boldsymbol{\beta}_0$ and $m(\cdot)$ under a quantile loss function when the dimension of $\mathbf{x}$ is finite.

In this paper, we focus on the ultrahigh dimensional situation, denote by p_n the dimension of $\mathbf{x}$ to emphasize the dependence of p_n on the sample size n. Denote by $F(y \mid \mathbf{x})$ the conditional distribution of Y given $\mathbf{x}$. We study a general class of single index models that includes models (1.1) and (1.2) as special cases. Specifically, we assume that there exists $\boldsymbol{\beta}_0 \in \mathbb{R}^{p_n}$ such that

$$F(y \mid \mathbf{x}) = F(y \mid \mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}_0), \text{ for all } y \in \mathbb{R}. \tag{1.3}$$

Here, the "curse of dimensionality" issue is avoided and model interpretability is maintained via a single index structure. Because the conditional distributional function $F(\cdot \mid \cdot)$ is unknown, the index parameter $\boldsymbol{\beta}_0$ is not identifiable. The direction of $\boldsymbol{\beta}_0$, instead of its true value, is our primary interest. We refer to (1.3) as a conditional distribution-based single index model (CDSIM for short) in order to distinguish it from (1.1) and (1.2).

When the covariate dimension is high, it is natural to assume that some covariates are irrelevant. The presence of irrelevant covariates may substantially deteriorate the precision of parameter estimation and the accuracy of response prediction (Altham (1984)). In the context of linear regression or generalized linear regression, many regularization methods, such as the LASSO (Tibshirani (1996)), the SCAD (Fan and Li (2001); Zou and Li (2008)), the adaptive LASSO (Zou (2006)), the MCP (Zhang (2010)), the hard thresholding penalty (Zheng, Fan, and Lv (2014)) and general penalty functions (Fan and Lv (2013)) have been proposed to remove irrelevant covariates and simultaneously estimate the nonzero coefficients. Naik and Tsai (2001), Kong and Xia (2007), Zhu, Qian, and Lin (2011) and Liang et al. (2010) developed some regularization methods for single index regression. Recently, Wang, Wu, and Li (2012) investigated nonconvex penalized quantile regression for analyzing heterogeneity in the ultrahigh-dimensional setting. Fan, Fan, and Barut (2014) proposed a two-step adaptive robust LASSO based on weighted L_1-penalized quantile regression to deal with heavy-tailed high-dimensional data.

We consider variable selection and feature screening for (1.3) when the covariate dimension p_n is ultrahigh. We further assume $\boldsymbol{\beta}_0$ is sparse. Denote by $\mathcal{A}$

the active index set, $\boldsymbol{\beta}_{\mathcal{A}}$ the nonzero entries of $\boldsymbol{\beta}_0$, and $\mathbf{x}_{\mathcal{A}}$ the collection of all active covariates. When $\boldsymbol{\beta}_0$ is sparse, (1.3) reduces to

$$F(y \mid \mathbf{x}) = F(y \mid \mathbf{x}_{\mathcal{A}}^{\mathrm{T}}\boldsymbol{\beta}_{\mathcal{A}}), \text{ for all } y \in \mathbb{R}. \tag{1.4}$$

Our goal is to identify $\mathcal{A}$ and if possible, to estimate $\boldsymbol{\beta}_{\mathcal{A}}$. To the best of our knowledge, there are few variable selection methods designed for models (1.3) or (1.4) with ultrahigh-dimensional covariates.

We introduce two approaches to accomplish our goal: a penalized linear quantile regression and an independence screening procedure. When (1.3) is true, the quantile functions of $(Y \mid \mathbf{x})$ vary with the realizations of $(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}_0)$. As the quantile function admits a single index structure, we implement a penalized quantile regression to exclude irrelevant covariates and simultaneously estimate the direction of $\boldsymbol{\beta}_0$. The advantage of using quantile regression is that the quantile function characterizes F at (1.3) and is resilient to outliers and extreme values in the response. We show that, although the true quantile functions of $(Y \mid \mathbf{x})$ are possibly nonlinear, the estimator obtained from penalized linear quantile regression remains consistent up to a proportionality constant. This strategy helps to reduce the computational complexity substantially in estimating (1.3) in that the linear quantile regression procedure avoids estimating nonlinear quantile functions; This is appealing for ultrahigh dimensional data analysis. We show that the penalized linear quantile regression estimate has the oracle property under mild regularity conditions, even when p_n tends to ∞ in an exponential rate in n.

From a practical perspective, when the covariate dimension is extremely large, penalized linear quantile regression has the drawbacks of computational inexpediency and algorithmic instability (Fan, Samworth, and Wu (2009)). To further reduce the computational complexity in selecting important covariates from ultrahigh dimensional candidates, we further introduce an independence screening procedure which ranks the importance of each covariate through its distance correlation with the marginal distribution F of the response at model (1.3) and the implicit model (1.4). Since F is bounded and monotone, we can reasonably expect that the procedure still works in the presence of outliers or extreme values in the response variable. It is computationally efficient and hence offers a useful complement, rather an alterative, to the penalized quantile regression approach since the proposed independence screening can precede the penalized quantile regression when the latter fails to produce a reliable solution within a tolerable time. Based on the reduced model, the penalized quantile regression may further refine selection of important covariates at different quantile levels. We show that this new independence screening procedure has the sure screening property even when p_n is ultrahigh.

The paper is organized as follows. In Section 2, we propose penalized linear quantile regression and study the consistency and the oracle property of the resulting estimator. We propose a robust independence screening procedure and establish its sure screening property in Section 3. We compare the finite sample performance of our proposals with several competitors in Section 4. Proofs are given in the Appendix.

2. Penalized Linear Quantile Regression

In this section, we construct an estimate for the direction of $\boldsymbol{\beta}_0$ at (1.3) via penalized linear quantile regression.

2.1. The methodology

Model (1.3) and its sparse structure (1.4) indicate that the quantile functions of $(Y \mid \mathbf{x})$ at different quantile levels are all functions of $(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}_0)$ and $(\mathbf{x}_{\mathcal{A}}^{\mathrm{T}}\boldsymbol{\beta}_{\mathcal{A}})$ if the sparsity principle applies. This motivates us to estimate $\boldsymbol{\beta}_0$ through quantile functions at different levels. Similar to Zhu, Huang, and Li (2012), we first show that linear quantile regression can be used to estimate the direction of $\boldsymbol{\beta}_0$ at (1.3). To be specific, let $\rho_\tau(r) = \tau r - rI(r < 0)$, the check loss function at the τth quantile, for $\tau \in (0,1)$. Let $\mathbf{b} = (b_1, \ldots, b_{p_n})^{\mathrm{T}} \in \mathbb{R}^{p_n}$ and put

$$\mathcal{L}_\tau(u, \mathbf{b}) = E\{\rho_\tau(Y - u - \mathbf{x}^{\mathrm{T}}\mathbf{b})\} \text{ and } (u_\tau, \boldsymbol{\beta}_\tau) = \underset{u,\mathbf{b}}{\operatorname{argmin}}\{\mathcal{L}_\tau(u, \mathbf{b})\}. \tag{2.1}$$

Lemma 1. *If $E\{\mathbf{x} - E(\mathbf{x}) \mid \mathbf{x}^{T}\boldsymbol{\beta}_0\} = var(\mathbf{x})\boldsymbol{\beta}_0\left\{\boldsymbol{\beta}_0^{T} var(\mathbf{x})\boldsymbol{\beta}_0\right\}^{-1}\boldsymbol{\beta}_0^{T}\{\mathbf{x} - E(\mathbf{x})\}$, then $\boldsymbol{\beta}_\tau$ is proportional to $\boldsymbol{\beta}_0$ at* (1.3).

This linearity condition is satisfied when $\mathbf{x}$ follows an elliptically contour distribution (Li (1991)). Hall and Li (1993) demonstrated that, regardless of the covariate distribution, the linearity condition always offers an ideal approximation to reality as long as p_n is sufficient large, and it is typically regarded as mild in an ultrahigh-dimensional setting. Lemma 1 implies that the indices of zero entries in $\boldsymbol{\beta}_0$ and $\boldsymbol{\beta}_\tau$ coincide. Estimating the direction of $\boldsymbol{\beta}_0$ at (1.3) amounts to estimating $\boldsymbol{\beta}_\tau$ at (2.1). This lemma can be proved using similar arguments as in Zhu, Huang, and Li (2012). We omit the proof.

When the covariate dimension is large, it is desirable to exclude irrelevant covariates and simultaneously estimate $\boldsymbol{\beta}_\tau$ at (2.1). Here $\boldsymbol{\beta}_\tau$ is identifiable because the linear quantile loss function $\mathcal{L}_\tau(u, \mathbf{b})$ is convex. Suppose that $\{(\mathbf{x}_i, Y_i), i = 1, 2, \ldots, n\}$ is a random sample from (1.3). We consider a penalized linear quantile regression to produce a sparse estimator of $\boldsymbol{\beta}_\tau$:

$$Q(u, \mathbf{b}) = n^{-1}\sum_{i=1}^{n}\rho_\tau(Y_i - u - \mathbf{x}_i^{\mathrm{T}}\mathbf{b}) + \sum_{j=1}^{p_n} p_\lambda(|b_j|), \tag{2.2}$$

where $p_\lambda(\cdot)$ is a penalty function with a regularization parameter λ. We use the SCAD penalty (Fan and Li (2001)) and the MCP penalty (Zhang (2010)). The MCP function is defined as

$$p_\lambda(b) = \lambda\left(|b| - \frac{b^2}{2a\lambda}\right) I(0 \le |b| < a\lambda) + \frac{a\lambda^2}{2} I(|b| \ge a\lambda),$$

where $a > 1$. The SCAD penalty is

$$\begin{aligned} p_\lambda(b) = {} & \lambda|b| I(0 \le |b| < \lambda) + \frac{a\lambda|b| - (b^2+\lambda^2)/2}{a-1} I(\lambda \le |b| \le a\lambda) \\ & + \frac{(a+1)\lambda^2}{2} I(|b| > a\lambda), \end{aligned}$$

where $a = 3.7$ is suggested by Fan and Li (2001). By minimizing the objective function $Q(u, \mathbf{b})$, we obtain the estimators $(\widehat{u}_\tau, \widehat{\boldsymbol{\beta}}_\tau)$ at the τ-th quantile, where

$$(\widehat{u}_\tau, \widehat{\boldsymbol{\beta}}_\tau) = \underset{u,\mathbf{b}}{\operatorname{argmin}}\{Q(u, \mathbf{b})\}. \tag{2.3}$$

2.2. The Oracle property

We study the oracle property of the estimators obtained from the penalized linear quantile regression. Without loss of generality, assume the first q_n components of $\mathbf{x}$ are active and the rest are inactive, where $q_n (\ll p_n)$ is a positive integer, so $\mathcal{A} = \{1, 2, \ldots, q_n\}$. We define the oracle estimator at the population level by

$$\mathcal{L}_\tau(u, \mathbf{b}_1) = E\{\rho_\tau(Y - u - \mathbf{x}_\mathcal{A}^{\mathrm{T}}\mathbf{b}_1)\} \ \text{ and } \ (u_\tau^o, \boldsymbol{\beta}_{\tau 1}^o) = \underset{u,\mathbf{b}_1}{\operatorname{argmin}}\{\mathcal{L}_\tau(u, \mathbf{b}_1)\}, \tag{2.4}$$

where $\mathbf{b}_1 = (b_1, \ldots, b_{q_n})^{\mathrm{T}} \in \mathbb{R}^{q_n}$. We further write $\boldsymbol{\beta}_\tau^o = (\boldsymbol{\beta}_{\tau 1}^{o\mathrm{T}}, \mathbf{0}^{\mathrm{T}})^{\mathrm{T}}$, where $\boldsymbol{\beta}_{\tau 1}^o$ represents a q_n-dimensional vector of nonzero components associated with the active covariates and $\mathbf{0}$ denotes a $(p_n - q_n)$-dimensional vector of zeros. Accordingly, we define the oracle estimator $\widehat{\boldsymbol{\beta}}_\tau^o = (\widehat{\boldsymbol{\beta}}_{\tau 1}^{o\mathrm{T}}, \mathbf{0}^{\mathrm{T}})^{\mathrm{T}}$ at the sample level by

$$\begin{aligned} & \mathcal{L}_{\tau n}(u, \mathbf{b}_1) = n^{-1}\sum_{i=1}^n \{\rho_\tau(Y_i - u - \mathbf{x}_{i,\mathcal{A}}^{\mathrm{T}}\mathbf{b}_1)\} \ \text{ and} \\ & (\widehat{u}_\tau^o, \widehat{\boldsymbol{\beta}}_{\tau 1}^o) = \underset{u,\mathbf{b}_1}{\operatorname{argmin}}\{\mathcal{L}_{\tau n}(u, \mathbf{b}_1)\}. \end{aligned} \tag{2.5}$$

We assume some regularity conditions.

(C1) There exist positive constants t_0 and C such that

$$\max_{1 \le k \le p_n} E\left\{\exp(t|X_k|)\right\} \le C < \infty, \ \text{ for } 0 < t \le t_0. \tag{2.6}$$

(C2) There exist positive constants $0 < C_1 \leq C_2 < \infty$, such that

$$C_1 \leq \lambda_{\min}\{E(\mathbf{x}_{\mathcal{A}}\mathbf{x}_{\mathcal{A}}^{\mathrm{T}})\} \leq \lambda_{\max}\{E(\mathbf{x}_{\mathcal{A}}\mathbf{x}_{\mathcal{A}}^{\mathrm{T}})\} \leq C_2,$$

where $\lambda_{\min}$ and $\lambda_{\max}$ represent the smallest and largest eigenvalues, respectively, where $\{(\mathbf{x}_{i,\mathcal{A}}, Y_i), i = 1, \ldots, n\}$ are in general position (Koenker (2005, Sec. 2.2)).

(C3) The probability density function of $(Y - \mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}_\tau)$ conditional on $\mathbf{x}$, denoted by $f(\cdot \mid \mathbf{x})$, is uniformly bounded away from 0 and ∞ in a neighborhood of u_τ^o.

(C4) The true model size q_n satisfies $q_n = O(n^{c_1})$ for $0 \leq c_1 < 1/2$.

(C5) For $\boldsymbol{\beta}_{\tau 1}^o = (\beta_{\tau,1}^o, \beta_{\tau,2}^o, \ldots, \beta_{\tau,q_n}^o)^{\mathrm{T}}$, there exist positive constants c_2 and C such that $2c_1 < c_2 \leq 1$ and $\min_{1\leq j\leq q_n} |\beta_{\tau,j}^o| \geq Cn^{-(1-c_2)/2}$.

Condition (C1) is concerned with the moments of the covariates; it holds when the covariates are bounded, or when $\mathbf{x}$ has a multivariate normal distribution. This condition is widely assumed in high dimensional inference. See, for instance, Bickel and Levina (2008). Condition (C2) requires that the design matrix of the true model at the population level be well behaved. Condition (C3) is a common assumption on the conditional distribution function of $(Y - \mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}_\tau)$ conditional on $\mathbf{x}$. Condition (C4) allows the sparsity size q_n can diverge as the sample size n goes to the infinity. Condition (C5) requires that the smallest true signal decay to zero at a slow rate.

Lemma 2. *Under* (C1)−(C4)*, the oracle estimators* $\widehat{u}_\tau^o$ *and* $\widehat{\boldsymbol{\beta}}_{\tau 1}^o$ *satisfy*

$$\|\widehat{\boldsymbol{\beta}}_{\tau 1}^o - \boldsymbol{\beta}_{\tau 1}^o\| = O_p\Big(\sqrt{\frac{q_n}{n}}\Big) \quad and \quad \|\widehat{u}_\tau^o - u_\tau^o\| = O_p\Big(\sqrt{\frac{q_n}{n}}\Big). \tag{2.7}$$

Theorem 1 (The Oracle Property)**.** *Suppose* (C1)−(C5) *hold, and* $\log p_n = o(n^{\min\{c_2-2\theta,\theta\}})$ *with* $0 < \theta < (c_2 - c_1)/2$ *and* $\lambda = o\{n^{-(1-c_2)/2}\}$*. Let* $\mathcal{B}_n(\lambda)$ *be the set of local minima* $\widehat{\boldsymbol{\beta}}_\tau$ *of the objective function* $Q(u, \mathbf{b})$ *defined at* (2.2) *with the SCAD or the MCP penalty and the tuning parameter* λ*. Then*

$$\Pr\Big\{\widehat{\boldsymbol{\beta}}_\tau^o \in \mathcal{B}_n(\lambda)\Big\} \to 1, \quad as \quad n \to \infty.$$

Theorem 1 has the oracle estimator $\widehat{\boldsymbol{\beta}}_\tau^o$ as a local minimizer of the objective function (2.2) with probability approaching one as $n \to \infty$. This result extends Theorem 2.4 of Wang, Wu, and Li (2012) from the linear quantile regression model to model (1.3). The results in Lemmas 1 and 2 and Theorem 1 imply that $\widehat{\boldsymbol{\beta}}_\tau$ from the penalized linear quantile regression is a consistent estimator of the direction of $\boldsymbol{\beta}_0$ at model (1.3). It can detect the non-zero components of $\boldsymbol{\beta}_0$ and simultaneously estimate its direction. From a technical perspective, Wang, Wu, and Li (2012) assumed all covariates uniformly bounded while in (C1) we only require that distributions of the covariates have sub-exponential tails. In practice, the linear quantile regression estimator obtained with the LASSO penalty can serve as an initial value in our algorithm to minimize the objective function $Q(u, \mathbf{b})$.

3. Robust SIS based on Distance Correlation

We propose a robust feature screening procedure for model (1.3) using distance correlation.

3.1. The methodology

We first review the definition of distance correlation (Szekely, Rizzo, and Bakirov (2007)). The distance covariance between random variables X and Y is

$$\mathrm{dcov}^2(X, Y) = S_1 + S_2 - 2S_3, \tag{3.1}$$

where $S_1 = E\big(|X-\widetilde{X}||Y-\widetilde{Y}|\big)$, $S_2 = E\big(|X-\widetilde{X}|\big)E\big(|Y-\widetilde{Y}|\big)$, $S_3 = E\big\{E\big(|X-\widetilde{X}| \mid X\big)E\big(|Y-\widetilde{Y}| \mid Y\big)\big\}$, and $(\widetilde{X}, \widetilde{Y})$ is an independent copy of (X, Y). The distance correlation between X and Y is

$$\mathrm{dcorr}(X, Y) = \frac{\mathrm{dcov}(X, Y)}{\sqrt{\mathrm{dcov}(X, Y)\mathrm{dcov}(Y, Y)}}. \tag{3.2}$$

Szekely, Rizzo, and Bakirov (2007) pointed out that $\mathrm{dcorr}(X, Y) = 0$ if and only if X and Y are independent and $\mathrm{dcorr}(X, Y)$ is strictly increasing in the absolute value of the Pearson correlation between X and Y. Motivated by these properties, Li, Zhong, and Zhu (2012) proposed a sure independence screening to rank all predictors using their distance correlations with the response variable, termed DC-SIS, and proved its sure screening property for ultrahigh-dimensional data.

We denote by X_k the kth predictor with $k = 1, \ldots, p_n$ and propose to quantify the importance of X_k through its distance correlation with the marginal distribution function of Y, denoted by $F(Y)$. That is,

$$\omega_k = \mathrm{dcorr}\{X_k, F(Y)\}, \tag{3.3}$$

where $F(y) = E\{\mathbf{1}(Y \le y)\}$ and $\mathbf{1}(\cdot)$ denotes an indicator function. This is a modification of the marginal utility in Li, Zhong, and Zhu (2012) in that here we use $F(Y)$ instead of Y.

The marginal utility at (3.3) has several advantages compared with existing measurements: $\mathrm{dcorr}\{X_k, F(Y)\} = 0$ if and only if X_k and Y are independent, and following Li, Zhong, and Zhu (2012), we can see that the screening procedure based on (3.3) is model-free and hence is applicable at (1.3) and (1.4); since $F(Y)$ is a bounded function for all types of Y, we can expect that the procedure using (3.3) has a reliable performance when the response is the heavy-tailed and when extreme values are present in the response values; If one suspects that the covariates also contain some extreme values, then one can use $\mathrm{dcorr}\{F_k(X_k), F(Y)\}$ to rank the importance of the X_k, where $F_k(x) = E\{\mathbf{1}(X_k \le x)\}$.

We now show how to implement the marginal utility (3.3) in the screening procedure. Let $\{(\mathbf{x}_i, Y_i), i = 1, \cdots, n\}$ be a random sample from the population $(\mathbf{x}, Y)$. We first estimate the distance covariance between X_k and $F(Y)$ through the moment estimation method,

$$\widehat{\mathrm{dcov}}^2\{X_k, F(Y)\} = \widehat{S}_{k,1} + \widehat{S}_{k,2} - 2\widehat{S}_{k,3}, \tag{3.4}$$

where

$$\widehat{S}_{k,1} = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} |X_{ik} - X_{jk}|\, |F_n(Y_i) - F_n(Y_j)|,$$

$$\widehat{S}_{k,2} = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} |X_{ik} - X_{jk}|\, \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} |F_n(Y_i) - F_n(Y_j)|, \text{ and}$$

$$\widehat{S}_{k,3} = \frac{1}{n^3}\sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{l=1}^{n} |X_{ik} - X_{lk}|\, |F_n(Y_j) - F_n(Y_l)|$$

are the corresponding estimators of $S_{k,1}$, $S_{k,2}$, $S_{k,3}$, and $F_n(y) = n^{-1}\sum_{i=1}^{n}\mathbf{1}(Y_i \le y)$. We estimate ω_k with

$$\widehat{\omega}_k = \widehat{\mathrm{dcorr}}\{X_k, F(Y)\} = \frac{\widehat{\mathrm{dcov}}(X_k, F(Y))}{\sqrt{\widehat{\mathrm{dcov}}(X_k, X_k)\widehat{\mathrm{dcov}}(F(Y), F(Y))}}. \tag{3.5}$$

Our independence screening procedure retains the covariates with the $\widehat{\omega}_k$ values larger than a user-specified threshold. Let $\widehat{\mathcal{A}} = \{k : \widehat{\omega}_k \ge cn^{-\kappa}, \text{ for } 1 \le k \le p_n\}$ for some pre-specified thresholds $c > 0$ and $0 \le \kappa < 1/2$. The constants c and κ control the signal strength and will be defined at (C6) below. We refer to this approach as the distance correlation based robust independence screening procedure (DC-RoSIS).

3.2. Sure screening property

We first state the consistency of $\widehat{\omega}_k$, which paves the road to proving the sure screening property of the DC-RoSIS procedure.

Theorem 2. *Under* (C1)*, for any* $0 < \gamma < 1/2 - \kappa$*, there exist positive constants* c_1 *and* c_2 *such that*

$$Pr\Big(\max_{1\le k\le p} |\widehat{\omega}_k - \omega_k| \ge cn^{-\kappa}\Big) \\ \le O\bigg(p\Big[\exp\Big\{-c_1 n^{1-2(\kappa+\gamma)}\Big\} + n\exp\left(-c_2 n^{\gamma}\right)\Big]\bigg). \tag{3.6}$$

We remark here that to derive the consistency of the estimated marginal utility, we do not need any moment condition on the response. To prove the sure screening property, we make of further assumption

(C6) The marginal utility at (3.3) satisfies $\min\limits_{k\in\mathcal{A}}\omega_k \ge 2cn^{-\kappa}$, for some constants $c > 0$ and $0 \le \kappa < 1/2$.

Condition (C6) allows the minimal signal of the active covariates to converge to zero as the sample size diverges, while it requires the minimum signal of active covariates be not too small.

Theorem 3 (Sure Screening Property)**.** *Under* (C6) *and the conditions in Theorem 2, it follows that*

$$Pr\left(\mathcal{A} \subseteq \widehat{\mathcal{A}}\right) \ge 1 - O\Big(s_n\Big[\exp\Big\{-c_1 n^{1-2(\kappa+\gamma)}\Big\} + n\exp\left(-c_2 n^{\gamma}\right)\Big]\Big), \tag{3.7}$$

where s_n *is the cardinality of* $\mathcal{A}$*. Thus,* $Pr\left(\mathcal{A} \subseteq \widehat{\mathcal{A}}\right) \to 1$ *as* $n \to \infty$*.*

4. Numerical Studies

We have conducted simulations to demonstrate the finite sample performance of our proposals. We further illustrate the proposed methodology through an empirical analysis of a real data example.

4.1. Simulations

In Example 1 we compare the performance of several independence screening procedures, and in Example 2 we assess the performance of penalized linear quantile regressions with different penalties and at different quantiles. Throughout the simulations we generated $\mathbf{x} = (X_1, X_2, \cdots, X_p)^{\mathrm{T}}$ from $\mathcal{N}(\mathbf{0}, \mathbf{\Sigma})$, where $\mathbf{\Sigma} = (\sigma_{ij})_{p\times p}$ with $\sigma_{ij} = 0.5^{|i-j|}$. We took $p = 1,000$ and $n = 200$.

Example 1. We compared the finite sample performance of DC-RoSIS with the existing procedures including SIS (Fan and Lv (2008)), SIRS (Zhu et al. (2011)), RRCS (Li et al. (2012)) and DC-SIS (Li, Zhong, and Zhu (2012)). We repeated each experiment 500 times and evaluated the performance with the the criteria $\mathcal{S}$, $\mathcal{P}_{sj}$, and $\mathcal{P}_a$. Here, $\mathcal{S}$ is the minimum model size to include all active covariates. We summarize the median of $\mathcal{S}$ with its associated robust estimate of the standard deviation (RSD = IQR/1.34). A smaller $\mathcal{S}$ value indicates a better performance. $\mathcal{P}_{sj}$ is the empirical probability that the active covariate X_j is selected for a given model size d. We set $d = 2[n/\log n]$ throughout. $\mathcal{P}_a$ is the empirical probability that all active covariates are selected for the given model size $d = 2[n/\log n]$. If the sure screening property holds true, both $\mathcal{P}_{sj}$ and $\mathcal{P}_a$ values are close to one when the estimated model size d is reasonably large.

We considered the models.

(1): $H(Y) = \mathbf{x}^{\mathrm{T}}\boldsymbol{\beta} + \varepsilon$,

(2): $Y = \exp(2 - \mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}/2) + (2 - \mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}/2)^2 + \exp(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}/2)\varepsilon$,

(3): $Y = \left\{1 + \exp\left(-3\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}\right)\right\}^{-1}\varepsilon$,

(4): $Y = \beta_1 X_1 + \beta_2 X_2 + \beta_7 X_7^2 + \varepsilon$,

where $\boldsymbol{\beta} = (3, 1.5, 0, 0, 0, 0, 2, 0, \ldots, 0)^{\mathrm{T}}$. In each model, only X_1, X_2 and X_7 are truly important. The random error ε was independently generated from either the standard normal or the standard Cauchy. In (1), $H(Y) = \{|Y|^{\lambda}\mathrm{sgn}(Y) - 1\}/\lambda$ is the Box-Cox transformation. These models were used in Li et al. (2012). We set $\lambda = 1$ and $\lambda = 0.25$. The single index $(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta})$ contributes both the conditional mean and variance of the response in (2), and are totally irrelevant to the mean regression function in model (3). The active covariate X_7 in (4) is quadratically related to the response. Though it is not a special case of model (1.3) or (1.4), we use it here to show that our procedure works quite well for a variety of regressions even when the model assumptions are violated.

The results are summarized in Table 1. It can be seen that SIS does not perform well when ε is Cauchy. Even when ε is normal, SIS still fails to behave well in the nonlinear models (3) and (4). SIRS performs very well for all single index models, but fails to identify X_7 as an important covariate in (4) because it is not capable of detecting symmetric patterns. The performance of RRCS is generally favorable for (1) and (2), but it hardly detects the active covariates that are only relevant to the conditional variance of the response in model (3) or that X_7 exhibits symmetric patterns with Y in (4). DC-RoSIS and DC-SIS have similar performances when ε is normal, and when ε is Cauchy, DC-RoSIS significantly improves DC-SIS. For example, in (1) with $\lambda = 0.25$, DC-SIS fails to detect the true relationship between two random variables when very extreme values are present.

Example 2. In this example, we examined the finite sample performance of the penalized linear quantile regression with different penalties, including LASSO (Tibshirani (1996)), SCAD (Fan and Li (2001)) and MCP (Zhang (2010)). We

Table 1. Performance comparison among different independence screening methods for four regression models with two different random errors.

		$\varepsilon \sim \mathcal{N}(0,1)$					$\varepsilon \sim$ Cauchy Distribution				
	Method	$\mathcal{S}$	$\mathcal{P}_{s1}$	$\mathcal{P}_{s2}$	$\mathcal{P}_{s7}$	$\mathcal{P}_a$	Size	$\mathcal{P}_{s1}$	$\mathcal{P}_{s2}$	$\mathcal{P}_{s7}$	$\mathcal{P}_a$
Model (1) ($\lambda = 1$)	SIS	3.0(0.0)	1.00	1.00	1.00	1.00	220.0(483.4)	0.67	0.62	0.49	0.39
	DC-SIS	3.0(0.0)	1.00	1.00	1.00	1.00	3.0(0.7)	0.98	0.98	0.95	0.95
	SIRS	3.0(0.0)	1.00	1.00	1.00	1.00	3.0(0.0)	1.00	1.00	1.00	1.00
	RRCS	3.0(0.0)	1.00	1.00	1.00	1.00	3.0(0.0)	1.00	1.00	1.00	1.00
	DC-RoSIS	3.0(0.0)	1.00	1.00	1.00	1.00	3.0(0.0)	1.00	1.00	1.00	1.00
Model (1) ($\lambda = 0.25$)	SIS	3.0(0.7)	1.00	1.00	0.99	0.99	794.5(210.5)	0.10	0.07	0.09	0.00
	DC-SIS	3.0(0.0)	1.00	1.00	1.00	1.00	702.5(246.6)	0.17	0.14	0.13	0.05
	SIRS	3.0(0.0)	1.00	1.00	1.00	1.00	3.0(0.0)	1.00	1.00	1.00	1.00
	RRCS	3.0(0.0)	1.00	1.00	1.00	1.00	3.0(0.0)	1.00	1.00	1.00	1.00
	DC-RoSIS	3.0(0.0)	1.00	1.00	1.00	1.00	3.0(0.0)	1.00	1.00	1.00	1.00
Model (2)	SIS	5.0(14.2)	0.99	0.99	0.90	0.90	29.0(122.4)	0.89	0.83	0.69	0.63
	DC-SIS	3.0(0.7)	1.00	1.00	0.99	0.99	3.0(2.9)	0.99	0.99	0.94	0.94
	SIRS	3.0(0.0)	1.00	1.00	1.00	1.00	3.0(0.7)	1.00	1.00	1.00	1.00
	RRCS	3.0(0.0)	1.00	1.00	1.00	1.00	3.0(0.7)	1.00	1.00	1.00	1.00
	DC-RoSIS	3.0(0.0)	1.00	1.00	1.00	1.00	3.0(0.7)	1.00	1.00	1.00	1.00
Model (3)	SIS	786.5(217.5)	0.08	0.06	0.07	0.00	791.0(213.2)	0.06	0.07	0.07	0.00
	DC-SIS	4.0(4.5)	1.00	1.00	0.97	0.97	70.0(130.8)	0.92	0.83	0.57	0.52
	SIRS	7.0(8.2)	1.00	1.00	0.99	0.99	8.0(8.9)	1.00	1.00	0.98	0.98
	RRCS	796.0(222.8)	0.10	0.10	0.09	0.00	782.0(253.3)	0.12	0.09	0.06	0.00
	DC-RoSIS	8.0(9.7)	1.00	1.00	0.96	0.96	9.0(11.9)	1.00	1.00	0.96	0.96
Model (4)	SIS	270.5(400.6)	1.00	1.00	0.25	0.25	594.0(358.0)	0.72	0.62	0.07	0.05
	DC-SIS	4.0(0.7)	1.00	1.00	1.00	1.00	8.0(18.7)	0.99	0.98	0.86	0.86
	SIRS	427.0(419.6)	1.00	1.00	0.11	0.11	493.5(387.7)	1.00	1.00	0.09	0.09
	RRCS	434.0(391.1)	1.00	1.00	0.13	0.13	477.5(394.0)	1.00	1.00	0.10	0.10
	DC-RoSIS	4.0(1.5)	1.00	1.00	1.00	1.00	6.0(5.2)	1.00	1.00	0.99	0.99

first utilized our procedure to select $d = 2[n/\log(n)]$ top ranked covariates and then applied penalized linear quantile regression to estimate the direction of $\boldsymbol{\beta}$. For conditional quantile regression, we considered three different quantiles $\tau = 0.25, 0.50$ and 0.75. Following Wang, Wu, and Li (2012), an additional independent data set of size $10n$ was generated to select the tuning parameter λ by minimizing the estimated prediction error based on the quantile check loss function.

We denoted the final estimator by $\widehat{\boldsymbol{\beta}}_\tau = (\widehat{\beta}_1, \widehat{\beta}_2, \ldots, \widehat{\beta}_p)^{\mathrm{T}}$. The coefficients of covariates removed by the screening procedure were directly set to be zero in the final estimator. Based on 100 repetitions, we evaluate the performance in terms of the following: Size: The average number of non-zero estimated regression coefficients $\widehat{\beta}_j \neq 0$ for $1 \leq j \leq p$; C: The average number of truly non-zero coefficients correctly estimated to be non-zero; IC: The average number of truly zero

Table 2. Simulation Results for Penalized Linear Quantile Regression at difference quantile levels ($25\%, 50\%$ and 75%) and with difference penalties (LASSO, SCAD, MCP).

$\varepsilon \sim \mathcal{N}(0,1)$				
Method	Size	C	IC	AE
LASSO($\tau = 0.25$)	18.16(6.28)	3.00(0.00)	15.16(6.28)	0.47(0.22)
LASSO($\tau = 0.50$)	18.14(6.33)	3.00(0.00)	15.14(6.33)	0.93(0.36)
LASSO($\tau = 0.75$)	13.97(6.16)	2.96(0.20)	11.01(6.15)	1.33(0.57)
SCAD($\tau = 0.25$)	3.46(0.86)	3.00(0.00)	0.46(0.86)	0.11(0.07)
SCAD($\tau = 0.50$)	3.68(1.58)	2.96(0.20)	0.72(1.56)	0.28(0.23)
SCAD($\tau = 0.75$)	3.47(1.58)	2.68(0.55)	0.79(1.52)	0.62(0.36)
MCP($\tau = 0.25$)	3.36(0.73)	3.00(0.00)	0.36(0.73)	0.11(0.07)
MCP($\tau = 0.50$)	3.53(1.23)	2.96(0.20)	0.57(1.21)	0.28(0.20)
MCP($\tau = 0.75$)	3.50(1.68)	2.68(0.55)	0.82(1.62)	0.63(0.36)
$\varepsilon \sim$ Cauchy Distribution				
Method	Size	C	IC	AE
LASSO($\tau = 0.25$)	23.75(6.63)	3.00(0.00)	20.75(6.63)	0.66(0.25)
LASSO($\tau = 0.50$)	19.29(7.66)	3.00(0.00)	16.29(7.66)	0.97(0.42)
LASSO($\tau = 0.75$)	14.01(6.89)	2.88(0.33)	11.13(6.82)	1.34(0.63)
SCAD($\tau = 0.25$)	3.56(1.19)	3.00(0.00)	0.56(1.19)	0.12(0.08)
SCAD($\tau = 0.50$)	3.66(1.36)	2.94(0.24)	0.72(1.36)	0.27(0.22)
SCAD($\tau = 0.75$)	3.33(1.91)	2.60(0.61)	0.73(1.84)	0.63(0.38)
MCP($\tau = 0.25$)	3.43(0.83)	3.00(0.00)	0.43(0.83)	0.11(0.07)
MCP($\tau = 0.50$)	3.70(1.67)	2.94(0.24)	0.76(1.66)	0.28(0.24)
MCP($\tau = 0.75$)	3.57(2.23)	2.64(0.53)	0.93(2.18)	0.65(0.42)

coefficients incorrectly estimated to be non-zero; AE: The average of absolute estimation error of $\widehat{\boldsymbol{\beta}}_\tau$, defined by $\sum_{j=1}^{p} \left| \widehat{\beta}_j \mathrm{sign}(\widehat{\beta}_{j,1})/\|\widehat{\boldsymbol{\beta}}_\tau\| - \beta_{0j}\mathrm{sign}(\beta_{0j,1})/\|\boldsymbol{\beta}_0\| \right|$.

We only report the results for model (2) in Example 1, which is a heteroscedastic single index model, as the results for other models lead to similar conclusion. The simulation results are charted in Table 2. In each column, the value represents the mean of 100 replicates with its sample standard deviation in the parentheses. For two random errors and different quantiles, the first three columns show that the LASSO is relatively conservative and tends to select larger models, while the SCAD and the MCP consistently select the true model. The relatively small values in the column labeled "AE" shows that our procedure can produce consistent estimators. The satisfactory simulation results suggest that the proposed robust two-stage procedure is indeed robust to the presence of heteroscedasticity and extreme values in the response.

4.2. An application

We conducted an empirical study of the Cardiomyopathy microarray dataset.

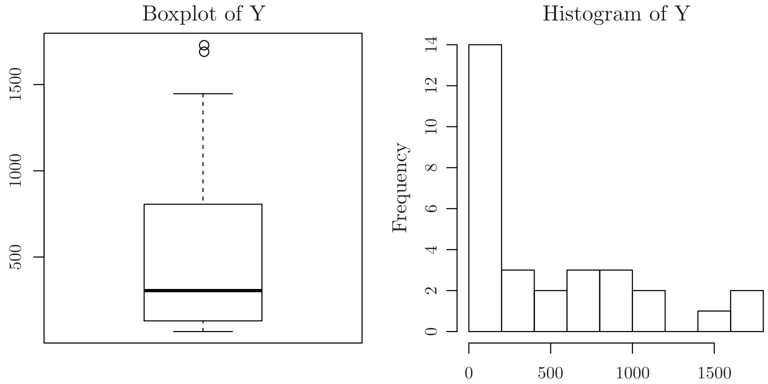

Figure 1. Exploratory data analysis: histogram and boxplot of Ro1.

This dataset was analyzed by Segal, Dahlquist, and Conklin (2003) , Hall and Miller (2009), and Li, Zhong, and Zhu (2012). The response variable is the genetic overexpression level of a G protein-coupled receptor (Ro1) in mice, which can sense molecules outside the cell and activate inside signal transduction pathways and cellular responses. The covariates are $6,319$ genetic expression levels. Only 30 specimens are observed. The main goal of the analysis is to determine the most influential genes for the response.

We display the boxplot and the histogram of Y in Figure 1. Both indicate that the response distribution may be heavy-tailed and the data contain outliers. We first implemented independence screening procedures to reduce the covariate dimension to the size of $2[n/\log n] = 16$. The performances of SIS and DC-SIS are similar to that of DC-RoSIS in this data analysis. We only present results of DC-RoSIS for regularized quantile regression with different penalties in this example. The DC-RoSIS ranks the genes, Msa.2877.0 and Msa.2134.0, in the top, which are same as the DC-SIS (Li, Zhong, and Zhu (2012)). The gene Msa.1166.0, identified by generalized correlation ranking (Hall and Miller (2009)), is also ranked in the top 10 by our screening procedure.

We further applied our procedure to the reduced model to estimate the direction of the index parameter and to simultaneously select important variables at different quantiles of the response. We chose the quantile levels $\tau = 0.25, 0.50$ and 0.75, and three different penalties, LASSO, SCAD and MCP. We used BIC to select the tuning parameters for each method. With the estimated single index, denoted $(\mathbf{x}^{\mathrm{T}}\widehat{\boldsymbol{\beta}}_{\tau})$, we applied cubic splines to estimate the quantile functions $\widehat{q}_{\tau}(\cdot)$ of model (1.3) or, equivalently, model (1.4). Figure 2 depicts the estimated curves of $\widehat{q}_{\tau}(\mathbf{x}^{\mathrm{T}}\widehat{\boldsymbol{\beta}}_{\tau})$ at different quantiles and for different penalties, which demonstrate the computational effectiveness of our proposals.

To compare the finite sample performances of different methods with different quantiles, we report the number of nonzero coefficients selected by each

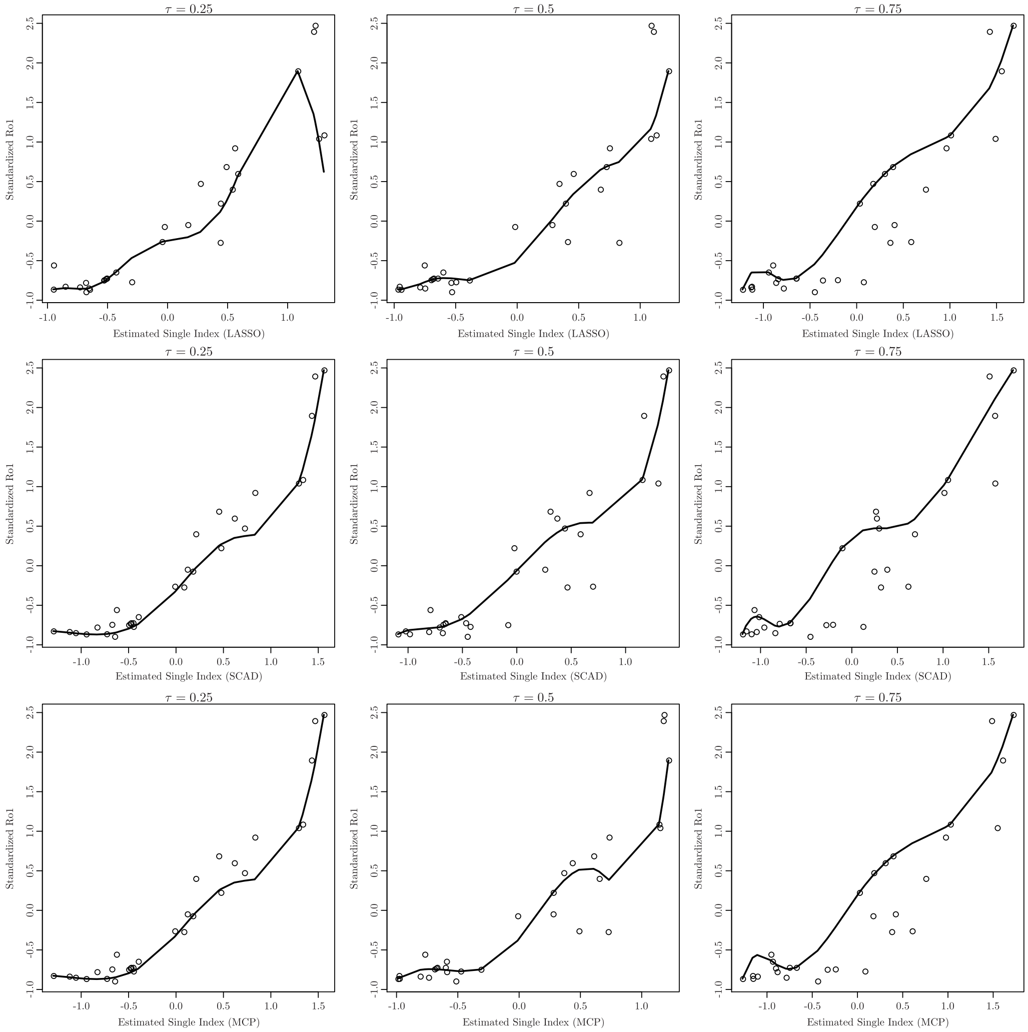

Figure 2. The estimated curves of $\widehat{q}_\tau(\mathbf{x}^{\mathrm{T}}\widehat{\boldsymbol{\beta}}_\tau)$ (the vertical axis) versus $(\mathbf{x}^{\mathrm{T}}\widehat{\boldsymbol{\beta}}_\tau)$ (the horizontal axis) at different quantiles for different penalties. From left to right, $\tau = 0.25$, 0.50 and 075; From up to down, LASSO, SCAD and MCP.

method, denoted by "Size" in Table 3. In addition, to evaluate the goodness of fit for each model, as with R^2 for the linear model, we take the quantile-adjusted R^2 ("Q-R^2") as

$$\text{Q-}R^2 = \left[1 - \frac{\sum_{i=1}^{n} \rho_\tau^2\{Y_i - \widehat{q}_\tau(X_i^{\mathrm{T}}\widehat{\boldsymbol{\beta}}_\tau)\}}{\sum_{i=1}^{n} \rho_\tau^2(Y_i - \widehat{Y}_\tau)}\right] \times 100\%, \tag{4.1}$$

where $\rho_\tau(\cdot)$ is the τth quantile check loss function, $\widehat{q}_\tau(\cdot)$ is the cubic-spline esti-

Table 3. Empirical analysis of Cardiomyopathy microarray dataset.

	All Data		Partitioned Data		
Method	Size	Q-R^2	Ave Size	Ave Q-R^2	PE
SCREEN($\tau = 0.25$)	16	97.6	16.00(0.00)	94.41(3.66)	0.48(0.21)
SCREEN($\tau = 0.50$)	16	93.1	16.00(0.00)	92.96(2.67)	0.67(0.27)
SCREEN($\tau = 0.75$)	16	94.4	16.00(0.00)	94.99(1.31)	0.59(0.33)
LASSO($\tau = 0.25$)	12	87.8	8.56(1.55)	78.59(10.75)	0.44(0.20)
LASSO($\tau = 0.50$)	8	89.1	7.21(1.54)	90.71(3.49)	0.55(0.17)
LASSO($\tau = 0.75$)	5	91.2	5.64(1.25)	94.54(1.71)	0.44(0.22)
SCAD($\tau = 0.25$)	10	96.9	8.29(2.81)	90.88(6.46)	0.44(0.18)
SCAD($\tau = 0.50$)	6	92.3	6.52(3.13)	92.33(3.24)	0.58(0.20)
SCAD($\tau = 0.75$)	3	93.0	3.82(1.46)	94.04(1.45)	0.50(0.25)
MCP($\tau = 0.25$)	10	96.9	8.69(2.64)	91.71(5.92)	0.43(0.14)
MCP($\tau = 0.50$)	5	89.3	6.91(2.81)	92.58(3.09)	0.55(0.26)
MCP($\tau = 0.75$)	4	92.8	4.13(1.64)	94.15(1.46)	0.51(0.28)

mate of $q_\tau(\cdot)$, $\widehat{q}_\tau(X_i^{\mathrm{T}}\widehat{\boldsymbol{\beta}}_\tau)$ is the τ-th quantile function of Y_i, and $\widehat{Y}_\tau$ is the sample τth quantile of Y. The larger Q-R^2 is, the better the model fit. For example, for $\tau = 0.75$, SCAD selected three covariates, which can explain 93.0% variance of the response in terms of the defined Q-R^2. As a benchmark, we also report the model with all 16 selected genes by our screening procedure, denoted by SCREEN in Table 3. In addition, we conducted 100 random partitions to examine prediction performance. For each partition, we randomly selected 90% of the data (27 observations) as the training set and the rest 10% (3 observations) as the test set. The average of the model sizes selected by each method, with its standard error across 100 partitions in the parenthesis, are reported in the third column ("Ave Size") of Table 3. In this table, we also report the average of quantile-adjusted R^2 for each method on the training set and its associated standard error, denoted by "Ave Q-R^2". The column labeled by "PE" denotes the median of prediction errors based on the quantile check loss function with the interquartile range/1.34 in the parentheses. We conclude that the penalized linear quantile regression improves both the model interpretability in terms of the model size and the model predictability in terms of the prediction errors.

5. Discussions

Our proposed method has reliable performance when the distribution of the response variable is heavily tailed or response realizations contain extreme values. A referee raised the question of the performance of the procedure in the presence of heavy-tail predictors or extreme outliers contained in the predictors. In this case, Condition (C1) is violated and the proposed method may fail. However, we can use $F_k(X_k)$, the distribution function of X_k, in place of X_k in the screening procedure. This replacement helps us remove (C1) and achieve the robustness feature in the x-direction; see Appendix A in the Supplement for more details. However, implementing penalized linear quantile regression when $\mathbf{x}$ contains outliers is not straightforward. How to remove condition (C1) in penalized linear quantile regression is an interesting topic for future research.

For statistical inference, one may be interested in the asymptotical distribution of the regularized quantile estimator and here we can adapt the idea of Theorem 2 in Wu and Liu (2009). They proved that the SCAD and adaptive-LASSO penalized linear quantile estimator is asymptotically normal if the number of important covariates is a fixed number. If the number of important covariates diverges to infinity, it is much more challenging to derive the asymptotic normality. This is an interesting research direction.

Supplementary Material

The supplementary material consists of three sections. In Section S1, we propose to quantify the importance of X_k through its distance correlation between the respective marginal distribution functions of X_k and Y. In Section S2 and S3, we provide more simulations.

Acknowledgement

Zhong's research was supported by National Natural Science Foundation of China (NNSFC) 11301435, 11401497, 71131008 and the Fundamental Research Funds for the Central Universities 20720140034. Zhu's research was supported by NNSFC 11371236 and 11422107, Henry Fok Education Foundation Fund of Young College Teachers 141002 and Innovative Research Team in University of China IRT13077, Ministry of Education of China. Li's research was supported by National Institute of Health (NIH) grants P50-DA10075, P50 DA036107, R01 CA168676, R01 MH096711 and NNSFC 11028103. Cui's research was supported by NNSFC 11071022, 11028103, 11231010 and 11471223, Key project of Beijing Municipal Educational Commission and Beijing Center for Mathematics and Information Interdisciplinary Sciences. The authors thank the editor, associate editor and three referees for their constructive comments, which have led to a significant improvement of the earlier version of this paper. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH or NNSFC.

Appendix: Proof of Theorems

Appendix A. Proof of Lemma 2

Lemma A.1. *The oracle estimator u_τ^o of u is the τth quantile of $Y - \mathbf{x}_{\mathcal{A}}^{\mathrm{T}}\boldsymbol{\beta}_{\tau 1}^o$ conditional on $\mathbf{x}_{\mathcal{A}}$, i.e. $E\{I(Y - \mathbf{x}_{\mathcal{A}}^{\mathrm{T}}\boldsymbol{\beta}_{\tau 1}^o \leq u_\tau^o) \mid \mathbf{x}_{\mathcal{A}}\} = \tau$.*

Proof of Lemma A.1. Let ξ_τ be the τth quantile of $Y - \mathbf{x}_{\mathcal{A}}^{\mathrm{T}}\boldsymbol{\beta}_{\tau 1}^o$ conditional on $\mathbf{x}_{\mathcal{A}}$. By definition, we have $E\{I(Y - \mathbf{x}_{\mathcal{A}}^{\mathrm{T}}\boldsymbol{\beta}_{\tau 1}^o \leq \xi_\tau) \mid \mathbf{x}_{\mathcal{A}}\} = \tau$. It suffices to show $\mathcal{L}_\tau(\xi_\tau, \boldsymbol{\beta}_{\tau 1}^o) \leq \mathcal{L}_\tau(u, \boldsymbol{\beta}_{\tau 1}^o)$ holds for any u. We have

$$\begin{aligned}
&\mathcal{L}_\tau(u, \boldsymbol{\beta}_{\tau 1}^o) - \mathcal{L}_\tau(\xi_\tau, \boldsymbol{\beta}_{\tau 1}^o) = E\{\rho_\tau(Y - u - \mathbf{x}_{\mathcal{A}}^{\mathrm{T}}\boldsymbol{\beta}_{\tau 1}^o)\} - E\{\rho_\tau(Y - \xi_\tau - \mathbf{x}_{\mathcal{A}}^{\mathrm{T}}\boldsymbol{\beta}_{\tau 1}^o)\} \\
&\quad = E\left[(u - \xi_\tau)\{I(Y - \xi_\tau - \mathbf{x}_{\mathcal{A}}^{\mathrm{T}}\boldsymbol{\beta}_{\tau 1}^o \leq 0) - \tau\}\right] \\
&\qquad + E\left[\int_0^{u-\xi_\tau} \{I(Y - \xi_\tau - \mathbf{x}_{\mathcal{A}}^{\mathrm{T}}\boldsymbol{\beta}_{\tau 1}^o \leq t) - I(Y - \xi_\tau - \mathbf{x}_{\mathcal{A}}^{\mathrm{T}}\boldsymbol{\beta}_{\tau 1}^o \leq 0)\}\, dt\right] \geq 0,
\end{aligned}$$

where the second equality follows from Knight (1998). Of the final two terms, the first is zero and the second is nonnegative. Thus $\xi_\tau = u_\tau^o$ and the desired conclusion follows.

Proof of Lemma 2. To prove Lemma 2, we borrow from He and Shao (2000) on M-estimation. It suffices to show that for any fixed $\eta > 0$, there exists two constants Δ_1 and Δ_2 such that for all sufficiently large n,

$$\Pr\left\{\inf_{\substack{\|\boldsymbol{\gamma}\|=\Delta_1 \\ |u|=\Delta_2}} \mathcal{L}_{\tau n}(u_\tau^o + n^{-1/2}q_n^{1/2}u, \boldsymbol{\beta}_{\tau 1}^o + n^{-1/2}q_n^{1/2}\boldsymbol{\gamma}) > \mathcal{L}_{\tau n}(u_\tau^o, \boldsymbol{\beta}_{\tau 1}^o)\right\} \geq 1 - \eta.$$

Here,

$$\begin{aligned}
&G_n(u, \boldsymbol{\gamma}) =: nq_n^{-1}\left\{\mathcal{L}_{\tau n}(u_\tau^o + n^{-1/2}q_n^{1/2}u, \boldsymbol{\beta}_{\tau 1}^o + n^{-1/2}q_n^{1/2}\boldsymbol{\gamma}) - \mathcal{L}_{\tau n}(u_\tau^o, \boldsymbol{\beta}_{\tau 1}^o)\right\} \\
&= q_n^{-1}\sum_{i=1}^n n^{-1/2}q_n^{1/2}(u + \mathbf{x}_{i,\mathcal{A}}^{\mathrm{T}}\boldsymbol{\gamma})\left\{I(Y_i - \mathbf{x}_{i,\mathcal{A}}^{\mathrm{T}}\boldsymbol{\beta}_{\tau 1}^o \leq u_\tau^o) - \tau\right\} \\
&+ q_n^{-1}\sum_{i=1}^n \int_0^{n^{-1/2}q_n^{1/2}(u + \mathbf{x}_{i,\mathcal{A}}^{\mathrm{T}}\boldsymbol{\gamma})} \left\{I(Y_i - \mathbf{x}_{i,\mathcal{A}}^{\mathrm{T}}\boldsymbol{\beta}_{\tau 1}^o \leq u_\tau^o + s) - I(Y_i - \mathbf{x}_{i,\mathcal{A}}^{\mathrm{T}}\boldsymbol{\beta}_{\tau 1}^o \leq u_\tau^o)\right\} ds \\
&=: I_{n1} + I_{n2},
\end{aligned}$$

where the second equality follows from Knight (1998)'s identity.
$E\{I(Y - \mathbf{x}_{\mathcal{A}}^{\mathrm{T}}\boldsymbol{\beta}_{\tau 1}^o \leq u_\tau^o) \mid \mathbf{x}_{\mathcal{A}}\} = \tau$ by Lemma A.1, so $E(I_{n1}) = 0$. Then,

$$\begin{aligned}
\mathrm{var}(I_{n1}) &= E\{\mathrm{var}(I_{n1} \mid \mathbf{x}_{\mathcal{A}})\} + \mathrm{var}\{E(I_{n1} \mid \mathbf{x}_{\mathcal{A}})\} \\
&= \tau(1-\tau)q_n^{-1}E\Big\{n^{-1}\sum_{i=1}^n (u + \mathbf{x}_{i,\mathcal{A}}^{\mathrm{T}}\boldsymbol{\gamma})^2\Big\}
\end{aligned}$$

$$\leq 2\tau(1-\tau)q_n^{-1}[u^2 + \lambda_{\max}\{E(\mathbf{x}_{\mathcal{A}}\mathbf{x}_{\mathcal{A}}^{\mathrm{T}})\}\,\|\boldsymbol{\gamma}\|^2] \leq Cq_n^{-1}(\Delta_1^2+\Delta_2^2),$$

where the last inequality follows by (C2). Thus, $I_{n1} = O_p\big(q_n^{-1/2}\big)\left(\Delta_1^2+\Delta_2^2\right)^{1/2}$.

We evaluate I_{n2}. Denote by $F(\cdot \mid \mathbf{x}_{\mathcal{A}})$ and $f(\cdot \mid \mathbf{x}_{\mathcal{A}})$ the conditional distribution and density of $(Y - \mathbf{x}_{\mathcal{A}}^{\mathrm{T}}\boldsymbol{\beta}_{\tau 1}^{o})$ given $\mathbf{x}_{\mathcal{A}}$, respectively

$$\begin{aligned}
E\left(I_{n2}\right) &= q_n^{-1}E\left[\sum_{i=1}^{n}\int_0^{n^{-1/2}q_n^{1/2}(u+\mathbf{x}_{i,\mathcal{A}}^{\mathrm{T}}\boldsymbol{\gamma})}\{F(u_\tau^o+s \mid \mathbf{x}_{i,\mathcal{A}}) - F(u_\tau^o \mid \mathbf{x}_{i,\mathcal{A}})\}\,ds\right]\\
&= q_n^{-1}E\left[\sum_{i=1}^{n}\int_0^{n^{-1/2}q_n^{1/2}(u+\mathbf{x}_{i,\mathcal{A}}^{\mathrm{T}}\boldsymbol{\gamma})} f(u_\tau^o+s' \mid \mathbf{x}_{i,\mathcal{A}})s ds\right]\\
&\geq Cq_n^{-1}E\left[\sum_{i=1}^{n}\left\{n^{-1/2}q_n^{1/2}(u+\mathbf{x}_{i,\mathcal{A}}^{\mathrm{T}}\boldsymbol{\gamma})\right\}^2\right]\\
&= CE(u+\mathbf{x}_{\mathcal{A}}^{\mathrm{T}}\boldsymbol{\gamma})^2 \geq C\left[1+\lambda_{\min}\{E(\mathbf{x}_{\mathcal{A}}\mathbf{x}_{\mathcal{A}}^{\mathrm{T}})\}\right](u^2+\|\boldsymbol{\gamma}\|^2) \geq C(\Delta_1^2+\Delta_2^2),
\end{aligned}$$

where the first inequality follows by (C3) and the last inequality follows by (C2). Therefore, $E\left(I_{n2}\right) = O(1)(\Delta_1^2+\Delta_2^2)$. Next we consider the variance of I_{n2},

$$\begin{aligned}
&\mathrm{var}\left(I_{n2}\right)\\
&\leq nq_n^{-2}E\left[\int_0^{n^{-1/2}q_n^{1/2}(u+\mathbf{x}_{\mathcal{A}}^{\mathrm{T}}\boldsymbol{\gamma})}\{I(Y-\mathbf{x}_{\mathcal{A}}^{\mathrm{T}}\boldsymbol{\beta}_{\tau 1}^{o}\leq u_\tau^o+s) - I(Y-\mathbf{x}_{\mathcal{A}}^{\mathrm{T}}\boldsymbol{\beta}_{\tau 1}^{o}\leq u_\tau^o)\}\,ds\right]^2\\
&\leq nq_n^{-2}E\{n^{-1/2}q_n^{1/2}(u+\mathbf{x}_{\mathcal{A}}^{\mathrm{T}}\boldsymbol{\gamma})\}^2 \leq q_n^{-1}\left[1+\lambda_{\min}\{E(\mathbf{x}_{\mathcal{A}}\mathbf{x}_{\mathcal{A}}^{\mathrm{T}})\}\right](u^2+\|\boldsymbol{\gamma}\|^2)\\
&\leq O(q_n^{-1})(\Delta_1^2+\Delta_2^2),
\end{aligned}$$

which converges to zero as $n\to\infty$ because $q_n = O(n^{c_1})$. This indicates that $|I_{n2} - E(I_{n2})| = o_p(1)$ by Chebyshev's inequality. Since I_{n2} is always nonnegative,

$$I_{n2} = E(I_{n2}) + o_p(1) \geq C(\Delta_1^2+\Delta_2^2) + o_p(1).$$

For sufficiently large Δ_1 and Δ_2, I_{n2} dominates I_{n1} asymptotically as $n\to\infty$. Therefore, for any fixed $\eta > 0$, there exists two constants Δ_1 and Δ_2 such that for all sufficiently large n, we have $G_n(u,\boldsymbol{\gamma}) > 0$ with probability at least $1-\eta$.

Appendix B. Proof of Theorem 1

We follow the idea of the proof of Theorem 2.4 in Wang, Wu, and Li (2012). Their moment conditions on $\mathbf{x}$ are different. With a slight notational abuse, we write $\mathbf{x}_{\mathcal{A}} = (1,\mathbf{x}_{\mathcal{A}})^{\mathrm{T}}$, $\boldsymbol{\beta}_\tau^o = (u_\tau^o, \boldsymbol{\beta}_\tau^{o\mathrm{T}})^{\mathrm{T}}$ as at (2.4), $\widehat{\boldsymbol{\beta}}_\tau = (\widehat{u}_\tau, \widehat{\boldsymbol{\beta}}_\tau^{\mathrm{T}})^{\mathrm{T}}$, and $\widehat{\boldsymbol{\beta}}_\tau^{o} = (\widehat{u}_\tau^o, \widehat{\boldsymbol{\beta}}_\tau^{o\mathrm{T}})^{\mathrm{T}}$, where $\widehat{\boldsymbol{\beta}}_\tau$ denotes the penalized linear quantile estimator at (2.3) and $\widehat{\boldsymbol{\beta}}_\tau^{o} = (\widehat{\boldsymbol{\beta}}_{\tau 1}^{o\mathrm{T}}, \mathbf{0}^{\mathrm{T}})^{\mathrm{T}}$ is the oracle estimator at (2.5). Accordingly, we write $\boldsymbol{\beta}_{\tau 1}^{o} = (u_\tau^o, \boldsymbol{\beta}_{\tau 1}^{o\mathrm{T}})^{\mathrm{T}}$ and $\widehat{\boldsymbol{\beta}}_{\tau 1}^{o} = (\widehat{u}_\tau^o, \widehat{\boldsymbol{\beta}}_{\tau 1}^{o\mathrm{T}})^{\mathrm{T}}$.

We first write the objective function (2.2) of the penalized linear quantile regression as the difference of two convex functions in $\boldsymbol{\beta}$. Here, we only consider the proof for the SCAD penalty, the proof for the MCP penalty can be achieved by the similar arguments. We have $Q(\boldsymbol{\beta}) = g(\boldsymbol{\beta}) - h(\boldsymbol{\beta})$, where $g(\boldsymbol{\beta}) = n^{-1}\sum_{i=1}^{n}\rho_\tau(Y_i - \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}) + \lambda\sum_{j=1}^{p_n}|\beta_j|$, and $h(\boldsymbol{\beta}) = \sum_{j=1}^{p_n} H_\lambda(\beta_j)$, with

$$H_\lambda(\beta_j) = \begin{cases} 0, & 0 \le |\beta_j| < \lambda; \\ \frac{(\beta_j^2 - 2\lambda|\beta_j| + \lambda^2)}{2(a-1)}, & \lambda \le |\beta_j| \le a\lambda; \\ \lambda|\beta_j| - \frac{(a+1)\lambda^2}{2}, & |\beta_j| > a\lambda. \end{cases}$$

Thus, the subdifferential of $h(\boldsymbol{\beta})$ at any $\boldsymbol{\beta}$ is

$$\partial h(\boldsymbol{\beta}) = \left\{\boldsymbol{\mu} = (\mu_0, \mu_1, \ldots, \mu_{p_n})^{\mathrm{T}} \in \mathbb{R}^{p_n+1} : \mu_0 = 0, \mu_j = \frac{\partial h(\boldsymbol{\beta})}{\partial \beta_j}, j=1,\ldots,p_n\right\}.$$

The subdifferential of $g(\boldsymbol{\beta})$ at any $\boldsymbol{\beta}$ is

$$\partial g(\boldsymbol{\beta}) = \Big\{\boldsymbol{\xi} = (\xi_0, \xi_1, \ldots, \xi_{p_n})^{\mathrm{T}} \in \mathbb{R}^{p_n+1} : \xi_j = (1-\tau)n^{-1}\sum_{i=1}^{n} X_{ij} I(Y_i - \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta} < 0)$$
$$-\tau n^{-1}\sum_{i=1}^{n} X_{ij} I(Y_i - \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta} > 0) - n^{-1}\sum_{i=1}^{n} X_{ij} v_i + \lambda l_j\Big\},$$

where $v_i = 0$ if $Y_i - \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta} \neq 0$ and $v_i \in [\tau - 1, \tau]$ otherwise; $l_0 = 0$; $l_j = \mathrm{sgn}(\beta_j)$ if $\beta_j \neq 0$ and $l_j \in [-1, 1]$ otherwise, for $1 \le j \le p_n$.

Let $s(\widehat{\boldsymbol{\beta}}) = \big\{s_0(\widehat{\boldsymbol{\beta}}), s_1(\widehat{\boldsymbol{\beta}}), \ldots, s_{p_n}(\widehat{\boldsymbol{\beta}})\big\}^{\mathrm{T}}$ be the set of the subgradient functions for the unpenalized quantile regression, where

$$s_j(\boldsymbol{\beta}) = (1-\tau)n^{-1}\sum_{i=1}^{n} X_{ij} I(Y_i - \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta} < 0) - \tau n^{-1}\sum_{i=1}^{n} X_{ij} I(Y_i - \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta} > 0)$$
$$-n^{-1}\sum_{i=1}^{n} X_{ij} v_i,$$

where $v_i = 0$ if $Y_i - \mathbf{x}_i^{\mathrm{T}}\widehat{\boldsymbol{\beta}} \neq 0$ and $v_i \in [\tau - 1, \tau]$ otherwise.

Lemmas B.1, B.2 and B.3 facilitate the proof of Theorem 1. Tao and An (1997) proposed the numerical algorithm based on the convex difference representation of Lemma B.1. Lemmas B.2 and B.3 characterize the properties of the oracle estimator $\widehat{\boldsymbol{\beta}}_\tau^o$ and the associated subgradient functions $s(\widehat{\boldsymbol{\beta}}_\tau^o)$ respectively.

Lemma B.1 (Difference Convex Program). *$g(\mathbf{x})$ and $h(\mathbf{x})$ are two convex functions. Let $\mathbf{x}^*$ be a point that admits a neighborhood U such that $\partial h(\mathbf{x}) \cap \partial g(\mathbf{x}^*) \neq \emptyset$, $\forall \mathbf{x} \in U \cap dom(g)$. Then $\mathbf{x}^*$ is a local minimizer of $g(\mathbf{x}) - h(\mathbf{x})$.*

Lemma B.2. *Suppose* (C4)−(C5) *holds and* $\lambda = o(n^{-(1-c_2)/2})$. *For the oracle estimator* $\widehat{\boldsymbol{\beta}}_\tau^o$, *there exist* v_i^* *with* $v_i^* = 0$ *if* $Y_i - \mathbf{x}_i^{\mathrm{T}}\widehat{\boldsymbol{\beta}}_\tau^o \neq 0$ *and* $v_i^* \in [\tau - 1, \tau]$ *otherwise, such that, with probability approaching one, we have*

$$s_j(\widehat{\boldsymbol{\beta}}_\tau^o) = 0, j = 0, 1, \ldots, q_n, \quad \text{and} \quad |\widehat{\beta}_j^{\,o}| \geq (a + \frac{1}{2})\lambda, \quad j = 1, \ldots, q_n.$$

Proof of Lemma B.2. The unpenalized quantile loss objective function is convex. By convex optimization theory, $\mathbf{0} \in \partial \sum_{i=1}^n \rho_\tau(Y_i - \mathbf{x}_i^{\mathrm{T}}\widehat{\boldsymbol{\beta}}_\tau^o)$. Therefore, there exists v_i^* such that $s_j(\widehat{\boldsymbol{\beta}}_\tau^o) = 0$ with $v_i = v_i^*$ for $j = 0, 1, \ldots, q_n$. On the other hand,

$$\min_{1\leq j\leq q_n} |\widehat{\beta}_j^{\,o}| \geq \min_{1\leq j\leq q_n} |\beta_{\tau,j}^o| - \max_{1\leq j\leq q_n} |\widehat{\beta}_j^{\,o} - \beta_{\tau,j}^o|.$$

Condition (C5) requires that $\min_{1\leq j\leq q_n} |\beta_{\tau,j}^o| \geq Cn^{-(1-c_2)/2}$. In addition, $\max_{1\leq j\leq q_n} |\widehat{\beta}_j^{\,o} - \beta_{\tau,j}^o| \leq \|\widehat{\boldsymbol{\beta}}_\tau^o - \boldsymbol{\beta}_{\tau 1}^o\| = O_p(\sqrt{q_n/n}) = O_p(n^{-(1-c_1)/2}) = o_p(n^{-(1-c_2)/2})$. Therefore, $\min_{1\leq j\leq q_n} |\widehat{\beta}_j^{\,o}| \geq Cn^{-(1-c_2)/2} - o_p(n^{-(1-c_2)/2})$, where c_1 and c_2 are defined at (C4) and (C5), respectively. For $\lambda = o\{n^{-(1-c_2)/2}\}$, we have that, with probability approaching one, $|\widehat{\beta}_j^{\,o}| \geq (a + 1/2)\lambda, j = 1, \ldots, q_n$, which completes the proof.

Lemma B.3. *Suppose* (C1)−(C5) *hold,* $\lambda = o(n^{-(1-c_2)/2})$, *and* $\log p_n = o(n^{\min\{c_2 - 2\theta, \theta\}})$ *with some constant* $0 < \theta < (c_2 - c_1)/2$. *For the oracle estimator* $\widehat{\boldsymbol{\beta}}_\tau^o$ *and the* $s_j(\widehat{\boldsymbol{\beta}}_\tau^o)$, *with probability approaching one, we have*

$$|s_j(\widehat{\boldsymbol{\beta}}_\tau^o)| \leq \lambda, \quad \text{and} \quad |\widehat{\beta}_j^{\,o}| = 0, j = q_n + 1, \ldots, p_n.$$

Proof of Lemma B.3. Since $\widehat{\boldsymbol{\beta}}_\tau^o$ is the oracle estimator, $|\widehat{\beta}_j^{\,o}| = 0, j = q_n + 1, \ldots, p_n$. It remains to show that

$$\Pr\Big(|s_j(\widehat{\boldsymbol{\beta}}_\tau^o)| > \lambda, \quad \text{for some } j = q_n + 1, \ldots, p_n\Big) \to 0, \quad \text{as } n \to \infty.$$

Let $\mathcal{D} = \{i : Y_i - \mathbf{x}_i^{\mathrm{T}}\widehat{\boldsymbol{\beta}}_\tau^o = 0\} = \{i : Y_i - \mathbf{x}_{i,\mathcal{A}}^{\mathrm{T}}\widehat{\boldsymbol{\beta}}_{\tau 1}^o = 0\}$, then for $j = q_n + 1, \ldots, p_n$,

$$\begin{aligned} s_j(\widehat{\boldsymbol{\beta}}_\tau^o) &= (1-\tau)n^{-1}\sum_{i=1}^n X_{ij} I(Y_i - \mathbf{x}_i^{\mathrm{T}}\widehat{\boldsymbol{\beta}}_\tau^o < 0) \\ &\quad -\tau n^{-1}\sum_{i=1}^n X_{ij} I(Y_i - \mathbf{x}_i^{\mathrm{T}}\widehat{\boldsymbol{\beta}}_\tau^o > 0) - n^{-1}\sum_{i=1}^n X_{ij} v_i, \\ &= n^{-1}\sum_{i=1}^n X_{ij}\big\{I(Y_i - \mathbf{x}_i^{\mathrm{T}}\widehat{\boldsymbol{\beta}}_\tau^o \leq 0) - \tau\big\} \\ &\quad -n^{-1}\sum_{i=1}^n X_{ij}\big\{v_i + (1-\tau)I(Y_i - \mathbf{x}_i^{\mathrm{T}}\widehat{\boldsymbol{\beta}}_\tau^o = 0)\big\} \end{aligned}$$

$$= n^{-1}\sum_{i=1}^{n} X_{ij}\big\{I(Y_i - \mathbf{x}_{i,\mathcal{A}}^{\mathrm{T}}\widehat{\boldsymbol{\beta}}_{\tau 1}^{o} \le 0) - \tau\big\} - n^{-1}\sum_{i\in\mathcal{D}} X_{ij}[v_i^* + (1-\tau)],$$

where $v_i^* \in [\tau-1, \tau]$ with $i \in \mathcal{D}$ satisfies $s_j(\widehat{\boldsymbol{\beta}}_{\tau}^{o}) = 0$ with $v_i = v_i^*$, for $j = 1, \ldots, q_n$, by Lemma B.2.

$$\begin{aligned}
&\Pr(|s_j(\widehat{\boldsymbol{\beta}}_{\tau}^{o})| > 2\lambda, \quad \text{for some } j = q_n + 1, \ldots, p_n)\\
&\le \Pr\Big(\Big|n^{-1}\sum_{i=1}^{n} X_{ij}\big\{I(Y_i - \mathbf{x}_{i,\mathcal{A}}^{\mathrm{T}}\widehat{\boldsymbol{\beta}}_{\tau 1}^{o} \le 0) - \tau\big\}\Big| > \lambda, \text{ for some } j = q_n+1, \ldots, p_n\Big)\\
&\quad + \Pr\Big(\Big|n^{-1}\sum_{i\in\mathcal{D}} X_{ij}\{v_i^* + (1-\tau)\}\Big| > \lambda, \quad \text{for some } j = q_n + 1, \ldots, p_n\Big)\\
&=: T_{n1} + T_{n2}.
\end{aligned}$$

First, we deal with T_{n2}. Let $M = O(n^{\theta})$ with some constant $0 < \theta < (c_2 - c_1)/2$. We have

$$\begin{aligned}
T_{n2} &\le \Pr\Big(\max_{j=q_n+1,\ldots,p_n}\Big|n^{-1}\sum_{i\in\mathcal{D}} X_{ij}\mathbf{1}\{|X_{ij}| \le M\}\{v_i^* + (1-\tau)\}\Big| > \frac{\lambda}{2}\Big)\\
&\quad + \Pr\Big(\max_{j=q_n+1,\ldots,p_n}\Big|n^{-1}\sum_{i\in\mathcal{D}} X_{ij}\mathbf{1}\{|X_{ij}| > M\}[v_i^* + (1-\tau)]\Big| > \frac{\lambda}{2}\Big)\\
&=: T_{n21} + T_{n22}.
\end{aligned}$$

Since $(\mathbf{x}_{i,\mathcal{A}}, Y_i)$ are in general positions Koenker (2005, Sec. 2.2), with probability tending to one there exists exactly $q_n + 1$ elements in $\mathcal{D}$. Thus, with probability tending to one,

$$\begin{aligned}
&\max_{q_n+1,\ldots,p_n}\Big|n^{-1}\sum_{i\in\mathcal{D}} X_{ij}\mathbf{1}\{|X_{ij}| \le M\}\{v_i^* + (1-\tau)\}\Big|\\
&\le M(q_n + 1)n^{-1} = O(n^{\theta + c_1 - 1}) = o(\lambda),
\end{aligned}$$

where the last equality holds for $\lambda = o(n^{-(1-c_2)/2})$ and $0 < \theta < (c_2 - c_1)/2$. Therefore, $T_{n21} \to 0$ as $n \to \infty$. For T_{n22}, the events satisfy

$$\Big\{\Big|n^{-1}\sum_{i\in\mathcal{D}} X_{ij}\mathbf{1}\{|X_{ij}| > M\}[v_i^* + (1-\tau)]\Big| > \frac{\lambda}{2}\Big\} \subseteq \{|X_{ij}| > M, \text{ for some } i \in \mathcal{D}\},$$

because if $|X_{ij}| \le M$ for all $i \in \mathcal{D}$, then $n^{-1}\sum_{i\in\mathcal{D}} X_{ij}\mathbf{1}\{|X_{ij}| > M\} = 0$. Therefore,

$$T_{n22} \le p_n \max_{j=q_n+1,\ldots,p_n} \Pr\Big(\Big|n^{-1}\sum_{i\in\mathcal{D}} X_{ij}\mathbf{1}\{|X_{ij}| > M\}[v_i^* + (1-\tau)]\Big| > \frac{\lambda}{2}\Big)$$

$$\begin{aligned}&\le p_n(q_n+1)\max_{i\in\mathcal{D},q_n+1\le j\le p_n}\Pr\left(|X_{ij}|>M\right)\\&\le p_n(q_n+1)\exp(-tM)E\{\exp(t|X_{ij}|)\}\\&\le Cp_n(q_n+1)\exp(-tM)=Cp_nO(n^{c_1})\exp(-tn^{\theta})\to 0,\end{aligned}$$

as $n\to\infty$, where $\log p_n=o(n^{\min\{c_2-2\theta,\theta\}})$ with some constant $0<\theta<(c_2-c_1)/2$, $0<t\le t_0$. The third inequality here holds from Markov's inequality and the fourth follows from (C1). Therefore, $T_{n2}=T_{n21}+T_{n22}\to 0,$ as $n\to\infty$.

It remains to show that

$$\Pr\left(\left|n^{-1}\sum_{i=1}^{n}X_{ij}\{I(Y_i-\mathbf{x}_i^{\mathrm{T}}\widehat{\boldsymbol\beta}_\tau^{o}<0)-\tau\}\right|>\lambda,\ \text{ for some } j=q_n+1,\ldots,p_n\right)\to 0,$$

as $n\to\infty$. We consider

$$\begin{aligned}T_{n1}\le{}&\Pr\left(\max_{j=q_n+1,\ldots,p_n}\left|n^{-1}\sum_{i=1}^{n}X_{ij}\{I(Y_i-\mathbf{x}_{i,\mathcal{A}}^{\mathrm{T}}\boldsymbol\beta_{\tau1}^{o}\le 0)-\tau\}\right|>\frac{\lambda}{2}\right)\\&+\Pr\Bigg(\max_{j=q_n+1,\ldots,p_n}\sup_{\|\boldsymbol\beta_1-\boldsymbol\beta_{\tau1}^{o}\|\le\Delta\sqrt{q_n/n}}\Bigg|n^{-1}\sum_{i=1}^{n}X_{ij}\Big[I(Y_i-\mathbf{x}_{i,\mathcal{A}}^{\mathrm{T}}\boldsymbol\beta_1\le 0)\\&-I(Y_i-\mathbf{x}_{i,\mathcal{A}}^{\mathrm{T}}\boldsymbol\beta_{\tau1}^{o}\le 0)-\big\{\Pr(Y_i-\mathbf{x}_{i,\mathcal{A}}^{\mathrm{T}}\boldsymbol\beta_1\le 0)-\Pr(Y_i-\mathbf{x}_{i,\mathcal{A}}^{\mathrm{T}}\boldsymbol\beta_{\tau1}^{o}\le 0)\big\}\Big]\Bigg|>\frac{\lambda}{4}\Bigg)\\&+\Pr\Bigg(\max_{j=q_n+1,\ldots,p_n}\sup_{\|\boldsymbol\beta_1-\boldsymbol\beta_{\tau1}^{o}\|\le\Delta\sqrt{q_n/n}}\Bigg|n^{-1}\sum_{i=1}^{n}X_{ij}\\&\qquad\Big\{\Pr(Y_i-\mathbf{x}_{i,\mathcal{A}}^{\mathrm{T}}\boldsymbol\beta_1\le 0)-\Pr(Y_i-\mathbf{x}_{i,\mathcal{A}}^{\mathrm{T}}\boldsymbol\beta_{\tau1}^{o}\le 0)\Big\}\Bigg|>\frac{\lambda}{4}\Bigg)\\&=:J_{n1}+J_{n2}+J_{n3}.\end{aligned}$$

For J_{n1}, let $M=O(n^{\theta})$ with $0<\theta<(c_2-c_1)/2$, then

$$\begin{aligned}J_{n1}\le{}&\Pr\left(\max_{j=q_n+1,\ldots,p_n}\left|n^{-1}\sum_{i=1}^{n}X_{ij}\mathbf{1}\{|X_{ij}|\le M\}\{I(Y_i-\mathbf{x}_{i,\mathcal{A}}^{\mathrm{T}}\boldsymbol\beta_{\tau1}^{o}<0)-\tau\}\right|>\frac{\lambda}{4}\right)\\&+\Pr\left(\max_{j=q_n+1,\ldots,p_n}\left|n^{-1}\sum_{i=1}^{n}X_{ij}\mathbf{1}\{|X_{ij}|>M\}\{I(Y_i-\mathbf{x}_{i,\mathcal{A}}^{\mathrm{T}}\boldsymbol\beta_{\tau1}^{o}\le 0)-\tau\}\right|>\frac{\lambda}{4}\right)\\&=:J_{n11}+J_{n12}.\end{aligned}$$

By Hoeffding's inequality,

$$\begin{aligned}&\Pr\left(\left|n^{-1}\sum_{i=1}^{n}X_{ij}\mathbf{1}(|X_{ij}|\le M)\{I(Y_i-\mathbf{x}_{i,\mathcal{A}}^{\mathrm{T}}\boldsymbol\beta_{\tau1}^{o}\le 0)-\tau\}\right|>\frac{\lambda}{4}\right)\\&\quad\le 2\exp\left(-\frac{n\lambda^2}{8M^2}\right).\end{aligned}$$

Thus, $J_{n11} \leq 2p_n \exp\{-n\lambda^2/(8M^2)\} = 2p_n \exp(-n^{1-2\theta}\lambda^2/8) \to 0$, as $n \to \infty$, because $\log p_n = o(n^{\min\{c_2-2\theta,\theta\}})$ with some constant $0 < \theta < (c_2 - c_1)/2$ and $\lambda = o\{n^{-(1-c_2)/2}\}$. On the other hand, we can similarly follow the arguments that deal with T_{n22} and have that

$$\begin{aligned} J_{n12} &\leq p_n \max_{j=q_n+1,\ldots,p_n} \Pr\left(\left|n^{-1}\sum_{i=1}^{n} X_{ij}\mathbf{1}\{|X_{ij}| > M\}\right| > \frac{\lambda}{4}\right) \\ &\leq p_n n \max_{1\leq i\leq n, j=q_n+1,\ldots,p_n} \Pr\left(|X_{ij}| > M\right) = O(p_n n)\exp(-tn^{\theta}) \to 0, \end{aligned}$$

as $n \to \infty$, because $\log p_n = o(n^{\min\{c_2-2\theta,\theta\}})$. Therefore, $J_{n1} = J_{n11} + J_{n12} = o(1)$.

Following arguments for proving Lemma 4.3 of Wang, Wu, and Li (2012) and the arguments that deal with T_{n22} and J_{n12}, we can show that $J_{n2} = o(1)$. It remains to deal with J_{n3}. For a fixed $M = O(n^{\theta})$ with $0 < \theta < (c_2 - c_1)/2$,

$$\begin{aligned} J_{n3} \leq \Pr\Bigg(&\max_{j=q_n+1,\ldots,p_n} \sup_{\|\boldsymbol{\beta}_1-\boldsymbol{\beta}^o_{\tau 1}\|\leq\Delta\sqrt{q_n/n}} \bigg|n^{-1}\sum_{i=1}^{n} X_{ij}\mathbf{1}\{|X_{ij}| \leq M\} \\ &\Big\{\Pr(Y_i - \mathbf{x}_{i,\mathcal{A}}^{\mathrm{T}}\boldsymbol{\beta}_1 \leq 0) - \Pr(Y_i - \mathbf{x}_{i,\mathcal{A}}^{\mathrm{T}}\boldsymbol{\beta}^o_{\tau 1} \leq 0)\Big\}\bigg| > \frac{\lambda}{8}\Bigg) \\ &+ \Pr\Bigg(\max_{j=q_n+1,\ldots,p_n} \sup_{\|\boldsymbol{\beta}_1-\boldsymbol{\beta}^o_{\tau 1}\|\leq\Delta\sqrt{q_n/n}} \bigg|n^{-1}\sum_{i=1}^{n} X_{ij}\mathbf{1}\{|X_{ij}| > M\} \\ &\Big\{\Pr(Y_i - \mathbf{x}_{i,\mathcal{A}}^{\mathrm{T}}\boldsymbol{\beta}_1 \leq 0) - \Pr(Y_i - \mathbf{x}_{i,\mathcal{A}}^{\mathrm{T}}\boldsymbol{\beta}^o_{\tau 1} \leq 0)\Big\}\bigg| > \frac{\lambda}{8}\Bigg) \\ =:&\ J_{n31} + J_{n32}. \end{aligned}$$

To handle J_{n31}, we observe that

$$\begin{aligned} &\max_{j=q_n+1,\ldots,p_n} \sup_{\|\boldsymbol{\beta}_1-\boldsymbol{\beta}^o_{\tau 1}\|\leq\Delta\sqrt{q_n/n}} \bigg|n^{-1}\sum_{i=1}^{n} X_{ij}\mathbf{1}\{|X_{ij}| > M\} \\ &\qquad \Big\{\Pr(Y_i - \mathbf{x}_{i,\mathcal{A}}^{\mathrm{T}}\boldsymbol{\beta}_1 \leq 0) - \Pr(Y_i - \mathbf{x}_{i,\mathcal{A}}^{\mathrm{T}}\boldsymbol{\beta}^o_{\tau 1} \leq 0)\Big\}\bigg| \\ &\leq M \sup_{\|\boldsymbol{\beta}_1-\boldsymbol{\beta}^o_{\tau 1}\|\leq\Delta\sqrt{q_n/n}} \Big|E\Big\{f(\zeta|\mathbf{x}_{\mathcal{A}})\mathbf{x}_{\mathcal{A}}^{\mathrm{T}}(\boldsymbol{\beta}_1 - \boldsymbol{\beta}^o_{\tau 1})\Big\}\Big| \\ &\leq M \sup_{\|\boldsymbol{\beta}_1-\boldsymbol{\beta}^o_{\tau 1}\|\leq\Delta\sqrt{q_n/n}} \lambda_{\max}^{1/2}\left\{E(\mathbf{x}_{\mathcal{A}}\mathbf{x}_{\mathcal{A}}^{\mathrm{T}})\right\}\|\boldsymbol{\beta}_1 - \boldsymbol{\beta}^o_{\tau 1}\| \\ &\leq O\big\{n^{\theta}\,(q_n/n)^{1/2}\big\} = O\{n^{-(1-c_1-2\theta)/2}\}, \end{aligned}$$

where $f(\cdot \mid \mathbf{x}_{\mathcal{A}})$ is defined in (C3) where ζ is between $u^o_{\tau} + \mathbf{x}_{\mathcal{A}}^{\mathrm{T}}(\boldsymbol{\beta}_1 - \boldsymbol{\beta}^o_{\tau 1})$ and u^o_{τ}, and thus the second inequality follows (C3) and Cauchy-Schwartz inequality, and the third inequality follows (C2). Consequently, together with $\lambda = o\{n^{-(1-c_2)/2}\}$, we have that $J_{n31} \leq \Pr\{O(n^{-(1-c_1-2\theta)/2}) > \lambda/8\} = o(1)$ if $0 < \theta < (c_2 - c_1)/2$. We can also follow similar arguments as with J_{n12} and obtain that $J_{n32} = o(1)$. Therefore, $J_{n3} = J_{n31} + J_{n32} = o(1)$. Consequently,

$$\begin{aligned}&\Pr\left\{\max_{q_n+1,\ldots,p_n}\left|n^{-1}\sum_{i=1}^{n} X_{ij}\{I(Y_i - \mathbf{x}_{i,\mathcal{A}}^{\mathrm{T}}\widehat{\boldsymbol{\beta}}_{\tau 1}^{o} < 0) - \tau\}\right| > \lambda\right\}\\ &\quad \leq J_{n1} + J_{n2} + J_{n3} = o(1),\end{aligned}$$

which implies that $\Pr\left\{|s_j(\widehat{\boldsymbol{\beta}}_{\tau}^{o})| > \lambda, \quad \text{for some } j = q_n + 1, \ldots, p_n\right\} \to 0$. This completes the proof of Lemma B.3.

With Lemmas B.2 and B.3 for random $\mathbf{x}$ with sub-exponential tail probability, (C1), we can follow the proof of Theorem 2.4 of Wang, Wu, and Li (2012) to obtain the oracle property and complete the proof.

Appendix C. Proof of Theorem 2.

We use c_1 and c_2 to denote two different generic positive constants. First we assume $F(y)$ is known. Then, $\widehat{\mathrm{dcov}}^{*2}\{X_k, F(Y)\} = \widehat{S}_{k1}^{*} + \widehat{S}_{k2}^{*} - 2\widehat{S}_{k3}^{*}$, where

$$\begin{aligned}\widehat{S}_{k,1}^{*} &= \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}|X_{ik} - X_{jk}|\,|F(Y_i) - F(Y_j)|\,,\\ \widehat{S}_{k,2}^{*} &= \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}|X_{ik} - X_{jk}|\,\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}|F(Y_i) - F(Y_j)|\,, \text{ and}\\ \widehat{S}_{k,3}^{*} &= \frac{1}{n^3}\sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{l=1}^{n}|X_{ik} - X_{lk}|\,|F(Y_j) - F(Y_l)|\,.\end{aligned}$$

Similarly we take $\widehat{\omega}_k^{*} = \widehat{\mathrm{dcorr}}^{*2}\{X_k, F(Y)\}$. Theorem 1 of Li, Zhong, and Zhu (2012) stated that, for any $0 < \gamma < 1/2 - \kappa$, there exist positive constants $c_1 > 0$ and $c_2 > 0$ such that

$$\begin{aligned}&\Pr\left(\max_{1\leq k\leq p}|\widehat{\omega}_k^{*} - \omega_k| \geq cn^{-\kappa}\right)\\ &\quad \leq O\left(p\left[\exp\left\{-c_1 n^{1-2(\kappa+\gamma)}\right\} + n\exp\left(-c_2 n^{\gamma}\right)\right]\right). \qquad \text{(C.1)}\end{aligned}$$

To prove Theorem 2, it thus suffices to show the difference between $\widehat{\omega}_k^{*}$ and $\widehat{\omega}_k$ at (3.5) is ignorable when n is large enough, which amounts to studying the differences between $\widehat{S}_{km}^{*}$ and $\widehat{S}_{km}$ for $m = 1, 2, 3$. We sketch the proof for the case $m = 1$ only because the proof of the other two cases is, in spirit, the same. We recall that $\widehat{S}^*_{k1} = (1/n^2)\sum_{i=1}^n\sum_{j=1}^n |X_{ik}-X_{jk}||F(Y_i)-F(Y_j)|$ and $\widehat{S}_{k1} = (1/n^2)\sum_{i=1}^n\sum_{j=1}^n |X_{ik}-X_{jk}||F_n(Y_i)-F_n(Y_j)|$. Then,

$$\begin{aligned}
&\Pr\left(\max_{1\le k\le p_n} |\widehat{S}^*_{k1}-\widehat{S}_{k1}| \ge \varepsilon\right)\\
&= \Pr\Big(\max_{1\le k\le p_n} n^{-2}\sum_{i=1}^n\sum_{j=1}^n |X_{ik}-X_{jk}|\Big||F(Y_i)-F(Y_j)|-|F_n(Y_i)-F_n(Y_j)|\Big| \ge \varepsilon\Big)\\
&\le \Pr\left(\max_{1\le k\le p_n} (A_nB_n)^{1/2} \ge \varepsilon\right)\\
&\le \Pr\left(\max_{1\le k\le p_n} (A_nB_n)^{1/2} \ge \varepsilon, |X_k|\le M\right) + \Pr\left(\max_{1\le k\le p_n} (A_nB_n)^{1/2} \ge \varepsilon, |X_k| > M\right)\\
&=: T_1 + T_2,
\end{aligned}$$

where M is a positive constant to be specified later, $A_n = n^{-2}\sum_{i=1}^n\sum_{j=1}^n (X_{ik}-X_{jk})^2$, and $B_n = n^{-2}\sum_{i=1}^n\sum_{j=1}^n \{|F(Y_i)-F(Y_j)|-|F_n(Y_i)-F_n(Y_j)|\}^2$.

Using $\big||x|-|y|\big| \le |x-y| \le |x|+|y|$, we obtain that

$$\begin{aligned}
\big|\,|F_n(Y_i)-F_n(Y_j)|-|F(Y_i)-F(Y_j)|\,\big| &\le |F_n(Y_i)-F(Y_i)| + |F_n(Y_j)-F(Y_j)|\\
&\le 2\max_{1\le i\le n} |F_n(Y_i)-F(Y_i)|\,.
\end{aligned}$$

Also because $\max\limits_{1\le k\le p_n} A_n \le \max\limits_{1\le k\le p_n} n^{-2}\sum_{i=1}^n\sum_{j=1}^n 2(X_{ik}^2+X_{jk}^2) \le 4M^2$, we have

$$\begin{aligned}
T_1 &\le \Pr\left[\max_{1\le k\le p_n} 2Mn^{-1}\Big\{\sum_{i=1}^n\sum_{j=1}^n (|F(Y_i)-F(Y_j)|-|F_n(Y_i)-F_n(Y_j)|)^2\Big\}^{1/2} \ge \varepsilon\right]\\
&\le \Pr\left\{4M\max_{1\le i\le n}|F_n(Y_i)-F(Y_i)| \ge \varepsilon\right\} \le \Pr\left\{\max_{y\in R}|F_n(y)-F(y)| \ge \frac{\varepsilon}{4M}\right\}\\
&\le 2\exp\left\{-2n(\frac{\varepsilon}{4M})^2\right\} = 2\exp(-\frac{n\varepsilon^2}{8M^2}),
\end{aligned} \tag{C.2}$$

where the last inequality follows by the Dvoretzky-Kiefer-Wolfowitz inequality.

For the second term, for all $0 < s \le 2s_0$, where s_0 is defined in (C.1),

$$\begin{aligned}
T_2 &\le \Pr\left(\max_{1\le k\le p_n}|X_k| > M\right) = \Pr\left\{\max_{1\le k\le p_n}\exp(s|X_k|) > \exp(sM)\right\}\\
&\le \max_{1\le k\le p_n} E\{\exp(s|X_k|)\}\exp(-sM) \le C\exp(-sM),
\end{aligned} \tag{C.3}$$

where C is a positive constant, the second inequality follows from Markov's inequality, and the last inequality is applied under (C.1).

Then, by choosing $M = O(n^\gamma)$ for $0 < \gamma < 1/2-\kappa$, (C.2) and (C.3) together imply that, for some positive constants c_1 and c_2,

$$\begin{aligned}\Pr\left(\max_{1\le k\le p_n}|\widehat{S}^*_{k1}-\widehat{S}_{k1}|\ge\varepsilon\right)&\le 2\exp(-\frac{n\varepsilon^2}{8M^2})+C\exp(-sM)\\&\le 2\exp(-c_1\varepsilon^2n^{1-2\gamma})+C\exp(-c_2n^{\gamma}).\quad\text{(C.4)}\end{aligned}$$

Thus, it is not difficult to show that

$$\Pr\left(\max_{1\le k\le p}|\widehat{\omega}^*_k-\widehat{\omega}_k|\ge cn^{-\kappa}\right)\le O\left(\exp\left\{-c_1n^{1-2(\kappa+\gamma)}\right\}+\exp\left(-c_2n^{\gamma}\right)\right).\quad\text{(C.5)}$$

Then (C.1) and (C.5) together complete the proof of Theorem 2.

References

Altham, P. M. E. (1984). Improving the precision of estimation by fitting a generalized linear model and quasi-likelihood. *J. Roy. Statist. Soc. Ser. B* **46**, 118-119.

Bickel, P. and Levina E. (2008). Regularized estimation of large covariance matrices. *Ann. Statist.* **36**, 199-227.

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

Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and it oracle properties. *J. Amer. Statist. Assoc.* **96**, 1348-1360.

Fan, J. and Lv, J. (2008). Sure independence screening for ultrahigh dimensional feature space (with discussion). *J. Roy. Statist. Soc. Ser. B* **70**, 849-911.

Fan, Y. and Lv, J. (2013). Asymptotic equivalence of regularization methods in thresholded parameter space. *J. Amer. Statist. Assoc.* **108**, 1044-1061.

Fan, J., Samworth, R. and Wu, Y. (2009). Ultrahigh dimensional feature selection: beyond the linear model. *J. Machine Learn. Res.* **10**, 1829-1853.

Hall, P. and Li, K. C. (1993). On almost linearity of low dimensional projection from high dimensional data. *Ann. Statist.* **21**, 867-889.

Hall, P. and Miller, H. (2009). Using generalized correlation to effect variable selection in very high dimensional problems. *J. Comput. Graph. Statist.* **18**, 533-550.

Härdle, W., Hall, P. and Ichimura, H. (1993). Optimal smoothing in single index models. *Ann. Statist.* **21**, 157-178.

He, X. and Shao, Q. (2000). On parameters of increasing dimensions. *J. Multivariate Anal.* **73**, 120-135.

Knight, K. (1998). Limiting distributions for L_1 regression estimators under general conditions. *Ann. Statist.* **26**, 755-770.

Koenker, R. (2005). *Quantile Regression* (*Econometric Society Monographs*). Cambridge University Press. ISBN: 0521608279.

Kong, E. and Xia, Y. (2007). Variable selection for the single index model. *Biometrika* **94**, 217-229.

Li, K.-C. (1991). Sliced inverse regression for dimension reduction. *J. Amer. Statist. Assoc.* **86**, 316-327.

Li, G., Peng. H., Zhang J. and Zhu, L. (2012). Robust rank correlation based screening. *Ann. Statist.* **40**, 1846-1877.

Li, R., Zhong, W. and Zhu, L. (2012). Feature screening via distance correlation learning. *J. Amer. Statist. Assoc.* **107**, 1129-1139.

Liang, H., Liu, X., Li, R. and Tsai, C. (2010). Estimation and testing for partially linear single index models. *Ann. Statist.* **38**, 3811-3836.

Naik, P. A. and Tsai, C.-L. (2001). Single-index model selections. *Biometrika* **88**, 821-832.

Powell, J. L., Stock, J. H. and Stoker, T. M. (1989). Semiparametric estimation of index coefficient. *Econometrica* **51**, 1403-1430.

Segal, M. R., Dahlquist, K. D. and Conklin, B. R. (2003). Regression approach for microarray data analysis. *Journal of Computational Biology* **10**, 961-980.

Székely, G. J., Rizzo, M. L. and Bakirov, N. K. (2007). Measuring and testing dependence by correlation of distances. *Ann. Statist.* **35**, 2769-2794.

Tao, P. D. and An, L. T. H. (1997). Convex analysis approach to D.C. programming: theory, algorithms and applications. *Acta Math. Vietnamica* **22**, 289-355.

Tibshirani, R. (1996). Regression shrinkage and selection via LASSO. *J. Roy. Statist. Soc. Ser. B* **58**, 267-288.

Wang, L., Wu, Y. and Li, R. (2012). Quantile regression for analyzing heterogeneity in ultra-high dimension. *J. Amer. Statist. Assoc.* **107**, 214-222.

Wu, Y. and Liu, Y. (2009). Variable selection in quantile regression. *Statist. Sinica* **19**, 801-817.

Zhang, C.(2010). Nearly unbiased variable selection under minimax concave penalty. *Ann. Statist.* **101**, 1418-1429.

Zheng, Z., Fan, Y. and Lv, J. (2014), High dimensional thresholded regression and shrinkage effect. *J. Roy. Statist. Soc. Ser. B* **76**, 627-649.

Zhu, L., Huang, M. and Li, R. (2012). Semiparametric quantile regression with high-dimensional covariates. *Statist. Sinica* **22**, 1379-1401.

Zhu, L., Qian, L. and Lin, J. (2011). Variable selection in a class of single index models. *Ann. Statist. Math.* **63**, 1277-1293.

Zhu, L. P., Li, L., Li, R. and Zhu, L. X. (2011). Model-free feature screening for ultrahigh dimensional data. *J. Amer. Statist. Assoc.* **106**, 1464-1475.

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

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

Wang Yanan Institute for Studies in Economics, Department of Statistics, School of Economics, and Fujian Key Laboratory of Statistical Science, Xiamen University, Xiamen 361005, China.

E-mail: wzhong@xmu.edu.cn

Institute of Statistics and Big Data, Renmin University of China, 59 Zhongguancun Avenue, Haidian District, Beijing 100872, P. R. China.

E-mail: zhuliping.stat@yahoo.com

Department of Statistics and The Methodology Center, The Pennsylvania State University, University Park, PA 16802, USA.

E-mail: rzli@psu.edu

School of Mathematical Science, Capital Normal University, Beijing 100048, China.

E-mail: hjcui@bnu.edu.cn

(Received February 2014; accepted January 2015)