<table>
<thead>
<tr>
<th>Statistica Sinica Preprint No: SS-14-049R1</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Title</strong></td>
</tr>
<tr>
<td><strong>Manuscript ID</strong></td>
</tr>
<tr>
<td><strong>URL</strong></td>
</tr>
<tr>
<td><strong>DOI</strong></td>
</tr>
</tbody>
</table>
| **Complete List of Authors** | Wei Zhong  
Liping Zhu  
Runze Li and  
Hengjian Cui |
| **Corresponding Author** | Runze Li |
| **E-mail** | rzli@psu.edu |

Notice: Accepted version subject to English editing.
Regularized Quantile Regression and Robust Feature Screening for Single Index Models

Wei Zhong, Liping Zhu, Runze Li and Hengjian Cui

Xiamen University, Shanghai University of Finance and Economics
Pennsylvania State University and Capital Normal University

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 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.

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. Traditional single index regression model is referred to as

$$Y = m(\mathbf{x}^T \beta_0) + \varepsilon,$$  \hspace{1cm} (1.1)

where $m(\cdot)$ is an unknown regression function, $\beta_0$ consists of unknown index parameters, and $\varepsilon$ is a random error with $E(\varepsilon \mid \mathbf{x}) = 0$ and $\text{var}(\varepsilon \mid \mathbf{x}) = \sigma^2$. 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

$$Y = m(\mathbf{x}^T \beta_0) + \sigma(\mathbf{x}^T \beta_0)\varepsilon,$$  \hspace{1cm} (1.2)

for unknown functions $m(\cdot)$ and $\sigma(\cdot)$, where $\varepsilon$ has mean zero and is assumed to be independent of $\mathbf{x}$. Zhu, Huang and Li (2012) developed an estimation procedure for $\beta_0$ 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 $\mathbf{x} = (X_1, \ldots, X_{p_n})^T$. 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 $F(y \mid \mathbf{x})$ 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

$$F(y \mid \mathbf{x}) = F(y \mid \mathbf{x}^T \beta_0), \text{ for all } y \in \mathbb{R}.$$  \hspace{1cm} (1.3)

That is, the conditional distribution of $(Y \mid \mathbf{x})$ 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 $F(\cdot \mid \cdot)$ is unknown, the index parameter $\beta_0$ is not identifiable. The direction of $\beta_0$, 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 $\beta_0$ is sparse. Denote by $\mathcal{A}$ the active index set, $\beta_{\mathcal{A}}$ the nonzero entries of $\beta_0$ and $x_{\mathcal{A}}$ the collection of all active covariates. When $\beta_0$ is sparse, model (1.3) reduces to

$$F(y \mid x) = F(y \mid x_{\mathcal{A}}^\top \beta_{\mathcal{A}}), \text{ for all } y \in \mathbb{R}. \quad (1.4)$$

Our goal is to identify $\mathcal{A}$ and if possible, to estimate $\beta_{\mathcal{A}}$. 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 $(Y \mid x)$ always vary with the realizations of $(x^\top \beta_0)$. 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 $\beta_0$. 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 \((Y \mid x)\) 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 \(\infty\) 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 $\beta_0$ 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 $(Y \mid x)$ at different quantile levels are all functions of $(x^T \beta_0)$ and $(x^T A \beta_A)$ if the sparsity principle applies. This motivates us to estimate $\beta_0$ 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 $\beta_0$ in model (1.3). To be specific, we denote by $\rho_\tau(r) = \tau r - I(r < 0)$, the check loss function at the $\tau$th quantile, for $\tau \in (0, 1)$. Define

$$L_{\tau}(u, b) = E\{\rho_\tau(Y - u - x^T b)\} \text{ and } (u_\tau, \beta_\tau) = \arg\min_{u, b} L_{\tau}(u, b), \tag{2.1}$$

where $b = (b_1, \ldots, b_p)^T \in \mathbb{R}^p$.

**Lemma 1.** If $x$ satisfies the linearity condition that

$$E\{x - E(x) \mid x^T \beta_0\} = \text{var}(x)\beta_0 \{\beta_0^T \text{var}(x)\beta_0\}^{-1} \beta_0^T \{x - E(x)\},$$

then $\beta_\tau$ is proportional to $\beta_0$ in model (1.3).

The linearity condition is satisfied when $x$ 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 $\beta_0$ and $\beta_\tau$ coincide. To estimate the direction of $\beta_0$ in model (1.3), it amounts to estimating $\beta_\tau$ 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 $\beta_\tau$ in (2.1). Note that $\beta_\tau$ is identifi-
able because the linear quantile loss function $L_\tau(u, b)$ is convex. Suppose that
\[ (x_i, Y_i), i = 1, 2, \ldots, n \] is a random sample from (1.3). We consider the following penalized linear quantile regression to produce a sparse estimator of $\beta_\tau$:
\[
Q(u, b) = n^{-1} \sum_{i=1}^{n} \rho_\tau(Y_i - u - x_i^T b) + \sum_{j=1}^{p_n} p_\lambda(|b_j|),
\] (2.2)
where $p_\lambda(\cdot)$ is a penalty function with a regularization parameter $\lambda$. 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,
\[
p_\lambda(b) = \lambda |b| I(0 \leq |b| \leq \lambda) + \frac{a\lambda |b| - (b^2 + \lambda^2)/2}{a - 1} I(\lambda \leq |b| \leq a\lambda) + \frac{(a + 1)\lambda^2}{2} I(|b| > a\lambda),
\]
where $a = 3.7$ is suggested by Fan and Li (2001). By minimizing the objective function $Q(u, b)$, we obtain the estimators $(\hat{u}_\tau, \hat{\beta}_\tau)$ at the $\tau$-th quantile. In symbols, the resulting estimators are given by
\[
(\hat{u}_\tau, \hat{\beta}_\tau) = \text{argmin}_{u, b} \{Q(u, b)\}. \tag{2.3}
\]

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 $x$ 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, $A = \{1, 2, \ldots, q_n\}$. We define the oracle estimator at the population level by
\[
L_\tau(u, b_1) = E\{\rho_\tau(Y - u - x_A^T b_1)\} \quad \text{and} \quad (u_\tau^0, \beta_\tau^0) = \text{argmin}_{u, b_1} \{L_\tau(u, b_1)\}, \tag{2.4}
\]
where $b_1 = (b_1, \ldots, b_{q_n})^T \in \mathbb{R}^{q_n}$. We further write $\beta_\tau^0 = (\beta_{\tau_1}^0, 0^T)^T$, where $\beta_{\tau_1}^0$ 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 $\hat{\beta}^o = (\hat{\beta}_{r1}^o, \mathbf{0}^T)^T$ at the sample level by

$$
\ell_{\tau n}(u, \mathbf{b}_1) = n^{-1} \sum_{i=1}^{n} \{\rho_{\tau}(Y_i - u - \mathbf{x}_i^T \mathbf{b}_1)\} \quad \text{and} \\
(\hat{u}_{\tau}^o, \hat{\beta}_{r1}^o) = \arg\min_{u, \mathbf{b}_1} \ell_{\tau n}(u, \mathbf{b}_1). 
$$

(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 $0 < C_1 \leq C_2 < \infty$, such that

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

where $\lambda_{\min}$ and $\lambda_{\max}$ represent the smallest and largest eigenvalues, respectively. Assume further that $\{(\mathbf{x}_i, Y_i), i = 1, \ldots, n\}$ 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 $\infty$ in the neighborhood around $u_{\tau}^o$.

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

(C5) For $\beta_{r1}^o = (\beta_{r1}^o, \beta_{r2}^o, \ldots, \beta_{r_{q_n}}^o)^T$, there exist positive constants $c_2$ and $C$ such that $2c_1 < c_2 \leq 1$ and

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

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 \((Y - x^\tau \beta)\) conditional on \(x\). 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 \(\hat{u}_r^o\) and \(\hat{\beta}_r^{oT}\).

**Lemma 2.** Under Conditions (C1)-(C4), the oracle estimators \(\hat{u}_r^o\) and \(\hat{\beta}_r^{oT}\) satisfy

\[
\|\hat{\beta}_r^{oT} - \beta_r^{oT}\| = O_p\left(\sqrt{q_n/n}\right) \quad \text{and} \quad \|\hat{u}_r^o - u_r^o\| = O_p\left(\sqrt{q_n/n}\right).
\]  

(2.7)

Next we study the theoretical property of the oracle estimator \(\hat{\beta}_r^{oT}\).

**Theorem 1.** (The Oracle Property) Suppose Conditions (C1)-(C5) hold, and \(\log p_n = o(n^{-\min\{c_2/2, \theta\}})\) with \(0 < \theta < (c_2 - c_1)/2\) and \(\lambda = o\left\{n^{-\left(1 - c_2\right)/2}\right\}\). Let \(B_n(\lambda)\) be the set of local minima \(\hat{\beta}_r\) of the objective function \(Q(u, b)\) defined in (2.2) with the SCAD or the MCP penalty and the tuning parameter \(\lambda\). The oracle estimator \(\hat{\beta}_r^o = (\hat{\beta}_r^{oT}, 0^T)^T\) satisfies

\[
\Pr\left\{\hat{\beta}_r^o \in B_n(\lambda)\right\} \to 1, \quad \text{as} \quad n \to \infty.
\]

Theorem 1 implies that the oracle estimator \(\hat{\beta}_r^o\) is a local minimizer of the objective function (2.2) with the 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, 2 and Theorem 1 imply that \(\hat{\beta}_r\) from the penalized linear quantile regression is a consistent estimator of the direction of \(\beta_0\) in model (1.3). It can detect the non-zero components of \(\beta_0\) 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 \(Q(u, b)\).
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 tolerable 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

\[
\text{dcov}^2(X, Y) = S_1 + S_2 - 2S_3, \tag{3.1}
\]

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

\[
\text{dcorr}(X, Y) = \frac{\text{dcov}(X, Y)}{\sqrt{\text{dcov}(X, Y) \text{dcov}(Y, Y)}}. \tag{3.2}
\]

Szekely, Rizzo and Bakirov (2007) pointed out that \( \text{dcorr}(X, Y) = 0 \) if and only
if \(X\) and \(Y\) are independent and \(\text{dcorr}(X,Y)\) 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 \(k = 1, \ldots, p_n\) 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,

\[
\omega_k = \text{dcorr}\{X_k, F(Y)\}, \tag{3.3}
\]

where \(F(y) = E\{1(Y \leq y)\}\) and \(1(\cdot)\) 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 \(F_k(x) = E\{1(X_k \leq x)\}\).

In the sequel we introduce how to implement the marginal utility (3.3) in the screening procedure. Let \(\{(x_i, Y_i), i = 1, \ldots, n\}\) be a random sample from the population \((x, Y)\). We first estimate the distance covariance between \(X_k\) and \(F(Y)\) through the moment estimation method,

\[
\hat{\text{dcov}}^2 \{X_k, F(Y)\} = \hat{S}_{k,1} + \hat{S}_{k,2} - 2\hat{S}_{k,3}, \tag{3.4}
\]
where

\[ \hat{S}_{k,1} = \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} \left| X_{ik} - X_{jk} \right| \left| F_n(Y_i) - F_n(Y_j) \right|, \]

\[ \hat{S}_{k,2} = \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} \left| X_{ik} - X_{jk} \right| \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} \left| F_n(Y_i) - F_n(Y_j) \right|, \]

\[ \hat{S}_{k,3} = \frac{1}{n^3} \sum_{i=1}^{n} \sum_{j=1}^{n} \sum_{l=1}^{n} \left| X_{ik} - X_{lk} \right| \left| F_n(Y_j) - F_n(Y_l) \right|. \]

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

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

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

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

for some pre-specified thresholds \( c > 0 \) and \( 0 \leq \kappa < 1/2 \). The constants \( c \) and \( \kappa \) 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

\[ \Pr \left( \max_{1 \leq k \leq p} \left| \hat{\omega}_k - \omega_k \right| \geq cn^{-\kappa} \right) \leq O \left( p \left[ \exp \left\{ -c_1 n^{1-2(\kappa+\gamma)} \right\} + n \exp \left\{ -c_2 n^\gamma \right\} \right] \right). \] (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 \( \min_{k \in A} \omega_k \geq 2c n^{-\kappa} \), for some constants \( c > 0 \) and \( 0 \leq \kappa < 1/2 \).

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

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

where \( s_n \) is the cardinality of \( \mathcal{A} \). Thus, \( Pr(\mathcal{A} \subseteq \hat{\mathcal{A}}) \to 1 \) as \( n \to \infty \).

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 \( x = (X_1, X_2, \cdots, X_p)^T \) from \( N(0, \Sigma) \), 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. \( \mathcal{S} \): 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 $S$ 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 $d = 2[n/\log n]$ throughout.

3. $P_a$: 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 $P_{sj}$ 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):** $H(Y) = x^T\beta + \varepsilon$,

**Model (2):** $Y = \exp(2 - x^T\beta/2) + (2 - x^T\beta/2)^2 + \exp(x^T\beta/2)\varepsilon$,

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

**Model (4):** $Y = \beta_1X_1 + \beta_2X_2 + \beta_7X_7^2 + \varepsilon$,

where $\beta = (3, 1.5, 0, 0, 0, 2, 0, \ldots, 0)^T$. In the above four models, only $X_1, X_2$ and $X_7$ are truly important. The random error $\varepsilon$ 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 $(x^T\beta)$ 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 $\varepsilon$ follows Cauchy distribution. Even when $\varepsilon$ 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 $X_7$. 
Table 4.1: Performance comparison among different independence screening methods for four regression models with two different random errors.

<table>
<thead>
<tr>
<th>Method</th>
<th>$\varepsilon \sim N(0, 1)$</th>
<th>$\varepsilon \sim \text{Cauchy Distribution}$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$S$ $P_{S1}$ $P_{S2}$ $P_{S7}$ $P_a$</td>
<td>$P_{S1}$ $P_{S2}$ $P_{S7}$ $P_a$</td>
</tr>
<tr>
<td>Model (1)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\lambda = 1$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SIS</td>
<td>3.0(0.0) 1.00 1.00 1.00 1.00</td>
<td>794.5(217.5) 0.08 0.06 0.07 0.00</td>
</tr>
<tr>
<td>DC-SIS</td>
<td>3.0(0.0) 1.00 1.00 1.00 1.00</td>
<td>791.0(213.2) 0.06 0.07 0.00 0.00</td>
</tr>
<tr>
<td>SIRS</td>
<td>3.0(0.0) 1.00 1.00 1.00 1.00</td>
<td>702.5(246.6) 0.17 0.14 0.13 0.05</td>
</tr>
<tr>
<td>DC-RoSIS</td>
<td>3.0(0.0) 1.00 1.00 1.00 1.00</td>
<td>782.0(253.3) 0.12 0.09 0.06 0.00</td>
</tr>
<tr>
<td>Model (2)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SIS</td>
<td>8.0(14.2) 0.99 0.99 0.90 0.90</td>
<td>594.0(358.0) 0.72 0.62 0.57 0.52</td>
</tr>
<tr>
<td>DC-SIS</td>
<td>3.0(0.7) 1.00 1.00 0.99 0.99</td>
<td>300.0(222.4) 0.09 0.04 0.09 0.00</td>
</tr>
<tr>
<td>SIRS</td>
<td>3.0(0.0) 1.00 1.00 1.00 1.00</td>
<td>8.0(8.9) 1.00 1.00 0.98 0.98</td>
</tr>
<tr>
<td>RRCS</td>
<td>3.0(0.0) 1.00 1.00 1.00 1.00</td>
<td>782.0(253.3) 0.12 0.09 0.06 0.00</td>
</tr>
<tr>
<td>DC-RoSIS</td>
<td>3.0(0.0) 1.00 1.00 1.00 1.00</td>
<td>9.0(11.9) 1.00 1.00 0.96 0.96</td>
</tr>
<tr>
<td>Model (3)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SIS</td>
<td>876.5(217.5) 0.08 0.06 0.07 0.00</td>
<td>782.0(253.3) 0.12 0.09 0.06 0.00</td>
</tr>
<tr>
<td>DC-SIS</td>
<td>4.0(4.5) 1.00 1.00 0.97 0.97</td>
<td>700.0(130.8) 0.92 0.53 0.57 0.52</td>
</tr>
<tr>
<td>SIRS</td>
<td>7.0(8.2) 1.00 1.00 0.99 0.99</td>
<td>8.0(8.9) 1.00 1.00 0.98 0.98</td>
</tr>
<tr>
<td>RRCS</td>
<td>796.0(222.8) 0.10 0.10 0.09 0.00</td>
<td>782.0(253.3) 0.12 0.09 0.06 0.00</td>
</tr>
<tr>
<td>DC-RoSIS</td>
<td>8.0(9.7) 1.00 1.00 0.96 0.96</td>
<td>9.0(11.9) 1.00 1.00 0.96 0.96</td>
</tr>
<tr>
<td>Model (4)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SIS</td>
<td>270.5(400.6) 0.10 0.10 0.25 0.25</td>
<td>594.0(358.0) 0.72 0.62 0.57 0.52</td>
</tr>
<tr>
<td>DC-SIS</td>
<td>4.0(1.7) 1.00 1.00 1.00 1.00</td>
<td>8.0(8.9) 0.99 0.98 0.86 0.86</td>
</tr>
<tr>
<td>SIRS</td>
<td>427.0(419.6) 1.00 1.00 0.11 0.11</td>
<td>493.5(387.7) 1.00 1.00 0.09 0.09</td>
</tr>
<tr>
<td>RRCS</td>
<td>434.0(391.1) 1.00 1.00 0.13 0.13</td>
<td>477.5(394.0) 1.00 1.00 0.10 0.10</td>
</tr>
<tr>
<td>DC-RoSIS</td>
<td>4.0(1.5) 1.00 1.00 1.00 1.00</td>
<td>6.0(5.2) 1.00 1.00 0.99 0.99</td>
</tr>
</tbody>
</table>

that exhibits symmetric patterns with $Y$ in model (4). DC-RoSIS and DC-SIS have similar performances when $\varepsilon$ follows standard normal distribution. When $\varepsilon$ 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 $d = 2[n/\log(n)]$ top ranked covariates and then apply the penalized linear quantile regression to estimate the direction of $\beta$. 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 $\lambda$ by minimizing the estimated prediction error based on the quantile check loss function.

We denote the final estimator by $\hat{\beta}_\tau = (\hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_p)^T$. 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 $\hat{\beta}_j \neq 0$ 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

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

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 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)
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).

\[
\epsilon \sim \mathcal{N}(0, 1)
\]

<table>
<thead>
<tr>
<th>Method</th>
<th>Size</th>
<th>C</th>
<th>IC</th>
<th>AE</th>
</tr>
</thead>
<tbody>
<tr>
<td>LASSO ((\tau = 0.25))</td>
<td>18.16(6.28)</td>
<td>3.00(0.00)</td>
<td>15.16(6.28)</td>
<td>0.47(0.22)</td>
</tr>
<tr>
<td>LASSO ((\tau = 0.50))</td>
<td>18.14(6.33)</td>
<td>3.00(0.00)</td>
<td>15.14(6.33)</td>
<td>0.93(0.36)</td>
</tr>
<tr>
<td>LASSO ((\tau = 0.75))</td>
<td>13.97(6.16)</td>
<td>2.96(0.20)</td>
<td>11.01(6.15)</td>
<td>1.33(0.57)</td>
</tr>
<tr>
<td>SCAD ((\tau = 0.25))</td>
<td>3.46(0.86)</td>
<td>3.00(0.00)</td>
<td>0.46(0.86)</td>
<td>0.11(0.07)</td>
</tr>
<tr>
<td>SCAD ((\tau = 0.50))</td>
<td>3.68(1.58)</td>
<td>2.96(0.20)</td>
<td>0.72(1.56)</td>
<td>0.28(0.23)</td>
</tr>
<tr>
<td>SCAD ((\tau = 0.75))</td>
<td>3.47(1.58)</td>
<td>2.68(0.55)</td>
<td>0.79(1.52)</td>
<td>0.62(0.36)</td>
</tr>
<tr>
<td>MCP ((\tau = 0.25))</td>
<td>3.36(0.73)</td>
<td>3.00(0.00)</td>
<td>0.36(0.73)</td>
<td>0.11(0.07)</td>
</tr>
<tr>
<td>MCP ((\tau = 0.50))</td>
<td>3.53(1.23)</td>
<td>2.96(0.20)</td>
<td>0.57(1.21)</td>
<td>0.28(0.20)</td>
</tr>
<tr>
<td>MCP ((\tau = 0.75))</td>
<td>3.50(1.68)</td>
<td>2.68(0.55)</td>
<td>0.82(1.62)</td>
<td>0.63(0.36)</td>
</tr>
</tbody>
</table>

\[
\epsilon \sim \text{Cauchy Distribution}
\]

<table>
<thead>
<tr>
<th>Method</th>
<th>Size</th>
<th>C</th>
<th>IC</th>
<th>AE</th>
</tr>
</thead>
<tbody>
<tr>
<td>LASSO ((\tau = 0.25))</td>
<td>23.75(6.63)</td>
<td>3.00(0.00)</td>
<td>20.75(6.63)</td>
<td>0.66(0.25)</td>
</tr>
<tr>
<td>LASSO ((\tau = 0.50))</td>
<td>19.29(7.66)</td>
<td>3.00(0.00)</td>
<td>16.29(7.66)</td>
<td>0.97(0.42)</td>
</tr>
<tr>
<td>LASSO ((\tau = 0.75))</td>
<td>14.01(6.89)</td>
<td>2.88(0.33)</td>
<td>11.13(6.82)</td>
<td>1.34(0.63)</td>
</tr>
<tr>
<td>SCAD ((\tau = 0.25))</td>
<td>3.56(1.19)</td>
<td>3.00(0.00)</td>
<td>0.56(1.19)</td>
<td>0.12(0.08)</td>
</tr>
<tr>
<td>SCAD ((\tau = 0.50))</td>
<td>3.66(1.36)</td>
<td>2.94(0.24)</td>
<td>0.72(1.36)</td>
<td>0.27(0.22)</td>
</tr>
<tr>
<td>SCAD ((\tau = 0.75))</td>
<td>3.33(1.91)</td>
<td>2.60(0.61)</td>
<td>0.73(1.84)</td>
<td>0.63(0.38)</td>
</tr>
<tr>
<td>MCP ((\tau = 0.25))</td>
<td>3.43(0.83)</td>
<td>3.00(0.00)</td>
<td>0.43(0.83)</td>
<td>0.11(0.07)</td>
</tr>
<tr>
<td>MCP ((\tau = 0.50))</td>
<td>3.70(1.67)</td>
<td>2.94(0.24)</td>
<td>0.76(1.66)</td>
<td>0.28(0.24)</td>
</tr>
<tr>
<td>MCP ((\tau = 0.75))</td>
<td>3.57(2.23)</td>
<td>2.64(0.53)</td>
<td>0.93(2.18)</td>
<td>0.65(0.42)</td>
</tr>
</tbody>
</table>

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 $(x^T\hat{\beta}_{\tau})$, 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 $\hat{q}_{\tau}(x^T\hat{\beta}_{\tau})$ 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,

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

where $\rho_{\tau}(\cdot)$ is the $\tau$th quantile check loss function, $\hat{q}_{\tau}(\cdot)$ is the cubic-spline esti-
Figure 4.2: The estimated curves of $\hat{q}_\tau(x^T \hat{\beta}_\tau)$ (the vertical axis) versus $(x^T \hat{\beta}_\tau)$ (the horizontal axis) at different quantiles and for different penalties. From left to right, $\tau = 0.25, 0.50$ and 0.75; From up to down, LASSO, SCAD and MCP.

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
Table 4.3: Empirical analysis of Cardiomyopathy microarray dataset.

<table>
<thead>
<tr>
<th>Method</th>
<th>All Data</th>
<th>Partitioned Data</th>
<th>PE</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Size</td>
<td>Q-R²</td>
<td>Ave Size</td>
</tr>
<tr>
<td>SCREEN(τ = 0.25)</td>
<td>16</td>
<td>97.6</td>
<td>16.00(0.00)</td>
</tr>
<tr>
<td>SCREEN(τ = 0.50)</td>
<td>16</td>
<td>93.1</td>
<td>16.00(0.00)</td>
</tr>
<tr>
<td>SCREEN(τ = 0.75)</td>
<td>16</td>
<td>94.4</td>
<td>16.00(0.00)</td>
</tr>
<tr>
<td>LASSO(τ = 0.25)</td>
<td>8</td>
<td>87.8</td>
<td>7.21(1.54)</td>
</tr>
<tr>
<td>LASSO(τ = 0.50)</td>
<td>8</td>
<td>89.1</td>
<td>5.64(1.25)</td>
</tr>
<tr>
<td>LASSO(τ = 0.75)</td>
<td>5</td>
<td>91.2</td>
<td>8.29(2.81)</td>
</tr>
<tr>
<td>SCAD(τ = 0.25)</td>
<td>10</td>
<td>96.9</td>
<td>6.52(3.13)</td>
</tr>
<tr>
<td>SCAD(τ = 0.50)</td>
<td>6</td>
<td>92.3</td>
<td>3.82(1.46)</td>
</tr>
<tr>
<td>SCAD(τ = 0.75)</td>
<td>3</td>
<td>93.0</td>
<td>8.69(2.64)</td>
</tr>
<tr>
<td>MCP(τ = 0.25)</td>
<td>10</td>
<td>96.9</td>
<td>6.91(2.81)</td>
</tr>
<tr>
<td>MCP(τ = 0.50)</td>
<td>5</td>
<td>89.3</td>
<td>4.13(1.64)</td>
</tr>
</tbody>
</table>

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-R²”. 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 ultra-high 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 DC-RoSIS. 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 $x$ 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 $n \to \infty$. 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.

**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 Ed-
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), $E \{ I(Y - x^T_A \beta^o_{\tau 1} \leq u^o_{\tau}) \mid x_A \} = \tau$. That is, the oracle estimator $u^o_{\tau}$ of $u_{\tau}$ is the $\tau$th quantile of $Y - x^T_A \beta^o_{\tau 1}$ conditional on $x_A$.

**Proof of Lemma 3.** Let $\xi_{\tau}$ be the $\tau$th quantile of $Y - x^T_A \beta^o_{\tau 1}$ conditional on $x_A$. By definition, we have that $E \{ I(Y - x^T_A \beta^o_{\tau 1} \leq \xi_{\tau}) \mid x_A \} = \tau$. It suffices to show $L_\tau(\xi_{\tau}, \beta^o_{\tau 1}) \leq L_\tau(u, \beta^o_{\tau 1})$ holds for any $u$. To be specific,

\[
L_\tau(u, \beta^o_{\tau 1}) - L_\tau(\xi_{\tau}, \beta^o_{\tau 1}) = E\{\rho_\tau(Y - u - x^T_A \beta^o_{\tau 1})\} - E\{\rho_\tau(Y - \xi_{\tau} - x^T_A \beta^o_{\tau 1})\} \\
= E\left[(u - \xi_{\tau})\{I(Y - \xi_{\tau} - x^T_A \beta^o_{\tau 1} \leq 0) - \tau\}\right] \\
+ E\left[\int_0^{u - \xi_{\tau}} \{I(Y - \xi_{\tau} - x^T_A \beta^o_{\tau 1} \leq t) - I(Y - \xi_{\tau} - x^T_A \beta^o_{\tau 1} \leq 0)\} dt\right] \geq 0,
\]

where the second equality follows from Knight (1998). In the second equality, the first term is zero and the second is nonnegative. Thus $\xi_{\tau} = u^o_{\tau}$ 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 $\Delta_1$ and $\Delta_2$ such that for all sufficiently large $n$,

$$
\Pr \left\{ \inf_{\|\gamma\|=\Delta_1} \mathcal{L}_{\tau n}(u_r^0 + n^{-1/2} q_n^{1/2} u, \beta_{\tau r}^0 + n^{-1/2} q_n^{1/2} \gamma) > \mathcal{L}_{\tau n}(u_r^0, \beta_{\tau r}^0) \right\} \geq 1 - \eta.
$$

We define that

$$
G_n(u, \gamma) := n q_n^{-1} \left\{ \mathcal{L}_{\tau n}(u_r^0 + n^{-1/2} q_n^{1/2} u, \beta_{\tau r}^0 + n^{-1/2} q_n^{1/2} \gamma) - \mathcal{L}_{\tau 0}(u_r^0, \beta_{\tau r}^0) \right\}
$$

$$
= q_n^{-1} \sum_{i=1}^{n} n^{-1/2} q_n^{1/2} (u + x_{i, A}^T \gamma) \left\{ I(Y_i - x_{i, A}^T \beta_{\tau r}^0 \leq u_r^0) - \tau \right\}
$$

$$
+ q_n^{-1} \sum_{i=1}^{n} \int_{0}^{n^{-1/2} q_n^{1/2} (u + x_{i, A}^T \gamma)} \left\{ I(Y_i - x_{i, A}^T \beta_{\tau r}^0 \leq u_r^0 + s) - I(Y_i - x_{i, A}^T \beta_{\tau r}^0 \leq u_r^0) \right\} ds
$$

$$
=: I_{n1} + I_{n2},
$$

where the second equality follows from Knight (1998)'s identity. Note that $E \left\{ I(Y - x_{i, A}^T \beta_{\tau r}^0 \leq u_r^0) \mid x_{i, A} \right\} = \tau$ by Lemma 3 and hence $E(I_{n1}) = 0$.

Next we evaluate $I_{n2}$. Denote by $F(\cdot \mid x_{i, A})$ and $f(\cdot \mid x_{i, A})$ the conditional distribution and density of $(Y - x_{i, A}^T \beta_{\tau r}^0)$ given $x_{i, A}$, respectively

$$
E(I_{n2}) = q_n^{-1} \sum_{i=1}^{n} \int_{0}^{n^{-1/2} q_n^{1/2} (u + x_{i, A}^T \gamma)} \left[ F(u_r^0 + s \mid x_{i, A}) - F(u_r^0 \mid x_{i, A}) \right] ds
$$

$$
= q_n^{-1} \sum_{i=1}^{n} \int_{0}^{n^{-1/2} q_n^{1/2} (u + x_{i, A}^T \gamma)} f(u_r^0 + s \mid x_{i, A}) ds
$$

$$
\geq C q_n^{-1} \sum_{i=1}^{n} \left\{ n^{-1/2} q_n^{1/2} (u + x_{i, A}^T \gamma) \right\}^2
$$

$$
= CE(u + x_{i, A}^T \gamma)^2 \geq C [1 + \lambda_{\min} \left\{ E(x_{i, A}^2) \right\}] (u^2 + \|\gamma\|^2) \geq C(\Delta_1^2 + \Delta_2^2),
$$

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 \( I_{n2} \),

\[
\text{var}(I_{n2}) \leq n q_n^{-2} E \left[ \int_0^{n^{-1/2} q_n^{1/2} (u + x_A) \gamma} \{I(Y - x_A \beta_{r1} \leq u + s) - I(Y - x_A \beta^0_{r1} \leq u)\} ds \right]^2
\]

\[
\leq n q_n^{-2} E(n^{-1/2} q_n^{1/2} (u + x_A) \gamma)^2 \leq q_n^{-1} E(\{E(x_A x_A^T)\}) (u^2 + ||\gamma||^2)
\]

\[
\leq O(q_n^{-1})(\Delta_1^2 + \Delta_2^2),
\]

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 \( \Delta_1 \) and \( \Delta_2 \), \( I_{n2} \) dominates \( I_{n1} \) asymptotically as \( n \to \infty \).

Therefore, for any fixed \( \eta > 0 \), there exists two constants \( \Delta_1 \) and \( \Delta_2 \) 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 \( x \) are different. With slightly notational abuse, we write \( x_A = (1, x_A)^T \), \( \beta^0 = (u^0, \beta^{0T})^T \) defined in (2.4), \( \beta^o = (\hat{u}_\tau, \hat{\beta}_\tau^T) \) and \( \hat{\beta}^o = (\hat{u}_\tau, \hat{\beta}_{1, \tau}^T)^T \), where \( \beta^0 \) denotes the penalized linear quantile estimator defined in (2.3) and \( \hat{\beta}^o = (\hat{\beta}_{1, \tau}, 0^T)^T \) is the oracle estimator defined in (2.5). Accordingly, we write \( \beta^{0T}_{r1} = (u^0, \beta^{0T}_{r1})^T \) and \( \hat{\beta}^o_{r1} = (\hat{u}_\tau, \hat{\beta}_{1, \tau})^T \).

We first write the objective function (2.2) of the penalized linear quantile regression as the difference of two convex functions in \( \beta \). 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 \( g(\beta) = n^{-1} \sum_{i=1}^{n} \rho_\tau(Y_i - x_i^T \beta) + \lambda \sum_{j=1}^{p_n} |\beta_j| \), and \( h(\beta) = \sum_{j=1}^{p_n} H_\lambda(\beta_j) \), with

\[
H_\lambda(\beta_j) = \begin{cases}
null, & 0 \leq |\beta_j| < \lambda; \\
(\beta_j^2 - 2\lambda |\beta_j| + \lambda^2)/(2(a - 1)), & \lambda \leq |\beta_j| \leq a\lambda; \\
\lambda |\beta_j| - (a + 1)\lambda^2/2, & |\beta_j| > a\lambda.
\end{cases}
\]
Thus, the subdifferential of $h(\beta)$ at any $\beta$ is

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

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

$$\partial g(\beta) = \left\{ \xi = (\xi_0, \xi_1, \ldots, \xi_{pn})^T \in \mathbb{R}^{pn+1} : \xi_j = (1-\tau)n^{-1} \sum_{i=1}^{n} X_{ij} I(Y_i - x_i^j \beta < 0) - \tau n^{-1} \sum_{i=1}^{n} X_{ij} v_i + \lambda l_j \right\},$$

where $v_i = 0$ if $Y_i - x_i^j \beta \neq 0$ 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 $1 \leq j \leq pn$.

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

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

where $v_i = 0$ if $Y_i - x_i^j \hat{\beta} \neq 0$ 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 $\hat{\beta}_r^0$ and the associated subgradient functions $s(\hat{\beta}_r^0)$ respectively.

**Lemma 4. (Difference Convex Program)** $g(x)$ and $h(x)$ are two convex functions. Let $x^*$ be a point that admits a neighborhood $U$ such that $\partial h(x) \cap \partial g(x^*) \neq \emptyset$, $\forall x \in U \cap \text{dom}(g)$. Then $x^*$ is a local minimizer of $g(x) - h(x)$.

**Lemma 5.** Assume the conditions (C4)-(C5) holds and $\lambda = o(n^{-(1-c_2)/2})$. For the oracle estimator $\hat{\beta}_r^0$, there exist $v^*_i$ which satisfies $v^*_i = 0$ if $Y_i - x_i^j \hat{\beta}_r^0 \neq 0$ and $v^*_i \in [\tau - 1, \tau]$ otherwise, such that, with probability approaching one, we have

$$s_j(\hat{\beta}_r^0) = 0, j = 0, 1, \ldots, q_n, \text{ and } |\hat{\beta}_j^0| \geq (a + 1/2)\lambda, j = 1, \ldots, q_n.$$
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, $0 \in \partial \sum_{i=1}^{n} \rho_{\tau}(Y_{i} - x_{i}^{T} \hat{\beta}_{\tau}^{0})$. Therefore, there exists $v_{i}^{*}$ such that $s_{j}(\hat{\beta}_{\tau}^{0}) = 0$ with $v_{i} = v_{i}^{*}$ for $j = 0, 1, \ldots, q_{n}$. On the other hand,

$$
\min_{1 \leq j \leq q_{n}} |\hat{\beta}_{j}^{0}| \geq \min_{1 \leq j \leq q_{n}} |\beta_{j}^{0}| - \max_{1 \leq j \leq q_{n}} |\hat{\beta}_{j}^{0} - \beta_{j}^{0}|.
$$

Condition (C5) requires that $\min_{1 \leq j \leq q_{n}} |\beta_{j}^{0}| \geq Cn^{-(1-c_{2})/2}$. In addition, $\max_{1 \leq j \leq q_{n}} |\hat{\beta}_{j}^{0} - \beta_{j}^{0}| \leq \|\hat{\beta}_{\tau}^{0} - \beta_{\tau}^{0}\| = O_{p}(\sqrt{q_{n}/n}) = O_{p}(n^{-(1-c_{1})/2}) = o_{p}(n^{-(1-c_{2})/2})$. Therefore, $\min_{1 \leq j \leq q_{n}} |\hat{\beta}_{j}^{0}| \geq Cn^{-(1-c_{2})/2} - o_{p}(n^{-(1-c_{2})/2})$, 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, $|\hat{\beta}_{j}^{0}| \geq (a+1/2)\lambda, j = 1, \ldots, q_{n}$, 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 $\hat{\beta}_{\tau}^{0}$ and the $s_{j}(\hat{\beta}_{\tau}^{0})$, with probability approaching one, we have

$$
|s_{j}(\hat{\beta}_{\tau}^{0})| \leq \lambda, \quad \text{and} \quad |\hat{\beta}_{j}^{0}| = 0, j = q_{n} + 1, \ldots, p_{n}.
$$

Proof of Lemma 6. This lemma is parallel to Lemma 2.3 in Wang, Wu and Li (2012). Since the $\hat{\beta}_{\tau}^{0}$ is the oracle estimator, $|\hat{\beta}_{j}^{0}| = 0, j = q_{n} + 1, \ldots, p_{n}$. It remains to show that

$$
\Pr\left(|s_{j}(\hat{\beta}_{\tau}^{0})| > \lambda, \text{ for some } j = q_{n} + 1, \ldots, p_{n}\right) \rightarrow 0, \text{ as } n \rightarrow \infty.
$$

Let $D = \{i : Y_{i} - x_{i}^{T} \hat{\beta}_{\tau}^{0} = 0\} = \{i : Y_{i} - x_{i}^{T} \hat{\beta}_{\tau 1}^{0} = 0\}$, then for $j = q_{n} + 1, \ldots, p_{n},$

$$
\begin{align*}
\hat{s}_{j}(\hat{\beta}_{\tau}^{0}) &= (1 - \tau)n^{-1} \sum_{i=1}^{n} X_{ij} I(Y_{i} - x_{i}^{T} \hat{\beta}_{\tau}^{0} < 0) - \tau n^{-1} \sum_{i=1}^{n} X_{ij} I(Y_{i} - x_{i}^{T} \hat{\beta}_{\tau}^{0} > 0) - n^{-1} \sum_{i=1}^{n} X_{ij} v_{i}, \\
&= n^{-1} \sum_{i=1}^{n} X_{ij} \{I(Y_{i} - x_{i}^{T} \hat{\beta}_{\tau}^{0} \leq 0) - \tau\} - n^{-1} \sum_{i=1}^{n} X_{ij} \{v_{i} + (1 - \tau)I(Y_{i} - x_{i}^{T} \hat{\beta}_{\tau}^{0} = 0)\} \\
&= n^{-1} \sum_{i=1}^{n} X_{ij} \{I(Y_{i} - x_{i}^{T} \hat{\beta}_{\tau 1}^{0} \leq 0) - \tau\} - n^{-1} \sum_{i \in D} X_{ij} [v_{i} + (1 - \tau)],
\end{align*}
$$

Let \(D = \{i : Y_i - x_i^T \hat{\beta}_{\tau}^0 = 0\} = \{i : Y_i - x_i^T \hat{\beta}_{\tau 1}^0 = 0\}\), then for \(j = q_n + 1, \ldots, p_n\),

\[
\hat{s}_j(\hat{\beta}_{\tau}^0) = (1 - \tau)n^{-1} \sum_{i=1}^{n} X_{ij} I(Y_i - x_i^T \hat{\beta}_{\tau}^0 < 0) - \tau n^{-1} \sum_{i=1}^{n} X_{ij} I(Y_i - x_i^T \hat{\beta}_{\tau}^0 > 0) - n^{-1} \sum_{i=1}^{n} X_{ij} v_i,
\]

\[
= n^{-1} \sum_{i=1}^{n} X_{ij} \{I(Y_i - x_i^T \hat{\beta}_{\tau}^0 \leq 0) - \tau\} - n^{-1} \sum_{i=1}^{n} X_{ij} \{v_i + (1 - \tau)I(Y_i - x_i^T \hat{\beta}_{\tau}^0 = 0)\} \\
= n^{-1} \sum_{i=1}^{n} X_{ij} \{I(Y_i - x_i^T \hat{\beta}_{\tau 1}^0 \leq 0) - \tau\} - n^{-1} \sum_{i \in D} X_{ij} [v_i + (1 - \tau)],
\]
where \( v_i^* \in [\tau - 1, \tau] \) with \( i \in D \) satisfies \( s_j(\hat{\beta}_0) = 0 \) with \( v_i = v_i^* \), for \( j = 1, \ldots, q_n \) by Lemma 5.

\[
\Pr(|s_j(\hat{\beta}_0)| > 2\lambda, \text{ for some } j = q_n + 1, \ldots, p_n) \\
\leq \Pr \left( \left| \sum_{i=1}^{n} X_{ij} \mathbb{1}_{\{|X_{ij}| \leq M\}} \{v_i^* + (1 - \tau)\} \right| > \lambda/2 \right) \\
+ \Pr \left( \left| \sum_{i \in D} X_{ij} \mathbb{1}_{\{|X_{ij}| > M\}} \{v_i^* + (1 - \tau)\} \right| > \lambda/2 \right) \\
=: T_{n1} + T_{n2}.
\]

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

\[
T_{n2} \leq \Pr \left( \max_{j=q_n+1,\ldots,p_n} \left| \sum_{i \in D} X_{ij} \mathbb{1}_{\{|X_{ij}| \leq M\}} \{v_i^* + (1 - \tau)\} \right| > \lambda/2 \right) \\
+ \Pr \left( \max_{j=q_n+1,\ldots,p_n} \left| \sum_{i \in D} X_{ij} \mathbb{1}_{\{|X_{ij}| > M\}} \{v_i^* + (1 - \tau)\} \right| > \lambda/2 \right) \\
=: T_{n21} + T_{n22}.
\]

Since \((x_{iA}, Y_i)\) 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,

\[
\max_{q_n+1,\ldots,p_n} \left| \sum_{i \in D} X_{ij} \mathbb{1}_{\{|X_{ij}| \leq M\}} \{v_i^* + (1 - \tau)\} \right| \\
\leq M(q_n + 1)n^{-1} = O(n^{\theta+c_1-1}) = o(\lambda),
\]

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 \( n \to \infty \). Next, we deal with \( T_{n22} \). Note that the events satisfy

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

because that if \(|X_{ij}| \leq M\) for all \( i \in D \), then \( n^{-1} \sum_{i \in D} X_{ij} \mathbb{1}_{\{|X_{ij}| > M\}} = 0 \).
Therefore,
\[
T_{n22} \leq p_n \max_{j=q_n+1, \ldots, p_n} \Pr \left( \left| n^{-1} \sum_{i \in D} X_{ij} 1 \{ |X_{ij}| > M \} (\tau_{ij}^* + (1 - \tau)) \right| \geq \lambda/2 \right)
\]
\[
\leq p_n (q_n + 1) \max_{i \in D, q_n + 1 \leq j \leq p_n} \Pr (|X_{ij}| > M) \leq p_n (q_n + 1) \exp(-tM) \exp\left(\exp\left(t|X_{ij}|\right)\right)
\]
\[
\leq Cp_n (q_n + 1) \exp(-tM) = C p_n O(n^{-1}) \exp(-t n^\theta) \rightarrow 0,
\]
as \(n \to \infty\), where \(\log p_n = o(n^{\min\{c_2 - 2\theta, \theta\}}\) with some constant \(0 < \theta < (c_2 - c_1)/2\), \(0 < t \leq t_0\), the third inequality holds from Markov’s inequality and the fourth inequality follows from Condition (C1). Therefore, \(T_{n2} = T_{n12} + T_{n22} \rightarrow 0\), as \(n \to \infty\).

It remains to show that
\[
\Pr \left( \left| n^{-1} \sum_{i=1}^n X_{ij} \left( I(Y_i - \hat{x}_i T^0 \beta_{ij}^* < 0) - \tau \right) \right| > \lambda, \text{ for some } j = q_n + 1, \ldots, p_n \right) \to 0,
\]
as \(n \to \infty\). We consider
\[
T_{n1}
\]
\[
\leq \Pr \left( \max_{j=q_n+1, \ldots, p_n} \left| n^{-1} \sum_{i=1}^n X_{ij} \left( I(Y_i - \hat{x}_i T^0 \beta_{ij}^* < 0) - \tau \right) \right| > \lambda/2 \right)
\]
\[
+ \Pr \left( \max_{j=q_n+1, \ldots, p_n} \sup_{\|\beta_1 - \beta_{ij}^*\| \leq \Delta \sqrt{q_n/n}} \left| n^{-1} \sum_{i=1}^n X_{ij} \left[ I(Y_i - \hat{x}_i ^T \beta_1 \leq 0) - I(Y_i - \hat{x}_i ^T \beta_{ij}^* \leq 0) \right] \right| > \lambda/4 \right)
\]
\[
+ \Pr \left( \max_{j=q_n+1, \ldots, p_n} \sup_{\|\beta_1 - \beta_{ij}^*\| \leq \Delta \sqrt{q_n/n}} \left| n^{-1} \sum_{i=1}^n X_{ij} \left[ \Pr(Y_i - \hat{x}_i ^T \beta_1 \leq 0) - \Pr(Y_i - \hat{x}_i ^T \beta_{ij}^* \leq 0) \right] \right| > \lambda/4 \right) =: J_{n1} + J_{n2} + J_{n3}.
\]
First, let us consider \(J_{n1}\). We choose a \(M = O(n^\theta)\) with \(0 < \theta < (c_2 - c_1)/2\), then
\[
J_{n1}
\]
\[
\leq \Pr \left( \max_{j=q_n+1, \ldots, p_n} \left| n^{-1} \sum_{i=1}^n X_{ij} 1 \{ |X_{ij}| \leq M \} \left( I(Y_i - \hat{x}_i ^T \beta_{ij}^* < 0) - \tau \right) \right| > \lambda/4 \right)
\]
\[
+ \Pr \left( \max_{j=q_n+1, \ldots, p_n} \left| n^{-1} \sum_{i=1}^n X_{ij} 1 \{ |X_{ij}| > M \} \left( I(Y_i - \hat{x}_i ^T \beta_{ij}^* \leq 0) - \tau \right) \right| > \lambda/4 \right)
\]
\[
=: J_{n11} + J_{n12}.
\]
By Hoeffding’s inequality, we have that

\[
\Pr \left( \left| \frac{1}{n} \sum_{i=1}^{n} X_{ij} \right| \leq M \left\{ I(Y_i - x_{i,A}^T \beta_{r1}^o \leq 0) - \tau \right\} > \lambda/4 \right) \leq 2 \exp \left( - \frac{n\lambda^2}{8M^2} \right).
\]

Thus, \( J_{n1} \leq 2p_n \exp \left\{ -n(\lambda/8)^2 \right\} = 2p_n \exp (-n^{1-2\theta}(\lambda/8)^2) \rightarrow 0 \), as \( n \rightarrow \infty \), because \( \log p_n = o(n\min(c_2-2\theta, 0)) \) with some constant \( 0 < \theta < (c_2 - c_1)/2 \) and \( \lambda = o\left(n^{-(1-c_2)/2}\right) \). On the other hand, we can similarly follow the arguments that deal with \( T_{n22} \) and have that

\[
J_{n12} \leq p_n \max_{j=q_n+1, \ldots, p_n} \Pr \left( \left| \frac{1}{n} \sum_{i=1}^{n} X_{ij} \right| > \lambda/4 \right) \leq p_n \max_{1 \leq i \leq n, j=q_n+1, \ldots, p_n} \Pr \left( |X_{ij}| > M \right) = O(p_n n \exp (-tn^\theta)) \rightarrow 0,
\]
as \( n \rightarrow \infty \), because \( \log p_n = o(n\min(c_2-2\theta, 0)) \). 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 \( 0 < \theta < (c_2-c_1)/2 \),

\[
J_{n3} \leq \Pr \left( \max_{j=q_n+1, \ldots, p_n} \sup_{\|\beta_1 - \beta_{11}\| \leq \Delta/\sqrt{q_n/n}} \left| \frac{1}{n} \sum_{i=1}^{n} X_{ij} \left| X_{ij} \right| \leq M \right\} \right.
\]

\[
\left. \left\{ \Pr(Y_i - x_{i,A}^T \beta_{11} \leq 0) - \Pr(Y_i - x_{i,A}^T \beta_{r1}^o \leq 0) \right\} > \lambda/8 \right).
\]

\[
+ \Pr \left( \max_{j=q_n+1, \ldots, p_n} \sup_{\|\beta_1 - \beta_{11}\| \leq \Delta/\sqrt{q_n/n}} \left| \frac{1}{n} \sum_{i=1}^{n} X_{ij} \left| X_{ij} \right| > M \right\} \right.
\]

\[
\left. \left\{ \Pr(Y_i - x_{i,A}^T \beta_{11} \leq 0) - \Pr(Y_i - x_{i,A}^T \beta_{r1}^o \leq 0) \right\} > \lambda/8 \right).
\]

\[
=: J_{n31} + J_{n32}.
\]

To handle \( J_{n31} \), we observe that

\[
\max_{j=q_n+1, \ldots, p_n} \sup_{\|\beta_1 - \beta_{11}\| \leq \Delta/\sqrt{q_n/n}} \left| \frac{1}{n} \sum_{i=1}^{n} X_{ij} \left| X_{ij} \right| > M \right\} \right.
\]

\[
\left. \left\{ \Pr(Y_i - x_{i,A}^T \beta_{11} \leq 0) - \Pr(Y_i - x_{i,A}^T \beta_{r1}^o \leq 0) \right\} \right| \leq M \sup_{\|\beta_1 - \beta_{11}\| \leq \Delta/\sqrt{q_n/n}} \left| \mathbb{E} \left\{ f(\zeta | x_A) x_A^T (\beta_1 - \beta_{11}) \right\} \right|
\]

\[
\leq M \sup_{\|\beta_1 - \beta_{11}\| \leq \Delta/\sqrt{q_n/n}} \lambda_{\max}^{1/2} \left\{ \mathbb{E} (x_A x_A^T) \right\} \| \beta_1 - \beta_{11}^o \|
\]

\[
\leq O\left(n^\theta (q_n/n)^{1/2}\right) = O\left(n^{-\left(1-c_1-2\theta\right)/2}\right),
\]

where \( f(\cdot | x_A) \) is defined in Condition (C3) with \( \zeta \) is between \( u^o + x_A^T (\beta_1 - \beta_{11}) \).
and \( n^2 \) and thus the second inequality follows Condition (C3) and Cauchy-Schwartz inequality, and the third inequality follows Condition (C2). Consequently, together with \( \lambda = o\left(n^{-1/2}\right) \), we have that \( J_{n31} \leq \Pr\left\{ O(n^{-1-2\theta}) > \lambda/8 \right\} = o(1) \) 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,

\[
\Pr\left\{ \max_{q_n+1, \ldots, p_n} \left| n^{-1} \sum_{i=1}^{n} X_{ij} \{I(Y_i - x_i^T \hat{\beta}_1^o < 0) - \tau\} \right| > \lambda \right\} \leq J_{n1} + J_{n2} + J_{n3} = o(1),
\]

which implies that \( \Pr\left\{ s_j(\hat{\beta}_j^o) > \lambda, \text{ for some } j = q_n + 1, \ldots, p_n \right\} \to 0. \) This completes the proof of Lemma 6. \( \square \)

With Lemmas 5 and 6 for the random \( x \) 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. \( \square \)

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, \( \text{dcorr} \{X_k, F(Y)\} = \hat{S}_{k1}^* + \hat{S}_{k2}^* - 2\hat{S}_{k3}^* \), where

\[
\hat{S}_{k1}^* = \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} |X_{ik} - X_{jk}| |F(Y_i) - F(Y_j)|, \\
\hat{S}_{k2}^* = \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} |X_{ik} - X_{jk}| \sum_{i=1}^{n} |F(Y_i) - F(Y_j)|, \text{ and} \\
\hat{S}_{k3}^* = \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} |X_{ik} - X_{jk}| |F(Y_i) - F(Y_j)|.
\]

Similarly we define \( \bar{\omega}_k^* = \text{dcorr} \{X_k, F(Y)\} \). 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

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

To prove Theorem 2, it thus suffices to show the difference between \( \bar{\omega}_k^* \) and \( \bar{\omega}_k \) 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 \( \hat{S}_{k1}^* = \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} |X_{ik} - X_{jk}| |F(Y_i) - F(Y_j)| \)
and $\hat{S}_{k1} = \frac{1}{m} \sum_{i=1}^{n} \sum_{j=1}^{n} |X_{ik} - X_{jk}| |F_n(Y_i) - F_n(Y_j)|$.

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

where $M$ is a positive constant specified later, $A_n = n^{-2} \sum_{i=1}^{n} \sum_{j=1}^{n} (X_{ik} - X_{jk})^2$, 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 the argument that $|x| - |y| \leq |x - y| \leq |x| + |y|$, we obtain that

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

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

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

(C.2)

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),

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

(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 $c_2$,
\[
\Pr \left( \max_{1 \leq k \leq p} |\hat{S}_{k1} - \hat{S}_{k1}| \geq \varepsilon \right) \leq 2 \exp(-n\varepsilon^2/8M^2) + C \exp(-sM) \\
\leq 2 \exp(-c_1 \varepsilon^2 n^{1-2\gamma}) + C \exp(-c_2 n^\gamma). \quad (C.4)
\]
Thus, it is not difficult to show that
\[
\Pr \left( \max_{1 \leq k \leq p} |\hat{\omega}_k - \hat{\omega}_k| \geq c n^{-n} \right) \leq O \left( \exp \left\{ -c_1 n^{1-2(\kappa + \gamma)} \right\} + \exp \left\{ -c_2 n^\gamma \right\} \right). \quad (C.5)
\]
Therefore, (C.1) and (C.5) together completes the proof of Theorem 2.

References


Wei Zhong
Wang Yanan Institute for Studies in Economics,
Department of Statistics and Fujian Key Laboratory of Statistical Science,
Xiamen University, Xiamen 361005, China.
E-mail: 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.
E-mail: zhu.liping@mail.shufe.edu.cn

Runze Li
Department of Statistics and The Methodology Center,
The Pennsylvania State University, University Park, PA 16802.
E-mail: rzli@psu.edu

Hengjian Cui
School of Mathematical Science,
Capital Normal University, Beijing 100048, China.
E-mail: hjcui@bnu.edu.cn