Published in final edited form as:

Stat Sin. 2016 January ; 26(1): 69–95. doi:10.5705/ss.2014.049.

Regularized Quantile Regression and Robust Feature Screening for Single Index Models

Wei Zhong,

Wang Yanan Institute for Studies in Economics, Department of Statistics and Fujian Key Laboratory of Statistical Science, Xiamen University, Xiamen 361005, China.
wzhong@xmu.edu.cn

Liping Zhu,

School of Statistics and Management and Key Laboratory of Mathematical Economics, Ministry of Education, Shanghai University of Finance and Economics, Shanghai 200433, China.
zhu.liping@mail.shufe.edu.cn

Runze Li, and

Department of Statistics and The Methodology Center, The Pennsylvania State University, University Park, PA 16802. rzli@psu.edu

Hengjian Cui

School of Mathematical Science, Capital Normal University, Beijing 100048, China.
hjcui@bnu.edu.cn

Abstract

We propose both 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 the 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 the penalized quantile regression since it helps to reduce the covariate dimension from ultrahigh dimensionality to a moderate scale. Based on the reduced model, the penalized linear quantile regression further refines selection of

Supplementary Materials

The supplementary material of this paper consists of Appendix S.A and S.B. In Appendix S.A, 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 Appendix S.B, We provide extra simulation on Example 1 with more active covariates.

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 real data set.

Keywords

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. Traditional single index regression model is referred to as

$$Y = m(\mathbf{x}^T \beta_0) + \varepsilon, \quad (1.1)$$

where $m(\cdot)$ is an unknown regression function, β_0 consists of unknown index parameters, and ε is a random error with $E(\varepsilon | \mathbf{x}) = 0$ and $\text{var}(\varepsilon | \mathbf{x}) = \sigma^2$. Model (1.1) has been well studied in the literature. See, for example, Powell, Stock and Stoker (1989) and Härdle, Hall and Ichimura (1993). Zhu, Huang and Li (2012) studied the following heteroscedastic single index regression model

$$Y = m(\mathbf{x}^T \beta_0) + \sigma(\mathbf{x}^T \beta_0) \varepsilon, \quad (1.2)$$

for unknown functions $m(\cdot)$ and $\sigma(\cdot)$, where ε has mean zero and is assumed to be independent of $\mathbf{x}$. Zhu, Huang and Li (2012) developed an estimation procedure for β_0 and $m(\cdot)$ under a quantile loss function when the dimension of $\mathbf{x}$ is finite.

The goal of regression analysis amounts to characterizing how the conditional distribution of the response variable Y varies with the realizations of the covariate vector $\mathbf{x} = (X_1, \dots, X_{p_n})^T$. In this paper, we focus on ultrahigh dimensional situation. Thus, we denote by p_n the dimension of $\mathbf{x}$ to emphasize the dependence of p_n on the sample size n . Denote by $F(y | \mathbf{x})$ the conditional distribution of Y given $\mathbf{x}$. In this paper, we study a general class of single index models that include models (1.1) and (1.2) as special cases. Specifically, we assume that there exists $\beta_0 \in \mathbb{R}^{p_n}$ such that

$$F(y | \mathbf{x}) = F(y | \mathbf{x}^T \beta_0), \quad \text{for all } y \in \mathbb{R}. \quad (1.3)$$

That is, the conditional distribution of $(Y | \mathbf{x})$ is fully characterized through a single linear combination of predictors $\mathbf{x}^T \beta_0$. Consequently, the “curse of dimensionality” issue is avoided and simultaneously the model interpretability is maintained via a single index structure. Because the conditional distributional function $F(\cdot | \cdot)$ is unknown, the index parameter β_0 is not identifiable. The direction of β_0 , instead of its true value, is of our

primary interest. We refer to model (1.3) as a conditional distribution based single index model (CDSIM for short) in order to distinguish it from models (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 those 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 the nonconvex penalized quantile regression for analyzing heterogeneity in the ultrahigh dimensional setting. Fan, Fan and Barut (2014) proposed two-step adaptive robust LASSO based on weighted L_1 -penalized quantile regression to deal with heavy-tailed high dimensional data.

In this paper, we consider variable selection and feature screening for model (1.3) when the covariate dimension p_n is ultrahigh. We further assume β_0 is sparse. Denote by A the active index set, β_A the nonzero entries of β_0 and $\mathbf{x}_A$ the collection of all active covariates. When β_0 is sparse, model (1.3) reduces to

$$F(y|\mathbf{x}) = F\left(y|\mathbf{x}_A^T \beta_A\right), \quad \text{for all } y \in \mathbb{R}. \quad (1.4)$$

Our goal is to identify A and if possible, to estimate β_A . To the best of our knowledge, there is few variable selection method designed for model (1.3) or (1.4) with ultrahigh-dimensional covariates.

In this paper we introduce two approaches to accomplish our goal: a penalized linear quantile regression and an independence screening procedure. When model (1.3) is true, the quantile functions of $(Y|\mathbf{x})$ always vary with the realizations of $(\mathbf{x}^T \beta_0)$. In other words, the quantile function admits a single index structure. This motivates us to implement a penalized quantile regression to exclude irrelevant covariates and simultaneously estimate the direction of β_0 . The advantage of using quantile regression is that the quantile function characterizes equivalently the distributional function (1.3) and it is resilient to outliers and extreme values in the response. We show that, although the true quantile functions of $(Y|\mathbf{x})$ are possibly nonlinear, the resulting 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. Due to its computational efficiency, it 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 of n .

From a practical perspective, when the covariate dimension is extremely large, even the penalized linear quantile regression may suffer from at least two serious drawbacks: 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 function of the response in model (1.3) and the implicit model (1.4). Since the distribution function is bounded and monotone, we can reasonably expect that this new independence screening procedure still works in the presence of outliers or extreme values in the response variable. In addition, it is computationally efficient and hence offers a useful complement, rather an alternative, 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 tolerant 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.

This paper is organized as follows. In Section 2, we propose the 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. All technical proofs are given in the Appendix.

2. Penalized Linear Quantile Regression

In this section, we will construct an estimate for the direction of β_0 in model (1.3) via the penalized linear quantile regression.

2.1 The Methodology

Model (1.3) and its sparse structure (1.4) indicate that the quantile functions of $(Y | \mathbf{x})$ at different quantile levels are all functions of $(\mathbf{x}^T \beta_0)$ and $(\mathbf{x}_A^T \beta_A)$ if the sparsity principle applies. This motivates us to estimate β_0 through the 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 β_0 in model (1.3). To be specific, we denote by $\rho_\tau(r) = \tau r - rI(r < 0)$, the check loss function at the τ th quantile, for $\tau \in (0, 1)$. Define

$$\mathcal{L}_\tau(u, \mathbf{b}) = E \left\{ \rho_\tau(Y - u - \mathbf{x}^T \mathbf{b}) \right\} \quad \text{and} \quad (u_\tau, \beta_\tau) = \underset{u, \mathbf{b}}{\operatorname{argmin}} \{ \mathcal{L}_\tau(u, \mathbf{b}) \}, \quad (2.1)$$

where $\mathbf{b} = (b_1, \dots, b_{p_n})^T \in \mathbb{R}^{p_n}$.

Lemma 1—If $\mathbf{x}$ satisfies the linearity condition that

$$E \left\{ \mathbf{x} - E(\mathbf{x}) | \mathbf{x}^T \beta_0 \right\} = \operatorname{var}(\mathbf{x}) \beta_0 \left\{ \beta_0^T \operatorname{var}(\mathbf{x}) \beta_0 \right\}^{-1} \beta_0^T \{ \mathbf{x} - E(\mathbf{x}) \},$$

then β_τ is proportional to β_0 in model (1.3).

The 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 the reality as long as p_n is sufficient large. Therefore, the linearity condition is typically regarded as mild in an ultrahigh dimensional setting. Lemma 1 implies that the indices of zero entries in both β_0 and β_τ coincide. To estimate the direction of β_0 in model (1.3), it amounts to estimating β_τ defined in (2.1). This lemma can be proved by using similar arguments used in Zhu, Huang and Li (2012). Thus, we omit its proof to save space.

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

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

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

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

where $a > 1$. The SCAD penalty is defined as follows,

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

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

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

2.2 The Oracle Property

In this section we study the oracle property of the estimators obtained from the penalized linear quantile regression. Without loss of generality, we assume the first q_n components of $\mathbf{x}$ are active and the rest are inactive, where $q_n (\ll p_n)$ is a small positive integer representing

the sparsity level. In other words, $A = \{1, 2, \dots, q_n\}$. We define the oracle estimator at the population level by

$$\mathcal{L}_\tau(u, \mathbf{b}_1) = E \left\{ \rho_\tau \left(Y - u - \mathbf{x}_A^T \mathbf{b}_1 \right) \right\} \quad \text{and} \quad (u_\tau^o, \beta_{\tau 1}^o) = \underset{u, \mathbf{b}_1}{\operatorname{argmin}} \{ \mathcal{L}_\tau(u, \mathbf{b}_1) \}, \quad (2.4)$$

where $\mathbf{b}_1 = (b_1, \dots, b_{q_n})^T \in \mathbb{R}^{q_n}$. We further write $\beta_\tau^0 = (\beta_{\tau 1}^{oT}, \mathbf{0}^T)^T$, where $\beta_{\tau 1}^o$ 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 $\hat{\beta}_\tau^0 = (\hat{\beta}_{\tau 1}^{oT}, \mathbf{0}^T)^T$ at the sample level by

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

We assume the following regularity conditions to investigate the consistency of the oracle estimator defined in (2.5) and the oracle property of the resulting estimator obtained from the penalized linear quantile regression (2.3).

(C1) The covariates $\mathbf{x}$ satisfy the sub-exponential tail probability uniformly in p_n . That is, there exist positive constants t_0 and C such that

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

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

$$C_1 \leq \lambda_{\min} \left\{ E \left(\mathbf{x}_A \mathbf{x}_A^T \right) \right\} \leq \lambda_{\max} \left\{ E \left(\mathbf{x}_A \mathbf{x}_A^T \right) \right\} \leq C_2,$$

where $\lambda_{\min}$ and $\lambda_{\max}$ represent the smallest and largest eigenvalues, respectively. Assume further that $\{(\mathbf{x}_{i,A}, Y_i), i = 1, \dots, n\}$ are in general positions (Koenker, 2005, Section 2.2).

(C3) The probability density function of $Y - \mathbf{x}^T \beta_\tau$ conditional on $\mathbf{x}$, denoted by $f(\cdot | \mathbf{x})$, is uniformly bounded away from 0 and ∞ in the neighborhood around u_τ^o .

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

(C5) For $\beta_{\tau 1}^o = (\beta_{\tau 1}^o, \beta_{\tau 2}^o, \dots, \beta_{\tau, q_n}^o)^T$, there exist positive constants c_2 and C such that $2c_1 < c_2 < 1$ and

$$\min_{1 \leq j \leq q_n} |\beta_{\tau, j}^o| \geq C n^{-(1-c_2)/2}.$$

Condition (C1) is concerned with the moments of the covariates, which follows immediately 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 $(Y - \mathbf{x}^T \boldsymbol{\beta}_D)$ 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 below states the consistency of the oracle estimators $\hat{u}_\tau^o$ and $\hat{\beta}_{\tau 1}^o$.

Lemma 2—Under Conditions (C1)-(C4), the oracle estimators $\hat{u}_\tau^o$ and $\hat{\beta}_{\tau 1}^o$ satisfy

$$\|\hat{\beta}_{\tau 1}^o - \beta_{\tau 1}^o\| = O_p\left(\sqrt{q_n/n}\right) \quad \text{and} \quad \|\hat{u}_\tau^o - u_\tau^o\| = O_p\left(\sqrt{q_n/n}\right). \quad (2.7)$$

Next we study the theoretical property of the oracle estimator $\hat{\beta}_\tau^0 = (\hat{\beta}_{\tau 1}^{oT}, \mathbf{0}^T)^T$.

Theorem 1—(THE ORACLE PROPERTY) Suppose Conditions (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 $B_n(\lambda)$ be the set of local minima $\hat{\beta}_\tau$ of the objective function $Q(u, \mathbf{b})$ defined in (2.2) with the SCAD or the MCP penalty and the tuning parameter λ . The oracle estimator $\hat{\beta}_\tau^0 = (\hat{\beta}_{\tau 1}^{oT}, \mathbf{0}^T)^T$ satisfies

$$Pr\left\{\hat{\beta}_\tau^0 \in B_n(\lambda)\right\} \rightarrow 1, \quad \text{as } n \rightarrow \infty.$$

Theorem 1 implies that the oracle estimator $\hat{\beta}_\tau^0$ is a local minimizer of the objective function (2.2) with the probability approaching one as $n \rightarrow \infty$. 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, 2 and Theorem 1 imply that $\hat{\beta}_\tau$ from the penalized linear quantile regression is a consistent estimator of the direction of $\boldsymbol{\beta}_0$ in 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 are bounded uniformly while we relax their assumption to condition (C1) which only requires the 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

Next we propose a new robust feature screening procedure for model (1.3) through using distance correlation.

3.1 The Methodology

Theorem 1 indicates that the oracle property of the penalized quantile regression holds asymptotically true for $\log p_n = o(n^\delta)$ for some $\delta > 0$. This suffices for many problems from

a theoretical perspective. From a practical perspective, however, when the covariate dimension is extremely large, even the penalized linear quantile regression suffers from at least two serious drawbacks: computational expediency and algorithmic stability (Fan, Samworth and Wu, 2009). When the penalized linear quantile regression fails to produce a reliable solution within a tolerant time, an independence screening procedure better precede the penalized linear quantile regression. This strategy helps to reduce the ultrahigh dimensionality down to a relatively moderate scale. To this end, we propose an independence screening procedure which excludes irrelevant covariates and hence reduces the computational complexity of subsequent penalized quantile regressions at all different quantile levels. In other words, the independence screening procedure is expected to have a sure screening property and to be independent of the quantile levels. In addition, it is expected to behave well when extreme values and/or outliers are present in the observed response values because subsequent quantile regression automatically has such a robustness property.

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

$$dcov^2(X, Y) = S_1 + S_2 - 2S_3, \quad (3.1)$$

$$\begin{aligned} S_1 &= E(|X - \tilde{X}| |Y - \tilde{Y}|), S_2 \\ &= E(|X - \tilde{X}|) E(|Y - \tilde{Y}|), S_3 \\ \text{where } &= E\{E(|X - \tilde{X}| |X) E(|Y - \tilde{Y}| |Y)\}, \text{ and } (\tilde{X}, \tilde{Y}) \text{ is an independent} \\ &\text{copy of } (X, Y). \text{ Then, the distance covariance between } X \text{ and } Y \text{ is defined by} \end{aligned}$$

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

Szekely, Rizzo and Bakirov (2007) pointed out that $dcorr(X, Y) = 0$ if and only if X and Y are independent and $dcorr(X, Y)$ 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 an sure independence screening to rank all predictors using their distance correlations with the response variable, called as DC-SIS, and proved its sure screening property for ultrahigh dimensional data.

Next, we denote by X_k the k th predictor with $k = 1, \dots, p_n$ 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,

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

where $F(y) = E\{\mathbf{1}(Y \leq y)\}$ 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 defined in (3.3) has several distinctive and appealing advantages comparing with the existing measurements.

1. It is obvious that $\text{dcorr}\{X_k, F(Y)\} = 0$ if and only if X_k and Y are independent. Following similar arguments in Li, Zhong and Zhu (2012), we can see that this new independence screening procedure based on (3.3) is model-free and hence is applicable to model (1.3) and its sparse model structure (1.4).
2. Since $F(Y)$ is a bounded function for all types of Y , we can naturally expect that the independence screening 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.
3. If one suspects that the covariates also contain some extreme values, then one can use the utility $\text{dcorr}\{F_k(X_k), F(Y)\}$ to rank the importance of X_k , where $F_k(x) = E\{\mathbf{1}(X_k \leq x)\}$.

In the sequel we introduce how to implement the marginal utility (3.3) in the screening procedure. Let $\{(\mathbf{x}_i, Y_i), i = 1, \dots, n\}$ 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,

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

where

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

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

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

Our proposed independence screening procedure retains the covariates with the $\hat{\omega}_k$ values larger than a user-specified threshold. Denote

$$\hat{A} = \{k: \hat{\omega}_k \geq cn^{-\kappa}, \quad \text{for } 1 \leq k \leq p_n\}$$

for some pre-specified thresholds $c > 0$ and $0 < \kappa < 1/2$. The constants c and κ control the signal strength and will be defined in Condition (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 $\hat{\omega}_k$ defined in (3.5), which paves the road for proving the sure screening property of the DC-RoSIS procedure.

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

$$Pr \left(\max_{1 \leq k \leq p} |\hat{\omega}_k - \omega_k| \geq cn^{-\kappa} \right) \leq O \left(p \left[\exp \left\{ -c_1 n^{1-2(\kappa+\gamma)} \right\} + n \exp(-c_2 n^\gamma) \right] \right) \quad (3.6)$$

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

(C6) The marginal utility defined in (3.3) satisfies $\min_{k \in A} \omega_k \geq 2cn^{-\kappa}$, for some constants $c > 0$ and $0 < \kappa < 1/2$.

Condition (C6) allows the minimal signal of the active covariates converges to zero as the sample size diverges, yet it requires the minimum signal of active covariates be not too small. The sure screening property is stated below.

Theorem 3—(SURE SCREENING PROPERTY) Under condition (C6) and the conditions in Theorem 2, it follows that

$$Pr \left(A \subseteq \hat{A} \right) \geq 1 - O \left(s_n \left[\exp \left\{ -c_1 n^{1-2(\kappa+\gamma)} \right\} + n \exp(-c_2 n^\gamma) \right] \right), \quad (3.7)$$

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

4. Numerical Studies

In this section, we first conduct 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 generate $\mathbf{x} =$

$(X_1, X_2, \dots, X_p)^T$ from $N(\mathbf{0}, \Sigma)$, where $\Sigma = (\sigma_{ij})_{p \times p}$ with $\sigma_{ij} = 0.5^{|i-j|}$. The dimensionality $p = 1,000$ and the sample size $n = 200$.

Example 1—This example is designed to compare the finite sample performance of our proposal DC-RoSIS with existing procedures including SIS (Fan and Lv, 2008), SIRS (Zhu, Li, Li and Zhu, 2011), RRCS (Li, Peng, Zhang and Zhu, 2012) and DC-SIS (Li, Zhong and

Zhu, 2012). We repeat each experiment 500 times and evaluate the performance with the following three criteria.

1. S : The minimum model size to include all active covariates. We summarize the median of S with its associated robust estimate of the standard deviation ($RSD = IQR/1.34$). A smaller S value indicates a better performance.
2. P_{sj} : The empirical probability that the active covariate X_j is selected for a given model size d . We set $d = 2\lceil n/\log n \rceil$ throughout.
3. P_a : The empirical probability that all active covariates are selected for the given model size $d = 2\lceil n/\log n \rceil$. If the sure screening property holds true, both P_{sj} and P_a values are close to one when the estimated model size d is reasonably large.

We consider the following four models:

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

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

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

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

where $\boldsymbol{\beta} = (3, 1.5, 0, 0, 0, 0, 2, 0, \dots, 0)^T$. In the above four models, only X_1 , X_2 and X_7 are truly important. The random error ε is independently generated from either standard normal or standard Cauchy distribution. In model (1), $H(Y) = \{|Y|^\lambda \text{sgn}(Y) - 1\}/\lambda$ is the Box-Cox transformation. This model was used in Li, Peng, Zhang and Zhu (2012). We set $\lambda = 1$ and $\lambda = 0.25$. Both models (2) and (3) are heteroscedastic single-index models. The single index $(\mathbf{x}^T \boldsymbol{\beta})$ contributes both the conditional mean and variance of the response in model (2), and are totally irrelevant to the mean regression function in model (3). The active covariate X_7 in model (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 independence screening procedure works quite well for a variety of regressions even when the model assumptions are violated.

The results are summarized in Table 4.1. It can be seen that SIS does not perform well when ε follows Cauchy distribution. Even when ε follows standard normal distribution, SIS still fails to behave well in nonlinear models (3) and (4). SIRS performs very well for all single-index models. However, SIRS fails to identify X_7 as an important covariate in model (4) because it is not capable of detecting symmetric patterns. The performance of RRCS is generally favorable for models (1) and (2). However, RRCS hardly detects the active covariates that are only relevant to the conditional variance of the response in model (3) or X_7 that exhibits symmetric patterns with Y in model (4). DC-RoSIS and DC-SIS have similar performances when ε follows standard normal distribution. When ε follows Cauchy distribution, DC-RoSIS significantly improves DC-SIS. For example, in model (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 examine 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 first utilize our proposed screening procedure to select $d = 2\lceil n/\log(n) \rceil$ top ranked covariates and then apply the penalized linear quantile regression to estimate the direction of β . For the conditional quantile regression, we consider three different quantiles $\tau = 0.25, 0.50$ and 0.75 , which correspond to the 1st quartile, the median and 3rd quartile of the response conditioning on the covariates. Following Wang, Wu and Li (2012), an additional independent data set of size $10n$ is generated to select the tuning parameter λ by minimizing the estimated prediction error based on the quantile check loss function.

We denote the final estimator by $\hat{\beta}_\tau = (\hat{\beta}_1, \hat{\beta}_2, \dots, \hat{\beta}_p)^T$. Note that the coefficients of covariates removed by the screening procedure are directly set to be zero in the final estimator. Based on 100 repetitions, we evaluate the performance in terms of the following criteria.

Size: The average number of non-zero estimated regression coefficients

$$\hat{\beta}_j \neq 0 \quad \text{for } 1 \leq j \leq p;$$

C: The average number of truly non-zero coefficients correctly estimated to be non-zero;

IC: The average number of truly zero coefficients incorrectly estimated to be non-zero;

AE: The average of absolute estimation error of $\hat{\beta}_\tau$, which is defined by

$$\sum_{j=1}^p |\hat{\beta}_j \text{sign}(\hat{\beta}_{j,1})| / \|\hat{\beta}_\tau\| - \beta_{0j} \text{sign}(\beta_{0j,1}) / \|\beta_0\|.$$

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 4.2. In each column, the value represents the mean of 100 replicates with its sample standard deviation in the parentheses. For two different random errors and different quantiles, the first three columns demonstrate that the LASSO is relatively conservative and tends to select larger models while the SCAD and the MCP are consistent to select the true model. The relatively small values in the column labeled “AE” shows that the proposed penalized linear quantile regression procedure can produce consistent estimators and support the theoretical findings in Theorem 1. In conclusion, the satisfactory simulation results demonstrate 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

In this section we conduct an empirical study of the Cardiomyopathy microarray dataset. 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 this analysis is to determine the most influential genes for the response.

We display both the boxplot and the histogram of Y in Figure 4.1. Both indicate that the response distribution may be heavy-tailed and contains outliers. We first implement 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 real data analysis. Thus, we only present results of DC-RoSIS with regularized quantile regression with different penalties in this example. The DC-RoSIS selects the two genes, labeled 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 apply our proposed penalized linear quantile regression 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 choose the quantile levels $\tau = 0.25, 0.50$ and 0.75 , and three different penalties, LASSO, SCAD and MCP. We use BIC to select the tuning parameters for each method. With the estimated single index, denoted $(\mathbf{x}^T \hat{\beta}_\tau)$, we apply the cubic splines to estimate the quantile functions $\hat{q}_\tau(\cdot)$ of model (1.3), or

equivalently, model (1.4). Figure 4.2 depicts the estimated curves of $\hat{q}_\tau(\mathbf{x}^T \hat{\beta}_\tau)$ 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 method, denoted by “Size” in Table 4.3. In addition, to evaluate the goodness of fit for each model, we follow the idea of R^2 for the linear model and define the quantile-adjusted R^2 (“Q- R^2 ”) as follows,

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

where $\rho_\tau(\cdot)$ is the τ th quantile check loss function, $\hat{q}_\tau(\cdot)$ is the cubic-spline estimate of $q_\tau(\cdot)$, $\hat{q}_\tau(\mathbf{x}^T \hat{\beta}_\tau)$ is the τ -th quantile function of Y_i , and $\hat{Y}_\tau$ is the sample τ th quantile of Y . The larger Q- R^2 is, the better the model fit is. For example, for $\tau = 0.75$, SCAD selected 3 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 4.3. In addition, we conduct 100 random partitions to examine the prediction performance. For each partition, we randomly select 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 4.3. 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 and its

associated robust estimate of the standard deviation (i.e. interquartile range/1.34) in the parentheses. In conclusion, 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

In this paper, we first study the regularized quantile regression for ultrahigh dimensional single-index models, in which the conditional distribution of the response depends on the covariates via a single-index structure. The consistency and the oracle property for the penalized linear quantile regression estimator have been established under the sparsity condition. Then, we propose a robust independence screening based on the distance correlation between the distribution function of the response variable and each covariate, called as DCRoSIS. The new DC-RoSIS enjoys the sure screening property under even milder conditions than the existing alternative methods. It can be applied before the regularized quantile regression to reduce the covariate dimension from ultrahigh dimensionality to a moderate scale. The numerical studies show that the proposed methodology has the excellent finite-sample performance compared with other methods.

The proposed method has reliable performance when the distribution of the response variable is heavily tailed or response realizations contain extreme values. An interesting point raised by a referee is concerned with the performance of proposed procedure in the presence of heavy-tail predictors or extreme outliers contained in the predictors. In this case, Condition (C1) will be violated and the proposed method may fail. However, we may simply use $F_k(X_k)$, the distribution function of X_k , in place of X_k in the proposed screening procedure. This replacement will help us remove condition (C1) and achieve the robustness feature in the x -direction. Please 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 the penalized linear quantile regression would be an interesting topic for future research.

Theorem 1 implies that the oracle estimator is a local minimizer of the objective function (2.2) with the probability approaching one as $n \rightarrow \infty$. Thus, the proposed method can detect the non-zero components of the true coefficient and simultaneously estimate its direction. For the statistical inference purpose, if one may be interested in the asymptotical distribution of the regularized quantile estimator, 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 becomes much more challenging to derive the asymptotic normality. But it is an interesting and potential research direction.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgment

Zhong's research was supported by National Natural Science Foundation of China (NNSFC) 11301435, 71131008 and the Fundamental Research Funds for the Central Universities. Zhu's research was supported by NNSFC 11371236 and 11422107, Innovation Program of Shanghai Municipal Education Commission 13ZZ055, Pujiang Project of Science and Technology Commission of Shanghai Municipality 12PJ1403200 and Program for New Century Excellent Talents, Ministry of Education of China NCET-12-0901. Runze Li is the corresponding author and his 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 and 11231010, Key project of Beijing Municipal Educational Commission and Beijing Center for Mathematics and Information Interdisciplinary Sciences. The authors thank the Editor, the Associate Editor and three anonymous 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

We require the following lemma to prove Lemma 2.

Lemma 3

According to (2.4), $E \left\{ I \left(Y - \mathbf{x}_A^T \beta_{\tau 1}^o \leq u_\tau^o \right) | \mathbf{x}_A \right\} = \tau$. That is, the oracle estimator u_τ^o of u is the τ th quantile of $Y - \mathbf{x}_A^T \beta_{\tau 1}^o$ conditional on $\mathbf{x}_A$.

Proof of Lemma 3—Let ξ_τ be the τ th quantile of $Y - \mathbf{x}_A^T \beta_{\tau 1}^o$ conditional on $\mathbf{x}_A$. By definition, we have that $E \left\{ I \left(Y - \mathbf{x}_A^T \beta_{\tau 1}^o \leq \xi_\tau \right) | \mathbf{x}_A \right\} = \tau$. It suffices to show $\mathcal{L}_\tau(\xi_\tau, \beta_{\tau 1}^o) \leq \mathcal{L}_\tau(u, \beta_{\tau 1}^o)$ holds for any u . To be specific,

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

where the second equality follows from Knight (1998). In the second equality, the first term 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 the idea of 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 ,

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

We define that

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

where the second equality follows from Knight (1998)'s identity.

Note that $E \left\{ I \left(Y - \mathbf{x}_A^T \beta_{\tau 1}^o \leq u_\tau^o \right) | \mathbf{x}_A \right\} = \tau$ by Lemma 3 and hence $E(I_{n1}) = 0$.

$$\begin{aligned}
\text{var}(I_{n1}) &= E \{ \text{var}(I_{n1} | \mathbf{x}_A) \} = \text{var} \{ E(I_{n1} | \mathbf{x}_A) \} \\
&= \tau(1 - \tau) q_n^{-1} E \left\{ n^{-1} \sum_{i=1}^n \left(u + \mathbf{x}_{i,A}^T \gamma \right)^2 \right\} \\
&\leq 2\tau(1 - \tau) q_n^{-1} \left[u^2 + \lambda_{\max} \left\{ E \left(\mathbf{x}_A \mathbf{x}_A^T \right) \right\} \|\gamma\|^2 \right] \leq C q_n^{-1} (\Delta_1^2 + \Delta_2^2),
\end{aligned}$$

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

Next we evaluate I_{n2} . Denote by $F(\cdot | \mathbf{x}_A)$ and $f(\cdot | \mathbf{x}_A)$ the conditional distribution and density of $(Y - \mathbf{x}_A^T \beta_{\tau 1}^o)$ given $\mathbf{x}_A$, respectively

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

where the first inequality follows by Condition (C3) and the last inequality follows by

Condition (C2). Therefore, $E(I_{n2}) = O(1) (\Delta_1^2 + \Delta_2^2)$. Next we consider the variance of I_{n2} ,

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

which converges to zero as $n \rightarrow \infty$ because $q_n = O(n^{c1})$. 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 $n \rightarrow \infty$. Therefore, for any fixed $\eta > 0$, there exists two constants Δ_1 and Δ_2 such that for all sufficiently large n , we have $G_n(u, \gamma) > 0$ 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) to demonstrate the technical proof of Theorem 1. Note that their moment conditions on $\mathbf{x}$ are different. With

slightly notational abuse, we write $\mathbf{x}_A = (1, \mathbf{x}_A)^T$, $\beta_\tau^o = (u_\tau^o, \beta_\tau^{oT})^T$ defined in (2.4),

$\hat{\beta}_\tau = (\hat{u}_\tau, \hat{\beta}_\tau^T)^T$ and $\hat{\beta}_\tau^o = (\hat{u}_\tau^o, \hat{\beta}_\tau^{oT})^T$, where $\hat{\beta}_\tau$ denotes the penalized linear quantile estimator defined in (2.3) and $\hat{\beta}_\tau^o = (\hat{\beta}_{\tau 1}^{oT}, \mathbf{0}^T)^T$ is the oracle estimator defined in

(2.5). Accordingly, we write $\beta_{\tau 1}^o = (u_\tau^o, \beta_{\tau 1}^{oT})^T$ and $\hat{\beta}_{\tau 1}^o = (\hat{u}_\tau^o, \hat{\beta}_{\tau 1}^{oT})^T$.

We first write the objective function (2.2) of the penalized linear quantile regression as the difference of two convex functions in β . Here, we only consider the proof for the SCAD penalty, and the proof for the MCP penalty can be achieved by the similar arguments. To be precise, $Q(\beta) = g(\beta) - h(\beta)$, where

$g(\beta) = n^{-1} \sum_{i=1}^n \rho_\tau(Y_i - \mathbf{x}_i^T \beta) + \lambda \sum_{j=1}^{p_n} |\beta_j|$, and $h(\beta) = \sum_{j=1}^{p_n} H_\lambda(\beta_j)$, with

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

Thus, the subdifferential of $h(\beta)$ at any β is

$$\partial h(\beta) = \left\{ \mu = (\mu_0, \mu_1, \dots, \mu_{p_n})^T \in \mathbb{R}^{p_n+1} : \mu_0 = 0, \mu_j = \frac{\partial h(\beta)}{\partial \beta_j}, j=1, 2, \dots, p_n \right\}.$$

The subdifferential of $g(\beta)$ at any β is

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

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

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

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

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

Next we present Lemmas 4, 5 and 6 to facilitate the proof of Theorem 1. Tao and An (1997) proposed the numerical algorithm based on the convex difference representation, which is stated in Lemma 4. Lemmas 5 and 6 characterize the properties of the oracle estimator $\hat{\beta}_\tau^o$ and the associated subgradient functions $s(\hat{\beta}_\tau^o)$ respectively.

Lemma 4

(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 $h(\mathbf{x}) \cap g(\mathbf{x}^*) = \emptyset, \forall \mathbf{x} \in U \cap \text{dom}(g)$. Then $\mathbf{x}^*$ is a local minimizer of $g(\mathbf{x}) - h(\mathbf{x})$.

Lemma 5

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

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

Proof of Lemma 5—This lemma is parallel to Lemma 2.2 in Wang, Wu and Li (2012). The unpenalized quantile loss objective function is convex. By the convex optimization theory, $0 \in \partial \Sigma_{i=1}^n \rho_\tau(\mathbf{Y}_i - \mathbf{x}_i^T \hat{\beta}_\tau^o)$. Therefore, there exists v_i^* such that $s_j(\hat{\beta}_\tau^o) = 0$ with $v_i = v_i^*$ for $j = 0, 1, \dots, q_n$. On the other hand,

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

Condition (C5) requires that $\min_{1 \leq j \leq q_n} |\beta_{\tau,j}^o| \geq C_n^{-(1-c_2)/2}$. In addition,

$$\max_{1 \leq j \leq q_n} |\hat{\beta}_j^o - \beta_{\tau,j}^o| \leq \|\hat{\beta}_\tau^o - \beta_{\tau 1}^o\| = O_p\left(\sqrt{q_n/n}\right) = O_p\left(n^{-(1-c_1)/2}\right) = o_p\left(n^{-(1-c_2)/2}\right).$$

Therefore, $\min_{1 \leq j \leq q_n} |\hat{\beta}_j^o| \geq C_n^{-(1-c_2)/2} - o_p\left(n^{-(1-c_2)/2}\right)$, where c_1 and c_2 are defined in conditions (C4) and (C5) respectively. For $\lambda = o\{n^{-(1-c_2)/2}\}$, we have that, with probability approaching one, $|\hat{\beta}_j^o| \geq (a+1/2)\lambda, j=1, \dots, q_n$, which completes the proof.

Lemma 6

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

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

Proof of Lemma 6—This lemma is parallel to Lemma 2.3 in Wang, Wu and Li (2012).

Since the $\hat{\beta}_{\tau}^o$ is the oracle estimator, $|\hat{\beta}_j^o| = 0, j = q_n + 1, \dots, p_n$. It remains to show that

$$Pr\left(|s_j(\hat{\beta}_{\tau}^o)| > \lambda, \quad \text{for some } j = q_n + 1, \dots, p_n\right) \rightarrow 0, \quad \text{as } n \rightarrow \infty.$$

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

$$\begin{aligned} & s_j(\hat{\beta}_{\tau}^o) \\ &= (1 - \tau) n^{-1} \sum_{i=1}^n X_{ij} I(Y_i - \mathbf{x}_i^T \hat{\beta}_{\tau}^o < 0) - \tau n^{-1} \sum_{i=1}^n X_{ij} I(Y_i - \mathbf{x}_i^T \hat{\beta}_{\tau}^o > 0) - n^{-1} \sum_{i=1}^n X_{ij} v_i, \\ &= n^{-1} \sum_{i=1}^n X_{ij} \{I(Y_i - \mathbf{x}_i^T \hat{\beta}_{\tau}^o \leq 0) - \tau\} - n^{-1} \sum_{i=1}^n X_{ij} \{v_i + (1 - \tau) I(Y_i - \mathbf{x}_i^T \hat{\beta}_{\tau}^o = 0)\} \\ &= n^{-1} \sum_{i=1}^n X_{ij} \{I(Y_i - \mathbf{x}_{i,A}^T \hat{\beta}_{\tau 1}^o \leq 0) - \tau\} - n^{-1} \sum_{i \in D} X_{ij} [v_i^* + (1 - \tau)], \end{aligned}$$

where $v_i^* \in [\tau - 1, \tau]$ with $i \in D$ satisfies $s_j(\hat{\beta}_{\tau}^o)$ with $v_i = v_i^*$, for $j = 1, \dots, q_n$ by Lemma 5.

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

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

$$\begin{aligned} T_{n2} & \leq Pr\left(\max_{j=q_n+1, \dots, p_n} |n^{-1} \sum_{i \in D} X_{ij} 1\{|X_{ij}| \leq M\} \{v_i^* + (1 - \tau)\}| > \lambda/2\right) \\ & \quad + Pr\left(\max_{j=q_n+1, \dots, p_n} |n^{-1} \sum_{i \in D} X_{ij} 1\{|X_{ij}| > M\} [v_i^* + (1 - \tau)]| > \lambda/2\right) \\ & =: T_{n21} + T_{n22} \end{aligned}$$

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

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

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

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

because that if $|X_{ij}| \leq M$ for all $i \in D$, then $n^{-1} \sum_{i \in D} X_{ij} 1\{|X_{ij}| < M\} = 0$.

Therefore,

$$\begin{aligned} T_{n22} & \leq p_n \max_{j=q_n+1, \dots, p_n} \Pr \left(\left| n^{-1} \sum_{i \in D} X_{ij} 1\{|X_{ij}| > M\} [v_i^* + (1 - \tau)] \right| > \lambda/2 \right) \\ & \leq p_n (q_n + 1) \max_{i \in D, q_n+1 \leq j \leq p_n} \Pr(|X_{ij}| > M) \leq p_n (q_n + 1) \exp(-tM) E\{\exp(t|X_{ij}|)\} \\ & \leq Cp_n (q_n + 1) \exp(-tM) = Cp_n O(n^{c_1}) \exp(-tn^\theta) \rightarrow 0, \end{aligned}$$

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

It remains to show that

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

as $n \rightarrow \infty$. We consider

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

First, let us consider J_{n1} . We choose a $M = O(n^\theta)$ with $0 < \theta < (c_2 - c_1)/2$, then

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

By Hoeffding's inequality, we have that

$$Pr \left(|n^{-1} \sum_{i=1}^n X_{ij} 1(|X_{ij}| \leq M) \{I(Y_i - \mathbf{x}_{i,A}^T \beta_{\tau 1}^o \leq 0) - \tau\} | > \lambda/4 \right) \leq 2 \exp \left(-\frac{n\lambda^2}{8M^2} \right).$$

Thus, $J_{n11} = 2p_n \exp\{-n\lambda^2/(8M^2)\} = 2p_n \exp(-n^{1-2\theta}\lambda^2/8) \rightarrow 0$, as $n \rightarrow \infty$, 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

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

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

Following similar 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 $0 < \theta < (c_2 - c_1)/2$,

$$\begin{aligned}
J_{n3} & \leq Pr \left(\max_{j=q_n+1, \dots, p_n} \sup_{\|\beta_1 - \beta_{\tau 1}^o\| \leq \Delta \sqrt{q_n/n}} |n^{-1} \sum_{i=1}^n X_{ij} 1\{|X_{ij}| \leq M\} \right. \\
& \quad \left. \{Pr(Y_i - \mathbf{x}_{i,A}^T \beta_1 \leq 0) - Pr(Y_i - \mathbf{x}_{i,A}^T \beta_{\tau 1}^o \leq 0)\} | > \frac{\lambda}{8} \right) \\
& \quad + Pr \left(\max_{j=q_n+1, \dots, p_n} \sup_{\|\beta_1 - \beta_{\tau 1}^o\| \leq \Delta \sqrt{q_n/n}} |n^{-1} \sum_{i=1}^n X_{ij} 1\{|X_{ij}| > M\} \right. \\
& \quad \left. \{Pr(Y_i - \mathbf{x}_{i,A}^T \beta_1 \leq 0) - Pr(Y_i - \mathbf{x}_{i,A}^T \beta_{\tau 1}^o \leq 0)\} | > \frac{\lambda}{8} \right) \\
& =: J_{n31} + J_{n32}.
\end{aligned}$$

To handle J_{n31} , we observe that

$$\begin{aligned}
& \max_{j=q_n+1, \dots, p_n} \sup_{\|\beta_1 - \beta_{\tau_1}^o\| \leq \Delta \sqrt{q_n/n}} |n^{-1} \sum_{i=1}^n X_{ij} 1\{|X_{ij}| > M\} \\
& \quad \left\{ Pr\left(Y_i - \mathbf{x}_{i,A}^T \beta_1 \leq 0\right) - Pr\left(Y_i - \mathbf{x}_{i,A}^T \beta_{\tau_1}^o \leq 0\right) \right\} | \\
& \leq M \sup_{\|\beta_1 - \beta_{\tau_1}^o\| \leq \Delta \sqrt{q_n/n}} |E\{f(\zeta | \mathbf{x}_A) \mathbf{x}_A^T (\beta_1 - \beta_{\tau_1}^o)\}| \\
& \leq M \sup_{\|\beta_1 - \beta_{\tau_1}^o\| \leq \Delta \sqrt{q_n/n}} \lambda_{\max}^{1/2} \left\{ E\left(\mathbf{x}_A \mathbf{x}_A^T\right) \right\} \|\beta_1 - \beta_{\tau_1}^o\| \\
& \leq O\left\{n^\theta (q_n/n)^{1/2}\right\} = O\left\{n^{-(1-c_1-2\theta)/2}\right\},
\end{aligned}$$

where $f(\cdot | \mathbf{x}_A)$ is defined in Condition (C3) with ζ is between $u_\tau^o + \mathbf{x}_A^T (\beta_1 - \beta_{\tau_1}^o)$ and u_τ^o and thus the second inequality follows Condition (C3) and Cauchy-Schwartz inequality, and the third inequality follows Condition (C2). Consequently, together with $\lambda = o\{n^{-(1-c_2)/2}\}$, we have that $J_{n31} \Pr\{O(n^{-(1-c_1-2\theta)/2}) > \lambda/8\} = o(1)$ if $0 < \theta < (c_2 - c_1)/2$. We can also follow similar arguments for handling J_{n12} and obtain that $J_{n32} = o(1)$. Therefore, $J_{n3} = J_{n31} + J_{n32} = o(1)$. Consequently,

$$Pr\left\{\max_{j=q_n+1, \dots, p_n} |n^{-1} \sum_{i=1}^n X_{ij} \left\{I\left(Y_i - \mathbf{x}_{i,A}^T \hat{\beta}_{\tau_1}^o < 0\right) - \tau\right\}| > \lambda\right\} \leq J_{n1} + J_{n2} + J_{n3} = o(1),$$

which implies that $Pr\left\{|s_j(\hat{\beta}_\tau^o)| < \lambda \text{ for some } j = q_n+1, \dots, p_n\right\} \rightarrow 0$. This completes the proof of Lemma 6.

With Lemmas 5 and 6 for the random $\mathbf{x}$ with sub-exponential tail probability Condition (C1), we can follow the technical proof of Theorem 2.4 of Wang, Wu and Li (2012) to obtain the oracle property and complete the proof.

Proof of Theorem 2

For notational clarity, we use c_1 and c_2 to denote two different generic positive constants.

First we assume $F(y)$ is known. That is, $\widehat{dcov}^{*2}\{X_k, F(Y)\} = \hat{S}_{k1}^* + \hat{S}_{k2}^* - 2\hat{S}_{k3}^*$ where

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

Similarly we define $\hat{\omega}_k^* = \widehat{dcorr}^{*2}\{X_k, F(Y)\}$. 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

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

To prove Theorem 2, it thus suffices to show the difference between $\hat{\omega}_k^*$ and $\hat{\omega}_k$ defined in (3.5) is ignorable when size n is large enough, which amounts to studying the differences between $\hat{S}_{km}^*$ and $\hat{S}_{km}$ for $m = 1, 2, 3$. Next, 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

$$\hat{S}_{k1}^* = \frac{1}{n^2} \sum_{i=1}^n \sum_{j=1}^n |X_{ik} - X_{jk}| |F(Y_i) - F(Y_j)| \quad \text{and} \quad \hat{S}_{k1} = \frac{1}{n^2} \sum_{i=1}^n \sum_{j=1}^n |X_{ik} - X_{jk}| |F_n(Y_i) - F_n(Y_j)|$$

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

where M is a positive constant specified later,

$$A_n = n^{-2} \sum_{i=1}^n \sum_{j=1}^n (X_{ik} - X_{jk})^2, \quad \text{and} \quad B_n = n^{-2} \sum_{i=1}^n \sum_{j=1}^n \{ |F(Y_i) - F(Y_j)| - |F_n(Y_i) - F_n(Y_j)| \}^2$$

Using the argument that $||x| - |y|| \leq |x - y| \leq |x| + |y|$, we obtain that

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

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

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

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

For the second term, for all $0 < s \leq 2s_0$, where s_0 is defined in Condition (C1),

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

where C is a positive constant, the second inequality follows Markov's inequality and the last inequality is applied under Condition (C1).

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 ,

$$\begin{aligned}
\Pr \left(\max_{1 \leq k \leq p_n} |\hat{S}_{k1}^* - \hat{S}_{k1}| \geq \varepsilon \right) &\leq 2 \exp(-n\varepsilon^2/8M^2) + C \exp(-sM) \\
&\leq 2 \exp(-c_1\varepsilon^2 n^{1-2\gamma}) + C \exp(-c_2 n^\gamma).
\end{aligned} \quad (C.4)$$

Thus, it is not difficult to show that

$$\Pr \left(\max_{1 \leq k \leq p} |\hat{\omega}_k^* - \hat{\omega}_k| \geq cn^{-\kappa} \right) \leq O \left(\exp \left\{ -c_1 n^{1-2(\kappa+\gamma)} \right\} + \exp(-c_2 n^\gamma) \right). \quad (C.5)$$

Therefore, (C.1) and (C.5) together completes the proof of Theorem 2.

References

- Altham PME. Improving the precision of estimation by fitting a generalized linear model and quasi-likelihood. *Journal of the Royal Statistical Society, Series B*. 1984; 46:118–119.
- Bickel P, Levina E. Regularized estimation of large covariance matrices. *Annals of Statistics*. 2008; 36:199–227.
- Dvoretzky A, Kiefer J, Wolfowitz J. Asymptotic minimax character of the sample distribution function and of the classical multinomial estimator. *Annals of Mathematical Statistics*. 1956; 27:642–669.
- Fan Y, Fan J, Barut E. Adaptive Robust Variable Selection. *Annals of Statistics*. 2014; 42:324–351. [PubMed: 25580039]
- Fan J, Li R. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*. 2001; 96:1348–1360.
- Fan J, Lv J. Sure independence screening for ultrahigh dimensional feature space (with discussion). *Journal of the Royal Statistical Society, Series B*. 2008; 70:849–911.
- Fan Y, Lv J. Asymptotic equivalence of regularization methods in thresholded parameter space. *Journal of the American Statistical Association*. 2013; 108:1044–1061.
- Fan J, Samworth R, Wu Y. Ultrahigh dimensional feature selection: beyond the linear model. *Journal of Machine Learning Research*. 2009; 10:1829–1853.
- Hall P, Li KC. On almost linearity of low dimensional projection from high dimensional data. *Annals of Statistics*. 1993; 21:867–889.
- Hall P, Miller H. Using generalized correlation to effect variable selection in very high dimensional problems. *Journal of Computational and Graphical Statistics*. 2009; 18:533–550.
- Härdle W, Hall P, Ichimura H. Optimal smoothing in single-index models. *Annals of Statistics*. 1993; 21:157–178.
- He X, Shao Q. On parameters of increasing dimensions. *Journal of Multivariate Analysis*. 2000; 73:120–135.
- Knight K. Limiting distributions for L_1 regression estimators under general conditions. *The Annals of Statistics*. 1998; 26:755–770.
- Koenker, R. *Quantile Regression*. Cambridge University Press; 2005.

- Kong E, Xia Y. Variable Selection for the single-index model. *Biometrika*. 2007; 94:217–229.
- Li G, Peng H, Zhang J, Zhu L. Robust rank correlation based screening. *The Annals of Statistics*. 2012; 40:1846–1877.
- Li K-C. Sliced inverse regression for dimension reduction. *Journal of the American Statistical Association*. 1991; 86:316–327.
- Li R, Zhong W, Zhu L. Feature screening via distance correlation learning. *Journal of the American Statistical Association*. 2012; 107:1129–1139. [PubMed: 25249709]
- Liang H, Liu X, Li R, Tsai C. Estimation and testing for partially linear single-index models. *Annals of Statistics*. 2010; 38:3811–3836. [PubMed: 21625330]
- Naik PA, Tsai C-L. Single-index model selections. *Biometrika*. 2001; 88:821–832.
- Powell JL, Stock JH, Stoker TM. Semiparametric estimation of index coefficient. *Econometrica*. 1989; 51:1403–1430.
- Segal MR, Dahlquist KD, Conklin BR. Regression approach for microarray data analysis. *Journal of Computational Biology*. 2003; 10:961–980. [PubMed: 14980020]
- Szekely GJ, Rizzo ML, Bakirov NK. Measuring and testing dependence by correlation of distances. *Annals of Statistics*. 2007; 35:2769–2794.
- Tao PD, An LTH. Convex analysis approach to D.C. programming: theory, algorithms and applications. *Acta Mathematica Vietnamica*. 1997; 22:289–355.
- Tibshirani R. Regression shrinkage and selection via LASSO. *Journal of the Royal Statistical Society, Series B*. 1996; 58:267–288.
- Wang L, Wu Y, Li R. Quantile regression for analyzing heterogeneity in ultra-high dimension. *Journal of the American Statistical Association*. 2012; 107:214–222. [PubMed: 23082036]
- Wu Y, Liu Y. Variable selection in quantile regression. *Statistica Sinica*. 2009; 19:801–817.
- Zhang C. Nearly unbiased variable selection under minimax concave penalty. *Annals of Statistics*. 2010; 101:1418–1429.
- Zheng Z, Fan Y, Lv J. High dimensional thresholded regression and shrinkage effect. *Journal of the Royal Statistical Society Series B*. 2014; 76:627–649.
- Zhu L, Huang M, Li R. Semiparametric quantile regression with high-dimensional covariates. *Statistica Sinica*. 2012; 22:1379–1401. [PubMed: 24501536]
- Zhu LP, Li L, Li R, Zhu LX. Model-free feature screening for ultrahigh dimensional data. *Journal of the American Statistical Association*. 2011; 106:1464–1475. [PubMed: 22754050]
- Zhu L, Qian L, Lin J. Variable selection in a class of single-index models. *Annals of the Institute of Statistical Mathematics*. 2011; 63:1277–1293.
- Zou H. The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*. 2006; 101:1418–1429.
- Zou H, Li R. One-step sparse estimates in nonconcave penalized likelihood models. *Annals of Statistics*. 2008; 36:1509–1533. [PubMed: 19823597]

Figure 4.1.
Exploratory Data Analysis: Histogram and Boxplot of Ro1.

Figure 4.2.

The estimated curves of $\hat{q}_\tau(\mathbf{x}^T \hat{\beta}_\tau)$ (the vertical axis) versus $(\mathbf{x}^T \hat{\beta}_\tau)$ (the horizontal axis) at different quantiles for different penalties. From left to right, $\tau = 0.25, 0.50$ and 0.75 ; From up to down, LASSO, SCAD and MCP.

Table 4.1

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

$\varepsilon \sim N(0, 1)$													$\varepsilon \sim \text{Cauchy Distribution}$												
Method		S	P_{s1}	P_{s2}	P_{s7}	P_a	Size	P_{s1}	P_{s2}	P_{s7}	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														

Table 4.2

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

$\varepsilon \sim N(0, 1)$				
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)
$\varepsilon \sim \text{Cauchy Distribution}$				
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)

Table 4.3

Empirical analysis of Cardiomyopathy microarray dataset.

Method	All Data		Partitioned Data		PE
	Size	Q-R ²	Ave Size	Ave Q-R ²	
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)