Biometrika (2015), **102**, 1, *pp.* 65–76
Printed in Great Britain
doi: 10.1093/biomet/asu068
Advance Access publication 16 February 2015

Conditional quantile screening in ultrahigh-dimensional heterogeneous data

BY YUANSHAN WU

School of Mathematics and Statistics, Wuhan University, Wuhan, Hubei 430072, China
shan@whu.edu.cn

AND GUOSHENG YIN

Department of Statistics and Actuarial Science, University of Hong Kong, Pokfulam Road, Hong Kong
gyin@hku.hk

SUMMARY

To accommodate the heterogeneity that is often present in ultrahigh-dimensional data, we propose a conditional quantile screening method, which enables us to select features that contribute to the conditional quantile of the response given the covariates. The method can naturally handle censored data by incorporating a weighting scheme through redistribution of the mass to the right; moreover, it is invariant to monotone transformation of the response and requires substantially weaker conditions than do alternative methods. We establish sure independent screening properties for both the complete and the censored response cases. We also conduct simulations to evaluate the finite-sample performance of the proposed method, and compare it with existing approaches.

Some key words: Censored data; Feature screening; Heterogeneity; Quantile regression; Sure independent screening; Transformation model; Ultrahigh dimensionality; Variable selection.

1. INTRODUCTION

Ultrahigh-dimensional data arise in fields such as genomics, imaging and economics. Because the dimensionality p_n of the covariates $Z = (Z_1, \ldots, Z_{p_n})^{\mathrm{T}}$ increases very rapidly with the sample size n, existing penalized variable selection methods (Tibshirani, 1996; Fan & Li, 2001; Zhang, 2010) may not perform well (Fan et al., 2009). To overcome the difficulties associated with ultrahigh dimensionality, Fan & Lv (2008) proposed a sure independent screening method to reduce the dimension, so that penalized variable selection procedures would be applicable. Such screening procedures have been studied extensively in various ultrahigh-dimensional contexts, such as generalized linear models (Fan & Song, 2010), additive models (Fan et al., 2011), Cox proportional hazards regression with survival data (Zhao & Li, 2012), and additive hazard models (Gorst-Rasmussen & Scheike, 2013). Furthermore, Zhu et al. (2011) proposed a sure independent ranking and screening procedure for ultrahigh-dimensional multi-index models, and Li et al. (2012b) developed an alternative approach based on the distance correlation.

To identify active predictors from p_n covariates, we define the active predictor set as

$$\mathcal{A} = \{k : F(y \mid Z) \text{ depends on } Z_k,\ k = 1, \ldots, p_n\},$$

where $F(y \mid Z) = \mathrm{pr}(Y \leqslant y \mid Z)$. Instead of $F(y \mid Z)$, if we focus on the τth conditional quantile, which is defined as $Q_\tau(Y \mid Z) = \inf\{t : \mathrm{pr}(Y \leqslant t \mid Z) \geqslant \tau\}$ for a fixed $\tau \in (0, 1)$, the active predictor set corresponding to a particular τ is defined as

$$\mathcal{A}_\tau = \{k : Q_\tau(Y \mid Z) \text{ depends on } Z_k,\ k = 1, \ldots, p_n\}.$$

Obviously, $\mathcal{A}_\tau \subset \mathcal{A}$, because $\mathcal{A}_\tau$ identifies only those predictors that contribute to the τth conditional quantile of Y, whereas $\mathcal{A}$ contains all the predictors associated with the conditional distribution of Y. Regression quantile estimation is known to be able to handle heterogeneity (Koenker, 2005), which is often a feature of ultrahigh-dimensional data.

We consider a nonparametric regression model with heterogeneous errors,

$$Y = q(Z) + \sigma(Z)\epsilon,$$

where $q(\cdot)$ and $\sigma(\cdot) > 0$ are unspecified functions and ϵ is assumed to be independent of Z. The τth quantile is then $Q_\tau(Y \mid Z) = q(Z) + \sigma(Z)Q_\tau(\epsilon)$, where $Q_\tau(\epsilon) = \inf\{t : \mathrm{pr}(\epsilon \leqslant t) \geqslant \tau\}$. Clearly, all the predictors in $q(\cdot)$ belong to $\mathcal{A}_\tau$, while those in $\sigma(\cdot)$ would belong to $\mathcal{A}_\tau$ if and only if $Q_\tau(\epsilon) \neq 0$. For a particular τth-quantile model satisfying $Q_\tau(\epsilon) = 0$, the covariates in $\sigma(\cdot)$ alone should not be selected. He et al. (2013) proposed a quantile adaptive sure independent screening procedure to estimate $\mathcal{A}_\tau$, which employs spline approximation to model the marginal covariate effects and the inverse probability weighting scheme to accommodate censored data. To identify $\mathcal{A}_\tau$, Shao & Zhang (2014) developed a martingale difference correlation screening procedure based on the distance correlation.

We propose a conditional quantile sure independent screening procedure to estimate $\mathcal{A}_\tau$. The proposed method can also handle ultrahigh-dimensional survival data by employing redistribution-of-mass weights for censored observations. Following the suggestion of a referee, we further extend the conditional quantile screening procedure to a new model-free approach to recovering the entire active set $\mathcal{A}$. Compared with existing methods, our approach enjoys several advantages. First, it does not involve any nonparametric approximation, which facilitates its practical implementation. With censored data, the Kaplan–Meier estimator is used, which is also straightforward. Second, the features selected using the proposed procedure are invariant to monotone transformation of the response, whereas those in He et al. (2013) are not. Third, our approach does not require any finite-moment assumption, although it achieves a higher exponential rate for the dimensionality with respect to the sample size. Theoretical proofs and some of the detailed numerical results are given in the Supplementary Material.

2. SCREENING PROCEDURES

2·1. *Conditional quantile screening for a complete response*

Suppose that for an independent and identically distributed sample $\{Y_i,\ (Z_{i1}, \ldots, Z_{ip_n})^{\mathrm{T}} : i = 1, \ldots, n\}$, the dimensionality p_n greatly exceeds the sample size n. Our goal is to identify a set of active covariates associated with the response, for which we propose a marginal utility to rank the covariates. If $Q_\tau(Y \mid Z_k) = Q_\tau(Y)$ $(k = 1, \ldots, p_n)$, then

$$E[\tau - I\{Y < Q_\tau(Y \mid Z_k)\} \mid Z_k] = E[\tau - I\{Y < Q_\tau(Y)\} \mid Z_k] = 0.$$

Motivated by the definition of conditional expectation, we define

$$d_k(t) = E\big([\tau - I\{Y < Q_\tau(Y)\}]\, I(Z_k < t)\big);$$

it is easy to see that $d_k(t) = 0$ for any $t \in \mathbb{R}$. The empirical counterpart of $d_k(t)$ is

$$\hat{d}_k(t) = n^{-1} \sum_{i=1}^{n} \left[\tau - I\{Y_i < \hat{Q}_\tau(Y)\}\right] I(Z_{ik} < t), \tag{1}$$

where $\hat{Q}_\tau(Y)$ is the estimate of the τth quantile based on $Y_1, \ldots, Y_n$. If the τth conditional quantile of Y given Z_k does not depend on Z_k, then $\hat{d}_k(t)$ is expected to fluctuate around zero. There is an analogy between (1) and the goodness-of-fit test based on the cumulative sum of residuals (Lin et al., 1993; He & Zhu, 2003).

Based on this rationale, we construct the marginal utility for the kth predictor,

$$\|\hat{d}_k\| = n^{-1} \sum_{i=1}^{n} \hat{d}_k(Z_{ik})^2.$$

Those predictors with a large value of $\|\hat{d}_k\|$ are considered important. We define the estimated active set as

$$\hat{\mathcal{A}}_\tau = \{k : \|\hat{d}_k\| \geqslant cn^{-\alpha},\ k = 1, \ldots, p_n\},$$

where the prespecified threshold values c and $\alpha \in [0, 1/2)$ are given in the following regularity condition.

Condition 1. For some constant $c > 0$, $\min_{k \in \mathcal{A}_\tau} \|d_k\| \geqslant 2cn^{-\alpha}$, where $\|d_k\| = E\{d_k(Z_k)^2\}$ is the population counterpart of $\|\hat{d}_k\|$.

The marginal utility $\|\hat{d}_k\|$ is invariant under any monotone transformation of the response. Furthermore, the proposed method can deal with transformation models without needing to directly involve nonparametric estimation (Li et al., 2012a).

PROPOSITION 1. *Let* $Z_{\mathcal{A}_\tau} = \{Z_j : j \in \mathcal{A}_\tau\}$ *and* $Z_{\mathcal{A}_\tau^c} = \{Z_j : j \notin \mathcal{A}_\tau\}$, *and assume that:*

(i) $I\{Y < Q_\tau(Y)\}$ *and* $Z_{\mathcal{A}_\tau^c}$ *are conditionally independent given* $Z_{\mathcal{A}_\tau}$; *and*
(ii) $Z_{\mathcal{A}_\tau}$ *is independent of* $Z_{\mathcal{A}_\tau^c}$.

Then, under Condition 1,

$$\max_{k \notin \mathcal{A}_\tau} \|d_k\| < \min_{k \in \mathcal{A}_\tau} \|d_k\|,$$

with $\|d_k\| = 0$ *if and only if* $k \notin \mathcal{A}_\tau$.

This result implies that $\|\hat{d}_k\|$ is useful for feature screening, as it tends to rank important covariates over unimportant ones. Due to the fact that $Q_\tau(Y \mid Z) = Q_\tau(Y \mid Z_{\mathcal{A}_\tau})$, the assumption (i) holds if $\mathcal{A}_\tau = \mathcal{A}$, and the assumption (ii) is the partial orthogonality condition (Huang et al., 2008).

Shao & Zhang (2014) ranked the distance correlation between $\tau - I\{Y < Q_\tau(Y)\}$ and Z_k, so their screening method is also invariant under any monotone transformation. However, since our method is based on the indicator $I(Z_k < t)$, it does not require a finite-moment assumption for each Z_k, and this yields more robustness with respect to heavy-tailed distributions. Like all existing sure independent screening procedures based on marginal utilities, our method suffers in situations where the predictors are jointly but not marginally important.

2·2. *Conditional quantile screening for a censored response*

To accommodate censoring, we extend the screening procedure to ultrahigh-dimensional survival data. Suppose that we observe the data $\{X_i, \Delta_i, (Z_{i1}, \ldots, Z_{ip_n})^{\mathrm{T}} : i = 1, \ldots, n\}$, consisting of independent copies of (X, Δ, Z) where $X = \min(Y, C)$ and $\Delta = I(Y \leqslant C)$, with C representing the censoring variable. For ease of exposition, we assume that the censoring distribution is independent of covariates. For censored conditional quantile screening, we propose a weight-adjusted version of $d_k(t)$,

$$r_k(t) = E\big\{[\tau - \omega(F) I\{X < Q_\tau(Y)\}]\, I(Z_k < t)\big\},$$

where $F(y) = \mathrm{pr}(Y \leqslant y)$ and the weight function

$$\omega(F) = \begin{cases} 1, & \Delta = 1 \text{ or } F(C) > \tau, \\ \dfrac{\tau - F(C)}{1 - F(C)}, & \Delta = 0,\ F(C) \leqslant \tau \end{cases}$$

redistributes the masses of censored observations to the right (Portnoy, 2003; Wang & Wang, 2009). For any $t \in \mathbb{R}$, if the τth conditional quantile of Y given Z_k does not depend on Z_k, then $r_k(t) = 0$. Let $\hat{F}_n(y) = 1 - \hat{S}_n(y)$, where $\hat{S}_n(y)$ is the Kaplan–Meier estimator of Y based on $\{(X_i, \Delta_i) : i = 1, \ldots, n\}$. The τth sample quantile $\hat{F}_n^{-1}(\tau)$ is an estimator of $Q_\tau(Y)$ when Y is subject to right censoring. Likewise, we define the empirical version of $r_k(t)$ as

$$\hat{r}_k(t) = n^{-1} \sum_{i=1}^{n} \big[\tau - \omega_i(\hat{F}_n) I\{X_i < \hat{F}_n^{-1}(\tau)\}\big]\, I(Z_{ik} < t).$$

We expect $\hat{r}_k(t)$ to be close to zero if the τth conditional quantile of Y given Z_k does not depend on Z_k. We define $\|\hat{r}_k\| = n^{-1} \sum_{i=1}^{n} \hat{r}_k(Z_{ik})^2$ and then select a set of active variables

$$\hat{\mathcal{A}}_\tau^* = \{k : \|\hat{r}_k\| \geqslant c^* n^{-\alpha},\ k = 1, \ldots, p_n\},$$

where c^* is a prespecified threshold value, given in Condition 5 below.

2·3. *Conditional distribution function screening*

The proposed conditional quantile utility can be extended to recover the whole active predictor set $\mathcal{A}$, which contains all the predictors associated with the conditional distribution of Y. Define

$$h_k(y, t) = E\big[\{F(y) - I(Y \leqslant y)\} I(Z_k < t)\big]$$

and its empirical version

$$\hat{h}_k(y, t) = n^{-1} \sum_{i=1}^{n} \left\{ n^{-1} \sum_{j=1}^{n} I(Y_j \leqslant y) - I(Y_i \leqslant y) \right\} I(Z_{ik} < t).$$

We propose a model-free screening procedure based on $\|\hat{h}_k\| = n^{-1} \sum_{i=1}^{n} \{h_k(Y_i, Z_{ik})\}^2$, and thus the set of active variables can be identified as

$$\hat{\mathcal{A}} = \{k : \|\hat{h}_k\| \geqslant \tilde{c} n^{-\alpha},\ k = 1, \ldots, p_n\},$$

where $\tilde{c}$ is the constant in Condition 6 below.

In fact, $h_k(y, t) = -\mathrm{cov}\{I(Y \leqslant y), I(Z_k < t)\}$, which is also the basis for the statistic used in Heller et al. (2013) to test independence of two variables. Similarly, our model-free screening method has distinctive features: invariance under any monotone transformation and no requirement of finite moments. By introducing appropriate weights, the model-free screening utility can also be modified to handle censoring. In particular, the screening utility for the kth predictor can be constructed by cumulatively summing $\hat{S}_n(y) - \Delta_i I(X_i > y)/\hat{R}_n(X_i)$ over Z_{ik} from $i = 1$ to $i = n$, where $\hat{R}_n(y)$ is the Kaplan–Meier estimator of the censoring time.

3. Theoretical properties

In addition to Condition 1, we introduce further regularity conditions.

Condition 2. In a neighbourhood of $Q_\tau(Y)$, $F(y)$ is twice differentiable; the density function of Y, $f(y)$, is uniformly bounded away from zero and infinity, and its derivative $f'(y)$ is bounded uniformly.

Condition 3. In a neighbourhood of $Q_\tau(Y)$, $G(x) = \mathrm{pr}(C \leqslant x)$ is twice differentiable; the density function of C, $g(x)$, is uniformly bounded away from zero and infinity, and its derivative $g'(x)$ is bounded uniformly.

Condition 4. Let L be the end time of the study; then τ satisfies $0 < \tau < \mathrm{pr}(Y \leqslant L)$.

Condition 5. For a constant $c^* > 0$, $\min_{k \in \mathcal{A}_\tau} \|r_k\| \geqslant 2c^* n^{-\alpha}$, where $\|r_k\| = E\{r_k(Z_k)^2\}$.

Condition 6. For a constant $\tilde{c} > 0$, $\min_{k \in \mathcal{A}} \|h_k\| \geqslant 2\tilde{c} n^{-\alpha}$, where $\|h_k\| = E\{h_k(Y, Z_k)^2\}$.

Conditions 2 and 3 are standard in censored quantile regression. Condition 4 ensures estimability of the τth regression quantile. Conditions 1 and 5, for complete and censored responses, respectively, require the marginal utilities carrying information for the features in the active set not to decay too fast, and Condition 6 corresponds to the model-free screening. Compared with the finite exponential moment conditions C3 in Zhu et al. (2011), C1 in Li et al. (2012b) and B2 in Shao & Zhang (2014), as well as the bounded support condition C1 in He et al. (2013), we impose no conditions on the predictor Z, so our screening procedures are more robust with respect to heavy-tailed predictors.

The sure independent screening properties of our procedures can be stated as follows.

Theorem 1. *Under Condition* 2*, there exist positive constants* c_1 *and* c_2 *such that*

$$\mathrm{pr}\left(\max_{1 \leqslant k \leqslant p_n} \left| \|\hat{d}_k\| - \|d_k\| \right| \geqslant cn^{-\alpha} \right) \leqslant O\left\{ p_n \exp(-c_1 n^{1-2\alpha}) + p_n \exp(-c_2 n^{3-2\alpha}) \right\},$$

and under Conditions 1 *and* 2*,*

$$\mathrm{pr}(\mathcal{A}_\tau \subseteq \hat{\mathcal{A}}_\tau) \geqslant 1 - O\left\{ a_n \exp(-c_1 n^{1-2\alpha}) + a_n \exp(-c_2 n^{3-2\alpha}) \right\},$$

where $a_n = |\mathcal{A}_\tau|$ *is the cardinality of* $\mathcal{A}_\tau$*.*

Let $(\tilde{Y}, \tilde{Z})$ be an independent copy of (Y, Z), and define $V = [\tau - I\{\tilde{Y} < Q_\tau(\tilde{Y})\}] I(\tilde{Z} < Z)$, where $I(\tilde{Z} < Z) = \{I(\tilde{Z}_1 < Z_1), \ldots, I(\tilde{Z}_{p_n} < Z_{p_n})\}^{\mathrm{T}}$. Let $\|\cdot\|_{\mathrm{E}}$ denote the Euclidean norm. The next result says that if $E(\|V\|_{\mathrm{E}}^2) = O(n^\gamma)$ for some $\gamma > 0$, the model after screening is of polynomial size with probability approaching 1.

THEOREM 2. *Under Conditions* 1 *and* 2*, there exist positive constants* b_1 *and* b_2 *such that*

$$\mathrm{pr}\{|\hat{\mathcal{A}}_\tau| \leqslant 2c^{-1}n^{\alpha}E(\|V\|_{\mathrm{E}}^2)\} \geqslant 1 - O\{p_n \exp(-b_1 n^{1-2\alpha}) + p_n \exp(-b_2 n^{3-2\alpha})\}.$$

We can also establish the sure independent screening properties for ultrahigh-dimensional survival data with censoring.

THEOREM 3. *Under Conditions* 2–4*, there exist positive constants* c_1^* *and* c_2^* *such that*

$$\mathrm{pr}\left(\max_{1\leqslant k\leqslant p_n}\left|\|\hat{r}_k\| - \|r_k\|\right| \geqslant c^* n^{-\alpha}\right) \leqslant O\{p_n \exp(-c_1^* n^{1-2\alpha}) + p_n \exp(-c_2^* n^{3-2\alpha})\},$$

and under Conditions 2–5*,*

$$\mathrm{pr}(\mathcal{A}_\tau \subseteq \hat{\mathcal{A}}_\tau^*) \geqslant 1 - O\{a_n \exp(-c_1^* n^{1-2\alpha}) + a_n \exp(-c_2^* n^{3-2\alpha})\}.$$

Theorems 1 and 3 imply that our screening procedures can handle nonpolynomial dimensionality of order $\log p_n = o(n^{1-2\alpha})$ with $\alpha \in [0, 1/2)$ for both complete and censored data. Compared with Zhu et al. (2011), Li et al. (2012b), He et al. (2013) and Shao & Zhang (2014), we can achieve a higher exponential rate for the dimensionality under weaker conditions on the predictors, mainly due to the use of indicator functions in our screening utility. Let $(\tilde{Y}, \tilde{C}, \tilde{Z})$ be an independent copy of (Y, C, Z), and write $\tilde{X} = \min(\tilde{Y}, \tilde{C})$ and $\tilde{\Delta} = I(\tilde{Y} \leqslant \tilde{C})$.

THEOREM 4. *Under Conditions* 2–5*, there exist positive constants* b_1^* *and* b_2^* *such that*

$$\mathrm{pr}\{|\hat{\mathcal{A}}_\tau^*| \leqslant 2c^{*-1}n^{\alpha}E(\|V^*\|_{\mathrm{E}}^2)\} \geqslant 1 - O\{p_n \exp(-b_1^* n^{1-2\alpha}) + p_n \exp(-b_2^* n^{3-2\alpha})\},$$

where $V^* = [\tau - \tilde{\omega}(F)I\{\tilde{X} < Q_\tau(\tilde{Y})\}]I(\tilde{Z} < Z)$ *with*

$$\tilde{\omega}(F) = \begin{cases} 1, & \tilde{\Delta} = 1 \text{ or } F(\tilde{C}) > \tau, \\ \dfrac{\tau - F(\tilde{C})}{1 - F(\tilde{C})}, & \tilde{\Delta} = 0,\ F(\tilde{C}) \leqslant \tau. \end{cases} \tag{2}$$

This suggests that with censored data, if $E(\|V^*\|_{\mathrm{E}}^2) = O(n^{\gamma})$ for some $\gamma > 0$, the model after screening is of polynomial size with probability approaching 1. For the model-free conditional distribution screening procedure, we can establish similar sure independent screening properties.

THEOREM 5. *Under Condition* 2*, there exist positive constants* $\tilde{c}_1$ *and* $\tilde{c}_2$ *such that*

$$\mathrm{pr}\left(\max_{1\leqslant k\leqslant p_n}\left|\|\hat{h}_k\| - \|h_k\|\right| \geqslant \tilde{c} n^{-\alpha}\right) \leqslant O\{p_n \exp(-\tilde{c}_1 n^{1-2\alpha}) + p_n \exp(-\tilde{c}_2 n^{3-2\alpha})\},$$

and under Conditions 2 *and* 6*,*

$$\mathrm{pr}(\mathcal{A} \subseteq \hat{\mathcal{A}}) \geqslant 1 - O\{\tilde{a}_n \exp(-\tilde{c}_1 n^{1-2\alpha}) + \tilde{a}_n \exp(-\tilde{c}_2 n^{3-2\alpha})\},$$

where $\tilde{a}_n = |\mathcal{A}|$.

The proposed model-free screening method can also handle nonpolynomial dimensionality without finite-moment assumptions. Let $(\tilde{Y}, \tilde{Z})$ be an independent copy of (Y, Z), and write $\tilde{V} = \{F(Y) - I(\tilde{Y} \leqslant Y)\}I(\tilde{Z} < Z)$.

THEOREM 6. *Under Conditions* 2 *and* 6, *there exist positive constants* $\tilde{b}_1$ *and* $\tilde{b}_2$ *such that*

$$\mathrm{pr}\{|\hat{\mathcal{A}}| \leqslant 2\tilde{c}^{-1}n^{\alpha}E(\|\tilde{V}\|_{\mathrm{E}}^2)\} \geqslant 1 - O\{p_n \exp(-\tilde{b}_1 n^{1-2\alpha}) + p_n \exp(-\tilde{b}_2 n^{3-2\alpha})\}.$$

If $E(\|\tilde{V}\|_{\mathrm{E}}^2) = O(n^{\gamma})$ for some $\gamma > 0$, the model based on the conditional distribution screening is of polynomial size with probability approaching 1.

4. SIMULATION STUDIES

We examine the finite-sample performance of the proposed methods and compare them with existing methods via simulation studies.

Example 1. We consider the following model adapted from Zhu et al. (2011), He et al. (2013) and Shao & Zhang (2014):

$$Y = Z_1 + 0{\cdot}8Z_2 + 0{\cdot}6Z_3 + 0{\cdot}4Z_4 + 0{\cdot}2Z_5 + \sigma(Z)\epsilon, \tag{3}$$

where the ultrahigh-dimensional covariates $Z = (Z_1, \ldots, Z_{p_n})^{\mathrm{T}}$ follow a multivariate normal distribution with mean zero and correlation matrix $\Sigma = (0{\cdot}8^{|i-j|})$ $(i, j = 1, \ldots, p_n)$. We considered the sample sizes $n = 100$ and 200, and set the number of covariates p_n to 2000. For the error term, we set $\sigma(Z) = \exp(Z_{20} + Z_{21} + Z_{22})$ and generated ϵ from the standard normal or standard Cauchy distribution. We took the censoring time C to be $\min(\tilde{C}, L)$, where $\tilde{C}$ was generated from $\mathrm{Un}(1, L + 2)$ with L being the study duration time, which was chosen to yield a censoring rate of 25%. If we take the τth quantile of model (3), we have

$$Q_{\tau}(Y \mid Z) = Z_1 + 0{\cdot}8Z_2 + 0{\cdot}6Z_3 + 0{\cdot}4Z_4 + 0{\cdot}2Z_5 + \sigma(Z)Q_{\tau}(\epsilon).$$

Considering two quantile levels $\tau = 0{\cdot}5$ and $\tau = 0{\cdot}75$, the sizes of the true active predictor sets $\mathcal{A}_{0{\cdot}5}$ and $\mathcal{A}_{0{\cdot}75}$ are $p_0 = 5$ and 8, respectively. We chose the model size to be $\lceil n/\log n \rceil$. For each configuration, we replicated 500 simulations.

To assess the performance of the screening procedures, we employed the evaluation criteria in Li et al. (2012b). First, we compare the minimum model size $\mathcal{S}$, which is the smallest number of covariates needed to include all the active predictors. Obviously, $\mathcal{S}$ can be used to measure the resulting model complexity for each screening procedure. The closer it is to the true minimum model size, the better the screening procedure. We present the median and interquartile range of $\mathcal{S}$ over 500 replications. The second criterion is the proportion, out of the 500 replications, that all of the active predictors are selected for a given model size; we denote this proportion by $\mathcal{P}_{\mathrm{All}}$. An effective screening procedure is expected to yield $\mathcal{P}_{\mathrm{All}}$ close to 1.

Table 1 shows that our distribution function screening method performs substantially better than both the sure independent ranking and screening approach of Zhu et al. (2011) and the distance correlation screening method of Li et al. (2012b). For conditional quantile screening at $\tau = 0{\cdot}5$ or $0{\cdot}75$, our quantile screening method and the martingale difference correlation quantile screening approach of Shao & Zhang (2014) perform comparably, and both are superior to the quantile adaptive screening procedure of He et al. (2013). For censored data, our quantile screening method delivers better results than the procedure of He et al. (2013) at $\tau = 0{\cdot}5$, whereas the opposite is true at $\tau = 0{\cdot}75$. The performances of all the screening procedures are enhanced as n is increased to 200.

Figure 1 plots $\hat{d}_k(t)$ and $\hat{r}_k(t)$ against t, with $k = 2$ corresponding to an active predictor, $k = 8$ to a nonactive predictor, and $k = 20$ to a nonactive predictor at $\tau = 0{\cdot}5$ but an active one at $\tau = 0{\cdot}75$. Clearly, for active predictors Z_2 and Z_{20} at $\tau = 0{\cdot}75$, the curves of $\hat{d}_k(t)$ deviate

Table 1. *Simulation results for Example* 1 *with true model size* p_0*: reported are the median and interquartile range of the minimum model size needed to include all active predictors, along with the proportion* $\mathcal{P}_{\text{All}}$ (%) *that all of the active predictors are selected for a given model size*

				$n = 100$			$n = 200$		
Error	τ	Method	p_0	Median	IQR	$\mathcal{P}_{\text{All}}$	Median	IQR	$\mathcal{P}_{\text{All}}$
Normal		DF-SIS	8	28	38	41·0	10	2	99·6
		SIRS	8	36	53	26·4	10	3	99·6
		DC-SIS	8	360	628	8·2	126	358	30·6
	0·5	Q-SIS	5	6	5	86·0	5	0	100·0
		QaSIS	5	28	47	39·6	7	3	99·8
		MDC-SISQ	5	6	4	87·8	5	0	100·0
		Q-SIS cens.	5	8	18	74·6	5	0	99·6
		QaSIS cens.	5	132	276	6·8	20	25	74·6
	0·75	Q-SIS	8	245	680	4·8	26	73	57·4
		QaSIS	8	270	357	0·0	50	62	38·6
		MDC-SISQ	8	212	657	6·4	24	64	61·6
		Q-SIS cens.	8	516	1020	1·6	66	244	40·4
		QaSIS cens.	8	320	956	9·2	14	73	67·8
Cauchy		DF-SIS	8	25	33	42·0	9	2	98·6
		SIRS	8	34	38	28·6	9	3	99·0
		DC-SIS	8	1017	802	0·0	828	755	0·8
	0·5	Q-SIS	5	8	18	73·0	5	0	98·6
		QaSIS	5	77	119	8·2	10	6	94·6
		MDC-SISQ	5	7	17	74·0	5	0	98·8
		Q-SIS cens.	5	16	50	57·2	5	1	97·4
		QaSIS cens.	5	292	456	0·6	50	118	38·6
	0·75	Q-SIS	8	410	744	1·6	52	130	40·2
		QaSIS	8	598	605	0·0	192	248	3·6
		MDC-SISQ	8	379	741	2·2	44	120	45·0
		Q-SIS cens.	8	622	1041	0·4	71	201	34·0
		QaSIS cens.	8	272	662	4·8	21	98	62·0

IQR, interquartile range of the minimum model size $\mathcal{S}$; DF-SIS, proposed conditional distribution function sure independent screening approach; SIRS, sure independent ranking and screening of Zhu et al. (2011); DC-SIS, distance correlation sure independent screening of Li et al. (2012b); Q-SIS, proposed conditional quantile sure independent screening approach; QaSIS, quantile adaptive sure independent screening of He et al. (2013); MDC-SISQ, martingale difference correlation sure independent screening of Shao & Zhang (2014); cens., censored data case.

substantially from the zero axis. For nonactive predictors Z_8 and Z_{20} at $\tau = 0{\cdot}5$, $\hat{d}_8(t)$ and $\hat{d}_{20}(t)$ fluctuate around the zero axis. Similar phenomena can be observed for the censored version $\hat{r}_k(t)$. These plots demonstrate the effectiveness of the proposed conditional quantile screening procedures.

Example 2. To examine the nonlinear scenario, we consider a model with interactions,

$$Y = Z_1^2 \sin(Z_2) + Z_3^3 + \cos^2(Z_4) + \sigma(Z)\epsilon,$$

while keeping the rest of the set-up the same as in Example 1. The simulation results are summarized in Table 2, from which we can draw similar conclusions to Example 1.

Under a given model size $\lceil n/\log n \rceil$, the selection proportions for each of Z_{20}, Z_{21} and Z_{20} at $\tau = 0{\cdot}5$ are reported in Table 3. These predictors do not belong to $\mathcal{A}_{0{\cdot}5}$, the active set at $\tau = 0{\cdot}5$, in

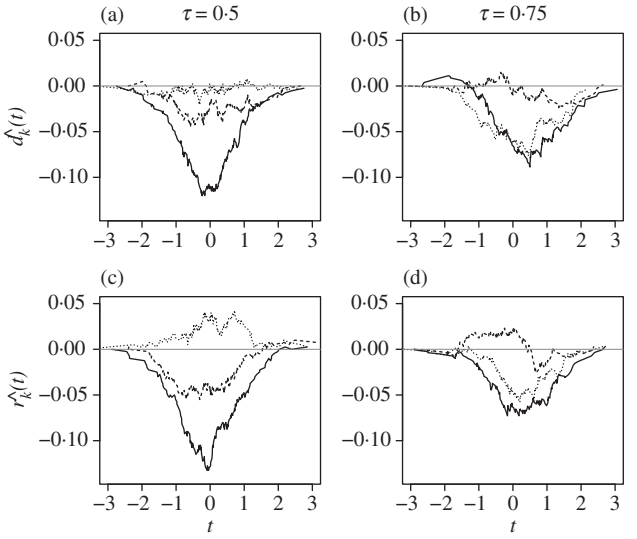

Fig. 1. Plots of $\hat{d}_k(t)$ with complete response and $\hat{r}_k(t)$ with censored response for $k = 2$ (solid), $k = 8$ (dashed) and $k = 20$ (dotted), based on one simulated dataset in Example 1 with heterogeneous normal errors and $n = 200$: (a) $\hat{d}_k(t)$ at $\tau = 0{\cdot}5$; (b) $\hat{d}_k(t)$ at $\tau = 0{\cdot}75$; (c) $\hat{r}_k(t)$ at $\tau = 0{\cdot}5$; (d) $\hat{d}_k(t)$ at $\tau = 0{\cdot}75$.

Table 2. *Simulation results for Example* 2 *with true model size* p_0*: reported are the median and interquartile range of the minimum model size needed to include all active predictors, along with the proportion* $\mathcal{P}_{\mathrm{All}}$ (%) *that all of the active predictors are selected for a given model size*

				$n = 100$			$n = 200$		
Error	τ	Method	p_0	Median	IQR	$\mathcal{P}_{\mathrm{All}}$	Median	IQR	$\mathcal{P}_{\mathrm{All}}$
Normal		DF-SIS	7	22	35	47·6	9	2	99·2
		SIRS	7	32	50	31·4	9	4	98·8
		DC-SIS	7	230	554	14·2	64	246	40·2
	0·5	Q-SIS	4	5	8	85·4	4	0	100·0
		QaSIS	4	14	17	69·4	6	2	100·0
		MDC-SISQ	4	5	5	88·8	4	0	100·0
		Q-SIS cens.	4	7	16	75·8	4	0	99·8
		QaSIS cens.	4	110	192	6·0	21	22	76·2
	0·75	Q-SIS	7	167	440	9·0	17	26	75·2
		QaSIS	7	284	330	0·0	49	56	37·6
		MDC-SISQ	7	146	410	10·6	15	24	76·6
		Q-SIS cens.	7	406	1009	2·6	40	114	49·2
		QaSIS cens.	7	78	575	21·8	10	10	84·8
Cauchy		DF-SIS	7	24	33	45·0	8	2	99·6
		SIRS	7	32	39	35·0	8	2	99·6
		DC-SIS	7	932	889	0·6	678	777	1·4
	0·5	Q-SIS	4	9	23	69·8	4	0	99·4
		QaSIS	4	36	55	33·2	7	3	99·2
		MDC-SISQ	4	8	21	73·2	4	0	99·4
		Q-SIS cens.	4	18	59	53·0	4	1	98·8
		QaSIS cens.	4	184	285	2·8	30	41	59·6
	0·75	Q-SIS	7	285	595	1·8	36	80	52·2
		QaSIS	7	539	601	0·0	162	180	6·2
		MDC-SISQ	7	251	594	2·8	29	68	56·6
		Q-SIS cens.	7	503	794	1·8	44	161	44·8
		QaSIS cens.	7	190	888	11·8	12	48	71·2

Table 3. *Selection proportions $\mathcal{P}_{20}$, $\mathcal{P}_{21}$ and $\mathcal{P}_{22}$ (%) corresponding to nonactive predictors Z_{20}, Z_{21} and Z_{22} at $\tau = 0{\cdot}5$*

			$n = 100$			$n = 200$		
Example	Error	Method	$\mathcal{P}_{20}$	$\mathcal{P}_{21}$	$\mathcal{P}_{22}$	$\mathcal{P}_{20}$	$\mathcal{P}_{21}$	$\mathcal{P}_{22}$
1	Normal	Q-SIS	2·0	1·8	0·4	2·2	3·0	2·2
		QaSIS	57·4	69·2	58·0	66·2	76·4	64·8
		MDC-SISQ	1·6	2·0	0·6	1·4	2·8	2·4
		Q-SIS cens.	2·0	1·8	1·4	2·2	3·2	3·0
		QaSIS cens.	93·2	96·8	92·6	99·4	100·0	99·8
	Cauchy	Q-SIS	1·4	1·0	1·0	2·8	3·0	2·2
		QaSIS	61·4	71·2	58·0	69·4	80·4	71·6
		MDC-SISQ	1·4	1·4	1·0	2·8	2·6	2·0
		Q-SIS cens.	0·8	1·6	1·2	3·6	2·8	2·6
		QaSIS cens.	96·0	98·4	95·8	100·0	100·0	100·0
2	Normal	Q-SIS	1·8	1·8	1·8	3·8	4·4	3·6
		QaSIS	63·0	71·0	57·0	73·6	84·6	70·4
		MDC-SISQ	2·0	1·8	1·4	4·2	4·2	3·4
		Q-SIS cens.	2·0	1·6	2·0	5·4	6·2	6·0
		QaSIS cens.	93·0	97·4	93·0	100·0	100·0	100·0
	Cauchy	Q-SIS	3·2	3·6	2·6	2·8	3·6	4·2
		QaSIS	59·2	69·4	57·2	71·2	80·8	73·8
		MDC-SISQ	3·2	3·6	2·2	3·4	4·0	4·2
		Q-SIS cens.	1·8	3·0	2·6	3·4	5·4	6·0
		QaSIS cens.	93·6	97·8	93·4	100·0	100·0	99·8

Table 4. *Selection results for the Msa.XXX.0 genes in the cardiomyopathy microarray data example; selected genes are indicated by a √ symbol*

		Model size 9					Model size 29				
τ	Method	2134	2877	26025	15442	10108	2134	2877	26025	15442	10108
	DF-SIS	√	√		√	√	√	√	√	√	√
	SIRS	√	√	√	√	√	√	√	√	√	√
	DC-SIS	√	√	√	√	√	√	√	√	√	√
0·3	Q-SIS				√		√			√	√
	QaSIS		√					√			
	MDC-SISQ						√			√	
0·5	Q-SIS	√		√	√		√	√	√	√	√
	QaSIS										
	MDC-SISQ	√		√	√		√	√	√	√	√
0·7	Q-SIS	√	√	√		√	√	√	√		√
	QaSIS										
	MDC-SISQ	√	√	√		√	√	√	√		√

either Example 1 or Example 2, although they are present in the model error. It can be seen that our quantile screening method, its censored version, and the martingale difference correlation quantile screening approach of Shao & Zhang (2014) tend to drop these nonactive predictors, which achieves the goal of conditional quantile screening. On the contrary, both the quantile adaptive screening approach and its censored version in He et al. (2013) select these nonactive predictors with high probability.

5. Real-data example

We illustrate the application of our procedure using a dataset from a cardiomyopathy microarray study (Segal et al., 2003; Li et al., 2012b). The main goal of the study was to determine which genes influence overexpression of a G protein-coupled receptor gene, Ro1, in mice; the findings could potentially help us to understand various types of heart disease in humans. The response is the Ro1 expression level, measured for $n = 30$ specimens, and the predictors are the expression levels of $p_n = 6319$ genes. We focus on the selection results for the five genes Msa.2134.0, Msa.2877.0, Msa.26025.0, Msa.15442.0 and Msa.10108.0, for a given model size of 9 or 29. From Table 4, it can be seen that under a given model size of 29, all of the five genes were selected by all three model-free screening methods, indicating that these genes could be strongly associated with gene Ro1. The results of our quantile screening and the martingale difference correlation quantile screening of Shao & Zhang (2014) coincide with each other: Msa.2134.0 is related to Ro1 across the considered quantiles and Msa.15442.0 displays some association with Ro1 at the median and lower tails of the conditional distribution. The genes Msa.2877.0, Msa.26025.0 and Msa.10108.0 tend to be associated with the median and upper quantile of the expression of Ro1. Moreover, our quantile screening procedure detected the association between Msa.10108.0 and the lower quantile of the expression of gene Ro1, which was missed by the martingale difference correlation quantile screening. By contrast, quantile adaptive screening (He et al., 2013) found only that Msa.2877.0 would affect the lower quantile of the expression of gene Ro1.

Acknowledgement

We thank the editor, associate editor, and two referees for comments that have led to significant improvements in the article. This research was supported in part by the National Natural Science Foundation of China and the Research Grants Council of Hong Kong.

Supplementary material

Supplementary material available at *Biometrika* online includes the proofs of Proposition 1 and Theorems 1–6, along with additional numerical studies.

References

Fan, J., Feng, Y. & Song, R. (2011). Nonparametric independence screening in sparse ultra-high-dimensional additive models. *J. Am. Statist. Assoc.* **106**, 544–57.

Fan, J. & Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Am. Statist. Assoc.* **96**, 1348–60.

Fan, J. & Lv, J. (2008). Sure independence screening for ultrahigh dimensional feature space (with Discussion). *J. R. Statist. Soc.* B **70**, 849–911.

Fan, J., Samworth, R. & Wu, Y. (2009). Ultrahigh dimensional feature selection: Beyond the linear model. *J. Mach. Learn. Res.* **10**, 2013–38.

Fan, J. & Song, R. (2010). Sure independence screening in generalized linear models with NP-dimensionality. *Ann. Statist.* **38**, 3567–604.

Gorst-Rasmussen, A. & Scheike, T. (2013). Independent screening for single-index hazard rate models with ultrahigh dimensional features. *J. R. Statist. Soc.* B **75**, 217–45.

He, X., Wang, L. & Hong, H. G. (2013). Quantile-adaptive model-free variable screening for high-dimensional heterogeneous data. *Ann. Statist.* **41**, 342–69.

He, X. & Zhu, L.-X. (2003). A lack-of-fit test for quantile regression. *J. Am. Statist. Assoc.* **98**, 1013–22.

Heller, R., Heller, Y. & Gorfine, M. (2013). A consistent multivariate test of association based on ranks of distances. *Biometrika* **100**, 503–10.

Huang, J., Horowitz, J. L. & Ma, S. (2008). Asymptotic properties of bridge estimators in sparse high-dimensional regression models. *Ann. Statist.* **36**, 587–613.

Koenker, R. (2005). *Quantile Regression*. Cambridge: Cambridge University Press.

LI, G., PENG, H., ZHANG, J. & ZHU, L.-X. (2012a). Robust rank correlation based screening. *Ann. Statist.* **40**, 1846–77.
LI, R., ZHONG, W. & ZHU, L. (2012b). Feature screening via distance correlation learning. *J. Am. Statist. Assoc.* **107**, 1129–39.
LIN, D. Y., WEI, L. J. & YING, Z. (1993). Checking the Cox model with cumulative sums of martingale-based residuals. *Biometrika* **80**, 557–72.
PORTNOY, S. (2003). Censored regression quantiles. *J. Am. Statist. Assoc.* **98**, 1001–12.
SEGAL, M. R., DAHLQUIST, K. D. & CONKLIN, B. R. (2003). Regression approaches for microarray data analysis. *J. Comp. Biol.* **10**, 961–80.
SHAO, X. & ZHANG, J. (2014). Martingale difference correlation and its use in high-dimensional variable screening. *J. Am. Statist. Assoc.* **109**, 1302–18.
TIBSHIRANI, R. J. (1996). Regression shrinkage and selection via the lasso. *J. R. Statist. Soc.* B **58**, 267–88.
WANG, J. H. & WANG, L. (2009). Locally weighted censored quantile regression. *J. Am. Statist. Assoc.* **104**, 1117–28.
ZHANG, C.-H. (2010). Nearly unbiased variable selection under minimax concave penalty. *Ann. Statist.* **38**, 894–942.
ZHAO, S. D. & LI, Y. (2012). Principled sure independence screening for Cox models with ultra-high-dimensional covariates. *J. Mult. Anal.* **105**, 397–411.
ZHU, L.-P., LI, L., LI, R. & ZHU, L.-X. (2011). Model-free feature screening for ultrahigh-dimensional data. *J. Am. Statist. Assoc.* **106**, 1464–74.

[*Received January* 2014. *Revised October* 2014]