应用概率统计 第 40 卷
第 5 期 2024 年 10 月

Chinese Journal of Applied Probability and Statistics
Oct., 2024, Vol. 40, No. 5, pp. 783–799
doi: 10.12460/j.issn.1001-4268.aps.2024.2022074

# Censored Composite Conditional Quantile Screening for High-Dimensional Survival Data*

LIU Wei†[1] LI Yingqiu[2]

[1] College of Mathematics and Statistics, Hunan University of Finance and Economics, Changsha, 410205, China

[2] School of Mathematics and Statistics, Changsha University of Science and Technology, Changsha, 410114, China

**Abstract:** In this paper, we introduce the censored composite conditional quantile coefficient (cCCQC) to rank the relative importance of each predictor in high-dimensional censored regression. The cCCQC takes advantage of all useful information across quantiles and can detect nonlinear effects including interactions and heterogeneity, effectively. Furthermore, the proposed screening method based on cCCQC is robust to the existence of outliers and enjoys the sure screening property. Simulation results demonstrate that the proposed method performs competitively on survival datasets of high-dimensional predictors, particularly when the variables are highly correlated.

**Keywords:** high-dimensional survival data; censored composite conditional quantile coefficient; sure screening property; rank consistency property

**2020 Mathematics Subject Classification:** 62H20

**Citation:** LIU W, LI Y Q. Censored composite conditional quantile screening for high-dimensional survival data [J]. *Chinese J Appl Probab Statist*, 2024, **40**(**5**): 783–799.

## 1 Inroduction

High dimensionality, heterogeneity, and the existence of outliers make variable selection for censored survival data challenging. There are numerous studies in the literature on variable selection for regression problems with and without censoring. Recently, various regularization methods have been proposed for feature selection in high-dimensional data analysis, which has become increasingly prominent and important across various research fields. These methods include, but are not limited to, the LASSO[1], the smoothly clipped absolute deviation (SCAD)[2–4], the least angle regression (LARS) algorithm[5], the elastic net[6–7], the adaptive LASSO[8], and the Dantzig selector[9]. On the other hand, variable

* This project was supported by the Outstanding Youth Foundation of Hunan Provincial Department of Education (Grant No. 22B0911).

† Corresponding author, E-mail: weiwei1631028@163.com.

Received on May 7, 2022. Revised on March 22, 2023. Accepted on April 20, 2023.

screening methods for high-dimensional survival data are mostly based on the partial-likelihood of the Cox model. For example, Fan et al.[10] and Sihai-Dave-Zhao and Li[11] investigated marginal screening based on the Cox proportional hazards model. However, in practice, the true models often remain unknown, and it is unclear whether these methods will perform well under model misspecification. More importantly, these penalized algorithms are effective for mean regressions and parametric models, yet face simultaneous challenges of computational efficiency, statistical accuracy and algorithmic stability when the predictors are ultrahigh dimensional and the sample size is relatively small[12].

A computationally simple method for very high-dimensional data that performs well in practice is sure independence screening, as demonstrated in the classical regression context in [13]. In this method, the outcome variable is regressed on each covariate separately. Sure independence screening recruits the features that have the best marginal utility. In the context of least squares regression for a linear model, this corresponds to the largest marginal absolute Pearson correlation between the response and the predictor. Correlation screening is a crude yet effective way to decrease the dimensionality of data. However, as pointed out in [14], the Pearson correlation might not work well for censored survival data because it cannot be reliably estimated , especially when the censoring rate is high. In addition, its performance can be significantly affected by outliers in predictors because correlation is not a robust measure for association. Such outliers pose challenges for theoretical studies of screening methods, most of which require tail probability conditions for the covariates. To address these challenges, Song et al.[14] proposed censored rank independence screening for high-dimensional survival data. However, their method may be adversely affected by the heterogeneity that is often present in high-dimensional data. To this end, Wu and Yin[15] propose a conditional quantile screening method for high-dimensional survival data with heterogeneity, which enables us to select features that contribute to the conditional quantile of the complete or censored response given the covariates. See [16–22] for further developments. It is worth noting that He et al.[23] also proposed a quantile adaptive sure independent screening procedure for high-dimensional survival data with heterogeneity. However, compared to He et al.[23], the computational cost in [15] is significantly lower, as the former involves fitting marginal spline-based quantile regression models, which are quite computationally expensive. In this paper, we propose a censored composite conditional quantile screening (cCCQC-SIS) method for high-dimensional survival data. Our proposed method has several advantages. First, it is robust against the existence of outliers. This robustness is derived from the censored conditional quantile coefficient. Second, it is a non-model-based method, so it works for a wide class of survival models. In particular, the cCCQC makes use of all

useful information across quantiles. There have existed several papers which utilize the composite quantile idea in other statistical problems. These methods include, but are not limited to, [24–26]. In this work, we apply the same principle in the context of feature screening for survival data.

The rest of this paper is organized as follows. In Section 2, we introduce the notion of the cCCQC, and the corresponding sure screening property and rank consistency property are rigorously justified. In Section 3, we evaluate the finite sample performance of our proposals through Monte Carlo simulations. The technical details are provided in the Appendix.

## 2 Method

### 2.1 The Censored Composite Conditional Quantile Coefficient

To introduce the notion of the censored composite conditional quantile coefficient (cCCQC), we shall provide a brief discussion on the censored composite conditional quantile coefficient. Let $Y$ denote the response variable of interest, $C$ denote the censoring variable, and $\mathbf{z} = (Z_1, Z_2, \cdots, Z_p)^\top$ denote the $p$-dimensional vector of covariates. Further, define $X = \min(Y, C)$ and $\Delta = I(Y < C)$. Here $I(\cdot)$ denotes the indicator function. The observed data are independent and identically distributed copies of $\{X, \Delta, (Z_1, Z_2, \cdots, Z_p)^\top\}$ and are denoted by $\{X_i, \Delta_i, (Z_{i1}, Z_{i2}, \cdots, Z_{ip})^\top\}_{i=1}^n$. Throughout the paper, we assume that the censoring variable $C$ is independent of the response $Y$ and the covariates $\mathbf{z}$.

The censored conditional quantile (CCQ) coefficient is given by

$$\mathrm{CCQ}(X, Z_k, \tau) = \mathsf{E}\left\{\mathsf{E}\left[\{\tau - w_\tau(F) I(X < Q_\tau(Y))\} I(Z_k < \widetilde{Z}_k) | \widetilde{Z}_k\right]\right\}^2, \tag{1}$$

where $\tau \in (0, 1)$, $F(y) = \mathsf{P}(Y \leqslant y)$, the weight function

$$w_\tau(F) = \begin{cases} 1, & \text{if } \Delta = 1 \text{ or } F(C) > \tau\ , \\ \frac{\tau - F(C)}{1 - F(C)}, & \text{if } \Delta = 0 \text{ and } F(C) \leqslant \tau\ , \end{cases}$$

redistributes the masses of censored observations to the right[15,27], $\widetilde{Z}_k$ is i.i.d. copy of $Z_k$ and $Q_\tau(Y)$ is the $\tau \times 100\%$th quantile of $Y$.

Motivated by Zou and Yuan[24], Kong and Xia[28] and Xu[29], we here propose the censored composite conditional quantile coefficient (cCCQC), i.e.,

$$\mathrm{cCCQC}(X, Z_k) = \mathsf{E} \int_0^1 \left\{\mathsf{E}\left[\{\tau - w_\tau(F) I(X < Q_\tau(Y))\} I(Z_k < \widetilde{Z}_k) | \widetilde{Z}_k\right]\right\}^2 \mathrm{d}\tau, \tag{2}$$

The CCQ in (1) is very useful for handling heterogeneity. However, with a limited sample size, there is variability in the set of selected variables as $\tau$ changes, even if just

slightly. Such variability is clearly undesirable for interpretation. More importantly, some important variables are likely to be missed, simply due to chance, if we perform variable selection at any given $\tau$. Therefore, one can anticipate that cCCQC is more stable than CCQ, as it takes advantage of all useful information across quantiles to enhance the stability of CCQ.

Let $\widehat{F}_n(y) = 1 - \widehat{S}_n(y)$, where $\widehat{S}_n(y)$ is the Kaplan-Meier estimator of $Y$ based on $\{(X_i, \Delta_i)\}_{i=1}^n$. The $\tau$th sample quantile $\widehat{F}_n^{-1}(\tau)$ is an estimator of $Q_\tau(Y)$ when $Y$ is subject to right censoring. By invoking (2), a natural estimator of cCCQC is given by

$$\begin{aligned}\widehat{\mathrm{cCCQC}}(X, Z_k) &\stackrel{\mathrm{def}}{=} \frac{1}{n}\sum_{j=1}^{n}\int_0^1 \left\{\frac{1}{n}\sum_{i=1}^{n}(\tau - w_{i\tau}(\widehat{F_n})I(X_i < \widehat{F}_n^{-1}(\tau)))I(Z_{ik} < Z_{jk})\right\}^2 \mathrm{d}\tau \\ &\approx \frac{1}{n^2}\sum_{j=1}^{n}\sum_{s=1}^{n}\left\{\frac{1}{n}\sum_{i=1}^{n}(\tau_s - w_{i\tau_s}(\widehat{F_n})I(X_i < \widehat{F}_n^{-1}(\tau_s)))I(Z_{ik} < Z_{jk})\right\}^2 \quad (3)\end{aligned}$$

where $\tau_s = \frac{s}{n+1}, s = 1, \cdots, n$ and $w_{i\tau_s}(\widehat{F}_n)$ is denoted in an obvious way. The integral approximation is straightforward by invoking the precursor work of [24,30]. For the purpose of high-dimensional screening, we focus on rather than the asymptotic properties of $\widehat{\mathrm{cCCQC}}(X, Z_k)$ but instead the desirable sure screening and rank consistency properties of $\widehat{\mathrm{cCCQC}}(X, Z_k)$.

Following the work of Kong and Xia[28], and for the sake of technical convenience, we focus on rather than the case $(0, 1)$ but instead the following truncated version $[\delta^*, 1-\delta^*]$:

$$\mathrm{cCCQC}_{\mathrm{T}}(X, Z_k) = \mathsf{E}\int_{\delta^*}^{1-\delta^*}\left\{\mathsf{E}\left[\{\tau - w_\tau(F)I(X < Q_\tau(Y))\}I(Z_k < \widetilde{Z}_k)|\widetilde{Z}_k\right]\right\}^2 \mathrm{d}\tau, \quad (4)$$

and

$$\widehat{\mathrm{cCCQC}}_T(X, Z_k) \stackrel{\mathrm{def}}{=} \frac{1}{n}\sum_{j=1}^{n}\int_{\delta^*}^{1-\delta^*}\left\{\frac{1}{n}\sum_{i=1}^{n}(\tau - w_{i\tau}(\widehat{F_n})I(X_i < \widehat{F}_n^{-1}(\tau)))I(Z_{ik} < Z_{jk})\right\}^2 \mathrm{d}\tau \quad (5)$$

for some small $\delta^* \in (0, 1)$. This is due to the fact that the uniformity in $\tau$ of the strong Bahadur-type representation of $\widehat{F}_n^{-1}(\tau)$ cannot be met by all $\tau \in (0, 1)$. See the proof given in the Appendix for more details. Nevertheless, such truncation need not cause much concern. The reasons are two-fold. On one hand, the integral in (2) is approximated by summing over a sequence of discretized $\tau$ values. On the other hand, the cCCQC which is derived based on $(0, 1)$ is expected to closely resemble, if not completely identical to, that based on $[\delta^*, 1 - \delta^*]$, provided that $\delta^*$ is small enough. In practice, we follow the work by [24,28] to choose $\delta^* = 1/n$.

### 2.2 A Screening Procedure

In this section we design a sure independence screening procedure based on the cCCQC for high-dimensional survival data. Let $\mathcal{A}$ denote the index set of the active variables:

$$\mathcal{A} = \{k : \mathsf{P}(Y > t|\mathbf{z}) \text{ depends functionally on } Z_k\}.$$

With a sample of size $n$, we aim to select the set of active variables $\mathbf{z}_{\mathcal{A}}$. The following assumptions are needed.

**Assumption** 1 The truly important predictors satisfy

$$\min_{k\in\mathcal{A}} \mathrm{cCCQC}_T(X, Z_k) \geqslant 2cn^{t-\frac{1}{2}}, \text{ for some constants } c > 0, \quad 0 < t \leqslant 1/2.$$

**Assumption** 2 $F(y)$ 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 on $[Q_{\delta^*}(Y) - \varepsilon, Q_{1-\delta^*}(Y) + \varepsilon]$ for some $0 < \varepsilon < 1$.

**Assumption** 3 $G(x) = \mathsf{P}(C \leqslant x)$ 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 on $[Q_{\delta^*}(Y) - \varepsilon, Q_{1-\delta^*}(Y) + \varepsilon]$ for some $0 < \varepsilon < 1$.

**Assumption** 4 Let $L$ denote the maximum follow-up variable; then $\mathsf{P}(L \geqslant Y) \geqslant \tau_0 > 0$ for some positive constant $\tau_0$.

Assumption 1 requires the signals of the important predictors to be strong enough to be detectable by the cCCQC. Similar conditions are widely assumed in the marginal screening literature. See, for example, [14–15]. Assumptions 2–4 are common in the survival analysis literature to ensure that the Kaplan-Meier estimator and its inverse function are well behaved.

If the signal level is not too small, i.e., Assumption 1 is true, we suggest the cCCQC-SIS procedure which retains the predictors indexed by

$$\widehat{\mathcal{A}} = \left\{k : \widehat{\mathrm{cCCQC}}_T(X, Z_k) \geqslant cn^{t-\frac{1}{2}}, k = 1, \cdots, p\right\}, \tag{6}$$

where $c$ and $t$ are specified in Assumption 1.

With the above Assumptions, we can easily establish the desirable sure screening property for the cCCQC-SIS procedure without assuming the marginal distribution functions of either $\mathbf{z}$ or $Y$ , or both, have exponential tails.

**Theorem** 1 (Sure Screening Property) Suppose the Assumptions 1–4 hold. Then, we can show that there exists a sufficiently small constant $s_n$ such that

$$\mathsf{P}\left(\mathcal{A} \subseteq \widehat{\mathcal{A}}\right) \geqslant 1 - O\left[|\mathcal{A}|\left\{\exp(-c_1 n^{2t}) + \exp(c_2 n \log(1 - \frac{1}{2}s_n n^{t-\frac{1}{2}}))\right\}\right],$$

where $|\mathcal{A}|$ denotes the cardinality of the index set $\mathcal{A}$,

One can expect that $Y$ depends more upon $\mathbf{z}_{\mathcal{A}}$ than upon $\mathbf{z}_{\mathcal{A}^c}$ , though such dependence can be nonlinear. Intuitively speaking, $\mathrm{cCCQC}_T(X, Z_k)$, for $k \in \mathbf{z}_{\mathcal{A}}$, is greater than $\mathrm{cCCQC}_T(X, Z_k)$, for $k \in \mathbf{z}_{\mathcal{A}^c}$, if we use the cCCQC to measure nonlinear dependence. Such an intuition is formulated in the following assumption.

**Assumption** 5 $\liminf_{p\to\infty}\left\{\min_{k\in\mathcal{A}}\mathrm{cCCQC}_T(X, Z_k) - \max_{k\in\mathcal{A}^c}\mathrm{cCCQC}_T(X, Z_k)\right\} \geqslant d_1$, where $d_1$ is a positive constant.

Assumption 5 imposes an assumption on the gap of signal strength between active and inactive features. With Assumption 5, we can easily establish the ranking consistency property for the cCCQC-SIS procedure.

**Theorem** 2 (Rank Consistency Property) In addition to the Assumptions 1–5, we further assume that $p = o\left\{\exp\left(an^{t+\frac{1}{2}}\right)\right\}$ for any fixed $a > 0$. Then

$$\liminf_{n\to\infty}\left\{\min_{k\in\mathcal{A}}\widehat{\mathrm{cCCQC}}_T(X, Z_k) - \max_{k\in\mathcal{A}^c}\widehat{\mathrm{cCCQC}}_T(X, Z_k)\right\} \geqslant 0,$$

almost surely.

Theorem 2 ensures that the important predictors will be ranked prior to the unimportant ones with an overwhelming probability, if the signals between the important predictors and the unimportant ones are distinguishable. We shall demonstrate the usefulness of these asymptotic properties in Section 3.

## 3 Numerical Studies

In this section, we conduct simulations and a real data illustration to evaluate the empirical performance of the proposed cCCQC-based screening method. Our simulation studies are conducted using Matlab code. We compare our screening procedure (cCCQC-SIS) with the following three competitors: the censored rank independence screening [14; CR-SIS], the censored conditional quantile coefficient based sure independence screening [15; $\mathrm{CCQ}_\tau$-SIS] and the sure independent ranking and screening procedure for censored regression [16; cSIRS]. We adopt the following three criteria to compare the performance of different independence screening procedures. These three criteria are generally correlated with each other, so we present the results based on only one or two criteria in some cases to conserve space.

The minimal model size which is required to include all truly important covariates. We denote this quantity by $\mathscr{S}$. If an independent screening procedure has the sure screening property, $\mathscr{S}$ is expected to be close to the number of truly important predictors. We

report the minimum, the first quartile, the median, the third quartile and the maximum number of $\mathscr{S}$ for each independence screening method out of 1000 replications.

The selection probability that all active predictors are ranked in either the top $[n/\log n]$ or $(n-1)$ positions, where $[a]$ denotes the integer part of $a$. We denote this quantity by $\mathscr{P}_A$. This measurement counts the proportion that all truly important predictors are selected out of 1000 replications. If an independence screening procedure has the sure screening property, $\mathscr{P}_A$ is expected to be close to 1.

The selection probability that an individual important predictor is ranked in either of the top $[n/\log n]$ or $(n-1)$ positions. We denote this quantity by $\mathscr{P}_S$. It can also be used to assess the sure screening property. In addition, it is helpful to understand which predictors are mostly likely missed by a specific independent screening procedure. We expect the value of $\mathscr{P}_S$ to be close to 1 if an independent screening procedure is able to identify each important covariate.

**Example** 1 Consider the simple linear model:

$$Y_i = Z_{i,1} + 0.8Z_{i,2} + 0.6Z_{i,3} + 0.4Z_{i,4} + 0.2Z_{i,5} + \varepsilon_i. \tag{7}$$

The high-dimensional covariates $\mathbf{z}_i = (Z_{i,1}, Z_{i,2}, \cdots, Z_{i,p})^\top$ is generated from a multivariate normal population with mean zero and covariance matrix $\Sigma = \left(0.8^{|k-k'|}\right)_{p\times p}$. The error term $\varepsilon_i$ is drawn from the standard normal or standard cauchy distribution. We consider a sample size of $n = 100$ and set the number of covariates to $p = 1000$ . We take the censoring variable $C$ to be $\min(\widetilde{C}, L)$, where $\widetilde{C}$ is generated from $Un(1, L+2)$ with $L$ being the study duration variable, which is chosen to yield a censoring rate of about 30%. We consider two quantile levels $\tau = 0.50$ and $\tau = 0.75$, respectively.

It can be seen from Tables 1 and 2 that in most scenarios, our proposed method performs the best for example 1, followed by $\text{CCQ}_{0.50}$-SIS, cSRIS, $\text{CCQ}_{0.75}$-SIS and CR-SIS. However, the difference among them is small. This indicates that cCCQC-SIS, cSRIS, $\text{CCQ}_\tau$-SIS and CR-SIS are all capable of detecting the linear relationship.

**Example** 2 Consider the following linear model with heterogeneity:

$$Y_i = Z_{i,1} + 0.8Z_{i,2} + 0.6Z_{i,3} + 0.4Z_{i,4} + 0.2Z_{i,5} + \exp(Z_{i,6} + Z_{i,7} + Z_{i,8})\varepsilon_i. \tag{8}$$

The covariates and the error term are simulated as in model (7). The $L$ is also chosen to yield a censoring rate of about 30%. However, to accommodate the heterogeneity, we consider two quantile levels $\tau = 0.40$ and $\tau = 0.75$, respectively.

It can be clearly seen from Tables 1 and 3 that the proposed cCCQC-SIS performs the best for Example 2. In particular, Table 3 indicates that our proposal can detect the

**Table 1 The quantiles of the minimum model size $\mathscr{S}$ for Examples 1, 2 and 3**

| Model | Error | Method | min | 25% | 50% | 75% | 95% | 99% | max |
|---|---|---|---|---|---|---|---|---|---|
| Model (3.1) | Normal | cCCQC-SIS | 5 | 5 | 5 | 5 | 5 | 5 | 16 |
| | | CR-SIS | 5 | 5 | 5 | 5 | 21 | 441 | 838 |
| | | $CCQ_{0.50}$-SIS | 5 | 5 | 5 | 5 | 5 | 7 | 28 |
| | | $CCQ_{0.75}$-SIS | 5 | 5 | 5 | 5 | 5 | 12 | 52 |
| | | cSRIS | 5 | 5 | 5 | 5 | 5 | 5 | 19 |
| | Cauchy | cCCQC-SIS | 5 | 5 | 5 | 5 | 6 | 27 | 132 |
| | | CR-SIS | 5 | 5 | 7 | 87 | 503 | 942 | 997 |
| | | $CCQ_{0.50}$-SIS | 5 | 5 | 5 | 5 | 8 | 33 | 136 |
| | | $CCQ_{0.75}$-SIS | 5 | 5 | 5 | 7 | 21 | 182 | 782 |
| | | cSRIS | 5 | 5 | 5 | 5 | 9 | 40 | 167 |
| Model (3.2) | Normal | cCCQC-SIS | 8 | 9 | 20 | 49 | 114 | 437 | 938 |
| | | CR-SIS | 9 | 141 | 616 | 871 | 959 | 994 | 1000 |
| | | $CCQ_{0.40}$-SIS | 30 | 114 | 274 | 482 | 712 | 921 | 986 |
| | | $CCQ_{0.50}$-SIS | 13 | 72 | 150 | 372 | 636 | 905 | 977 |
| | | $CCQ_{0.75}$-SIS | 8 | 9 | 19 | 61 | 213 | 845 | 966 |
| | | cSRIS | 8 | 11 | 29 | 88 | 305 | 910 | 978 |
| | Cauchy | cCCQC-SIS | 8 | 8 | 22 | 56 | 144 | 581 | 945 |
| | | CR-SIS | 24 | 161 | 717 | 926 | 979 | 998 | 1000 |
| | | $CCQ_{0.40}$-SIS | 8 | 184 | 405 | 548 | 772 | 961 | 998 |
| | | $CCQ_{0.75}$-SIS | 8 | 8 | 28 | 81 | 281 | 801 | 959 |
| | | cSRIS | 8 | 12 | 32 | 94 | 326 | 922 | 989 |
| Model (3.3) | Normal | cCCQC-SIS | 6 | 6 | 10 | 26 | 60 | 404 | 845 |
| | | CR-SIS | 7 | 18 | 166 | 658 | 909 | 991 | 1000 |
| | | $CCQ_{0.50}$-SIS | 6 | 8 | 19 | 61 | 151 | 493 | 861 |
| | | $CCQ_{0.75}$-SIS | 6 | 6 | 13 | 60 | 210 | 614 | 969 |
| | | cSRIS | 6 | 6 | 18 | 44 | 89 | 453 | 890 |
| | Cauchy | cCCQC-SIS | 6 | 6 | 21 | 50 | 133 | 446 | 958 |
| | | CR-SIS | 6 | 12 | 101 | 553 | 895 | 992 | 1000 |
| | | $CCQ_{0.50}$-SIS | 6 | 11 | 48 | 105 | 275 | 674 | 987 |
| | | $CCQ_{0.75}$-SIS | 6 | 7 | 23 | 106 | 321 | 740 | 956 |
| | | cSRIS | 6 | 9 | 39 | 90 | 117 | 522 | 980 |

**Table 2 The empirical probabilities $\mathscr{P}_S$ and $\mathscr{P}_A$ for Example 1**

| | | | $\mathscr{P}_S$ | | | | | $\mathscr{P}_A$ |
|---|---|---|---|---|---|---|---|---|
| Model Size | Error | Method | $X_1$ | $X_2$ | $X_3$ | $X_4$ | $X_5$ | |
| $(n-1)$ | Normal | cCCQC-SIS | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |
| | | CR-SIS | 1.00 | 1.00 | 0.98 | 0.95 | 0.92 | 0.88 |
| | | $CCQ_{0.50}$-SIS | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |
| | | $CCQ_{0.75}$-SIS | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |
| | | cSRIS | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |
| | Cauchy | cCCQC-SIS | 1.00 | 1.00 | 1.00 | 1.00 | 0.96 | 0.96 |
| | | CR-SIS | 1.00 | 0.95 | 0.90 | 0.86 | 0.80 | 0.71 |
| | | $CCQ_{0.50}$-SIS | 1.00 | 1.00 | 1.00 | 0.97 | 0.94 | 0.89 |
| | | $CCQ_{0.75}$-SIS | 1.00 | 1.00 | 0.98 | 0.93 | 0.86 | 0.78 |
| | | cSRIS | 1.00 | 1.00 | 1.00 | 0.97 | 0.93 | 0.91 |

**Table 3 The empirical probabilities $\mathscr{P}_S$ and $\mathscr{P}_A$ for Example 2**

| | | | $\mathscr{P}_S$ | | | | | | | | $\mathscr{P}_A$ |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Model Size | Error | Method | $X_1$ | $X_2$ | $X_3$ | $X_4$ | $X_5$ | $X_6$ | $X_7$ | $X_8$ | |
| $(n-1)$ | Normal | cCCQC-SIS | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 0.98 | 0.91 | 0.80 | 0.74 |
| | | CR-SIS | 0.69 | 0.76 | 0.73 | 0.58 | 0.38 | 0.19 | 0.14 | 0.11 | 0.04 |
| | | $\text{CCQ}_{0.40}$-SIS | 1.00 | 1.00 | 0.99 | 0.95 | 0.64 | 0.31 | 0.24 | 0.25 | 0.05 |
| | | $\text{CCQ}_{0.50}$-SIS | 1.00 | 1.00 | 0.98 | 0.97 | 0.73 | 0.49 | 0.46 | 0.38 | 0.16 |
| | | $\text{CCQ}_{0.75}$-SIS | 0.98 | 1.00 | 1.00 | 0.99 | 0.96 | 0.90 | 0.80 | 0.68 | 0.63 |
| | | cSRIS | 1.00 | 1.00 | 1.00 | 1.00 | 0.97 | 0.95 | 0.88 | 0.75 | 0.64 |
| | Cauchy | cCCQC-SIS | 0.98 | 0.98 | 0.98 | 0.96 | 0.95 | 0.95 | 0.94 | 0.88 | 0.70 |
| | | CR-SIS | 0.52 | 0.55 | 0.48 | 0.35 | 0.27 | 0.18 | 0.09 | 0.08 | 0.03 |
| | | $\text{CCQ}_{0.40}$-SIS | 0.98 | 0.98 | 0.95 | 0.80 | 0.43 | 0.27 | 0.28 | 0.32 | 0.03 |
| | | $\text{CCQ}_{0.50}$-SIS | 0.94 | 0.95 | 0.92 | 0.86 | 0.47 | 0.35 | 0.32 | 0.36 | 0.09 |
| | | $\text{CCQ}_{0.75}$-SIS | 0.89 | 0.90 | 0.91 | 0.90 | 0.87 | 0.89 | 0.74 | 0.63 | 0.53 |
| | | cSRIS | 0.97 | 0.96 | 0.97 | 0.94 | 0.94 | 0.92 | 0.87 | 0.72 | 0.57 |

**Table 4 The empirical probabilities $\mathscr{P}_S$ and $\mathscr{P}_A$ for Example 2 with $p = 4000$, cauchy error, model size $n-1$ and the active covariates spread out**

| | $\mathscr{P}_S$ | | | | | | | | $\mathscr{P}_A$ |
|---|---|---|---|---|---|---|---|---|---|
| Method | $X_{2001}$ | $X_{2002}$ | $X_{2003}$ | $X_{2004}$ | $X_{2005}$ | $X_{2006}$ | $X_{2007}$ | $X_{2008}$ | |
| cCCQC-SIS | 0.95 | 0.97 | 0.96 | 0.95 | 0.91 | 0.92 | 0.90 | 0.89 | 0.76 |
| CR-SIS | 0.50 | 0.52 | 0.49 | 0.37 | 0.23 | 0.22 | 0.14 | 0.13 | 0.06 |
| $\text{CCQ}_{0.40}$-SIS | 0.96 | 0.95 | 0.96 | 0.86 | 0.45 | 0.30 | 0.29 | 0.34 | 0.11 |
| $\text{CCQ}_{0.50}$-SIS | 0.95 | 0.95 | 0.94 | 0.89 | 0.46 | 0.37 | 0.36 | 0.40 | 0.18 |
| $\text{CCQ}_{0.75}$-SIS | 0.91 | 0.92 | 0.92 | 0.93 | 0.88 | 0.90 | 0.78 | 0.66 | 0.59 |
| cSRIS | 0.96 | 0.94 | 0.96 | 0.95 | 0.90 | 0.91 | 0.89 | 0.75 | 0.65 |

**Table 5 The empirical probabilities $\mathscr{P}_S$ and $\mathscr{P}_A$ for Example 2 with the heavy censoring case, normal error and model size $n-1$**

| | $\mathscr{P}_S$ | | | | | | | | $\mathscr{P}_A$ |
|---|---|---|---|---|---|---|---|---|---|
| Method | $X_1$ | $X_2$ | $X_3$ | $X_4$ | $X_5$ | $X_6$ | $X_7$ | $X_8$ | |
| cCCQC-SIS | 0.89 | 0.90 | 0.86 | 0.91 | 0.94 | 0.86 | 0.84 | 0.77 | 0.62 |
| CR-SIS | 0.54 | 0.60 | 0.61 | 0.49 | 0.30 | 0.13 | 0.09 | 0.08 | 0.01 |
| $\text{CCQ}_{0.40}$-SIS | 0.88 | 0.87 | 0.85 | 0.90 | 0.52 | 0.20 | 0.15 | 0.13 | 0.02 |
| $\text{CCQ}_{0.50}$-SIS | 0.90 | 0.92 | 0.86 | 0.90 | 0.64 | 0.40 | 0.37 | 0.28 | 0.10 |
| $\text{CCQ}_{0.75}$-SIS | 0.87 | 0.89 | 0.85 | 0.89 | 0.92 | 0.82 | 0.63 | 0.56 | 0.41 |
| cSRIS | 0.81 | 0.85 | 0.83 | 0.88 | 0.90 | 0.81 | 0.76 | 0.63 | 0.49 |

**Table 6 The empirical probabilities $\mathscr{P}_S$ and $\mathscr{P}_A$ for Example 3**

| | | | $\mathscr{P}_S$ | | | | | | $\mathscr{P}_A$ |
|---|---|---|---|---|---|---|---|---|---|
| Model Size | Error | Method | $X_1$ | $X_2$ | $X_3$ | $X_4$ | $X_5$ | $X_6$ | |
| $[n/\log n]$ | Normal | cCCQC-SIS | 0.59 | 0.83 | 0.84 | 0.87 | 0.92 | 0.88 | 0.42 |
| | | CR-SIS | 0.26 | 0.43 | 0.46 | 0.44 | 0.23 | 0.18 | 0.06 |
| | | $CCQ_{0.50}$-SIS | 0.49 | 0.70 | 0.68 | 0.68 | 0.82 | 0.68 | 0.26 |
| | | $CCQ_{0.75}$-SIS | 0.43 | 0.60 | 0.69 | 0.70 | 0.87 | 0.70 | 0.34 |
| | | cSRIS | 0.44 | 0.57 | 0.63 | 0.69 | 0.89 | 0.68 | 0.30 |
| | Cauchy | cCCQC-SIS | 0.44 | 0.75 | 0.76 | 0.70 | 0.82 | 0.77 | 0.31 |
| | | CR-SIS | 0.37 | 0.45 | 0.41 | 0.42 | 0.29 | 0.15 | 0.11 |
| | | $CCQ_{0.50}$-SIS | 0.40 | 0.60 | 0.56 | 0.52 | 0.65 | 0.45 | 0.13 |
| | | $CCQ_{0.75}$-SIS | 0.39 | 0.53 | 0.55 | 0.60 | 0.78 | 0.65 | 0.23 |
| | | cSRIS | 0.36 | 0.52 | 0.54 | 0.57 | 0.73 | 0.60 | 0.19 |
| $(n-1)$ | Normal | cCCQC-SIS | 0.87 | 0.96 | 0.93 | 0.96 | 1.00 | 0.98 | 0.82 |
| | | CR-SIS | 0.42 | 0.62 | 0.63 | 0.62 | 0.47 | 0.38 | 0.18 |
| | | $CCQ_{0.50}$-SIS | 0.78 | 0.90 | 0.89 | 0.91 | 0.97 | 0.92 | 0.60 |
| | | $CCQ_{0.75}$-SIS | 0.65 | 0.82 | 0.86 | 0.89 | 0.99 | 0.96 | 0.56 |
| | | cSRIS | 0.66 | 0.80 | 0.83 | 0.85 | 0.98 | 0.94 | 0.52 |
| | Cauchy | cCCQC-SIS | 0.76 | 0.93 | 0.89 | 0.95 | 0.96 | 0.94 | 0.69 |
| | | CR-SIS | 0.47 | 0.60 | 0.60 | 0.55 | 0.41 | 0.29 | 0.24 |
| | | $CCQ_{0.50}$-SIS | 0.60 | 0.80 | 0.81 | 0.83 | 0.92 | 0.87 | 0.48 |
| | | $CCQ_{0.75}$-SIS | 0.57 | 0.76 | 0.79 | 0.85 | 0.94 | 0.90 | 0.47 |
| | | cSRIS | 0.62 | 0.83 | 0.82 | 0.84 | 0.90 | 0.81 | 0.48 |

heteroscedasitic errors with an over-whelming probability. As expected, cCCQC-SIS is more stable than $CCQ_\tau$-SIS in that the former takes advantage of all useful information across quantiles. Also, the CR-SIS and cSRIS have unsatisfactory performance in this example due to the heterogeneity.

**Example** 3 Consider the following nonlinear model including a three-way interaction term:

$$Y_i = X_{i,1}^2 + 3X_{i,2}X_{i,3}X_{i,4} + 5X_{i,5}X_{i,6} + \varepsilon_i. \tag{9}$$

We keep the rest of the set-up the same as in model (7). From Tables 1 and 4, it is evident that the proposed cCCQC-SIS performs best for Example 3 in comparison with the existing counterparts. However, the differences among them are substantial. This indicates that compared to the existing choices, cCCQC-SIS has an excellent capability of identifying the interactions.

**Example** 4 Upon the suggestion of a reviewer, we reconsider the model (9) using $\log(Y_i)$ instead of $Y_i$. A small $\mathscr{S}$ tends to be associated with high proportions for $\mathscr{P}_A$ and $\mathscr{P}_S$. So we present the results based on the criterion $\mathscr{S}$ in this example to conserve space. From the simulation results summarized in Tables 4–6, we can draw similar conclusions to Example 1.

**Table 7 The quantiles of the minimum model size $\mathscr{S}$ for Example 4**

| Error | Method | min | 25% | 50% | 75% | 95% | 99% | max |
|---|---|---|---|---|---|---|---|---|
| Normal | cCCQC-SIS | 6 | 6 | 6 | 12 | 27 | 116 | 622 |
| | CR-SIS | 6 | 8 | 55 | 129 | 420 | 604 | 1000 |
| | $CCQ_{0.50}$-SIS | 6 | 6 | 10 | 35 | 86 | 216 | 813 |
| | $CCQ_{0.75}$-SIS | 6 | 6 | 7 | 21 | 71 | 189 | 916 |
| | cSRIS | 6 | 6 | 11 | 39 | 95 | 289 | 848 |
| Cauchy | cCCQC-SIS | 6 | 6 | 7 | 17 | 38 | 224 | 889 |
| | CR-SIS | 6 | 10 | 65 | 198 | 476 | 752 | 1000 |
| | $CCQ_{0.50}$-SIS | 6 | 8 | 20 | 46 | 98 | 314 | 933 |
| | $CCQ_{0.75}$-SIS | 6 | 8 | 15 | 40 | 85 | 233 | 908 |
| | cSRIS | 6 | 8 | 24 | 56 | 102 | 335 | 950 |

**Example** 5 As an illustration, we apply the proposed screening method to the analysis of microarray diffuse large-B-cell lymphoma (DLBCL) data of [31]. The DLBCL is one of the most common types of lymphoma in adults of United States. However, the survival rate after the standard chemotherapy is only about 35 to 40%. Thus it is of interest in studying how the survival rate depends on an individual's gene information The outcome in the study was the survival variable of $n = 240$ DLBCL patients after chemotherapy. Measurements of $p = 7399$ genes obtained from cDNA microarrays for each individual patient were the predictors. Given such a large number of predictors and small sample size, feature screening seems a necessary initial step as a prelude to any other sophisticated statistical modeling that does not cope well with such high dimensionality.

In this data set, all gene expression levels are standardized to have mean zero and standard deviation one during the exploratory data analysis. We split these data set into a training set with $n_1$ subjects and a test setwith $n_2$ subjects. Here $n_1 + n_2 = 240$. We first apply the screening procedures to the training data set, and retain $[n_1/\log n_1]$ covariates during this screening stage. Considering that some truly unimportant covariates are also retained in the screening stage, we next perform the lasso penalization to further remove those irrelevant covariates. We then build an un-penalized Cox proportional hazards model using the selected genes. We next apply the log-rank test to compare the prediction power of different screening methods. Table 6 describes the Kaplan–Meier estimate of survival curves for the two risk groups of patients in the testing data with the log-rank test yielding different p-values. These results indicate our good prediction of the fitted model.

## Appdendix A: Proof of Theorem 1

The following Lemma paves the road for proving Theorem 1. Lemma 3 is the modified version of Lemma S1 of [15]. Hence, the details are omitted here and a detailed technical

**Table 8 The p-values of the log-rank test for Example 5 with several combinations of $(n_1, n_2)$**

| $(n_1, n_2)$ | cCCQC-SIS | CR-SIS | $\text{CCQ}_{0.50}$-SIS | $\text{CCQ}_{0.75}$-SIS | cSRIS |
|---|---|---|---|---|---|
| $(120, 120)$ | 0.001 | 0.034 | 0.120 | 0.060 | 0.021 |
| $(180, 60)$ | 0.003 | 0.019 | 0.082 | 0.113 | 0.009 |
| $(80, 160)$ | 0.004 | 0.134 | 0.107 | 0.085 | 0.115 |

report is available from the author.

**Lemma 3** Let $\mathcal{F}$ be a class of distribution functions whose support is the same as that of $F$, and let $\mathcal{Y}$ be the support of $Y$. For any $\varepsilon > 0$, let

$$\mathcal{H}_\tau(\varepsilon) \stackrel{\text{def}}{=} \left\{ F^* \in \mathcal{F} : \sup_{y\in\mathcal{Y}} |F^*(y) - F(Y)| \leqslant \varepsilon \text{ and } |Q_\tau(Y^*) - Q_\tau(Y)| \leqslant \varepsilon \right\},$$

where $Y^*$ follows the distribution $F^*$. Then

$$\begin{aligned}
&\sup_{\tau\in[\delta^*,1-\delta^*]} \sup_{\mathcal{H}_\tau(\varepsilon)} |w(F^*) I(X \leqslant Q_\tau(Y^*)) - w(F^*) I(X \leqslant Q_\tau(Y))| \\
\leqslant & c_{01}\varepsilon + \sup_{\tau\in[\delta^*,1-\delta^*]} I(Q_\tau(Y) - \varepsilon < Y \leqslant Q_\tau(Y) + \varepsilon) \\
&+ 3 \sup_{\tau\in[\delta^*,1-\delta^*]} I(Q_\tau(Y) - \varepsilon < C \leqslant Q_\tau(Y) + \varepsilon) \\
&+ \sup_{\tau\in[\delta^*,1-\delta^*]} I\left(F^{-1}(\tau-\varepsilon) < C \leqslant F^{-1}(\tau+\varepsilon)\right),
\end{aligned}$$

where the constant $c_{01}$ is independent of $\tau$.

**Proof of Theorem 1** Define

$$\widetilde{\text{cCCQC}}_T(Y, X_k) \stackrel{\text{def}}{=} \frac{1}{n}\sum_{j=1}^{n} \int_{\delta^*}^{1-\delta^*} \left\{ \frac{1}{n}\sum_{i=1}^{n} (\tau - w_{i\tau}(F) I(X_i < Q_\tau(Y))) I(Z_{ik} < Z_{jk}) \right\}^2 \mathrm{d}\tau.$$

Simple calculations yield

$$\widetilde{\text{cCCQC}}_T(Y, X_k) = \frac{(n-1)(n-2)}{n^2} \left( \frac{1}{n-2} \widetilde{R}_{k1} + \widetilde{R}_{k2} \right), \tag{A.1}$$

where

$$\begin{aligned}
\widetilde{R}_{k1} &= \frac{2}{n(n-1)} \sum_{i<j} \frac{1}{2} \left\{ \int_{\delta^*}^{1-\delta^*} (\tau - w_{i\tau}(F) I(X_i < Q_\tau(Y)))^2 \mathrm{d}\tau I(Z_{ik} < Z_{jk}) \right. \\
&\quad \left. + \int_{\delta^*}^{1-\delta^*} (\tau - w_{j\tau}(F) I(X_j < Q_\tau(Y)))^2 \mathrm{d}\tau I(Z_{jk} < Z_{ik}) \right\} \\
&\stackrel{\text{def}}{=} \frac{2}{n(n-1)} \sum_{i<j} h_1(Z_{ik}; X_i; Z_{jk}; X_j; F)
\end{aligned}$$

and

$$\begin{aligned}\widetilde{R}_{k2} =& \frac{6}{n(n-1)(n-2)} \sum_{i<j<l} \frac{1}{3} \Bigg\{ I\left(Z_{ik} < Z_{lk}\right) I\left(Z_{jk} < Z_{lk}\right) \\ & \times \int_{\delta^*}^{1-\delta^*} \left(\tau - w_{i\tau}(F) I\left(X_i < Q_\tau(Y)\right)\right) \left(\tau - w_{j\tau}(F) I\left(X_j < Q_\tau(Y)\right)\right) \mathrm{d}\tau \\ & + \int_{\delta^*}^{1-\delta^*} \left(\tau - w_{j\tau}(F) I\left(X_j < Q_\tau(Y)\right)\right) \left(\tau - w_{l\tau}(F) I\left(X_l < Q_\tau(Y)\right)\right) \mathrm{d}\tau \\ & \times I\left(Z_{jk} < Z_{ik}\right) I\left(Z_{lk} < Z_{ik}\right) \\ & + \int_{\delta^*}^{1-\delta^*} \left(\tau - w_{l\tau}(F) I\left(X_l < Q_\tau(Y)\right)\right) \left(\tau - w_{i\tau}(F) I\left(X_i < Q_\tau(Y)\right)\right) \mathrm{d}\tau \\ & \times I\left(Z_{lk} < Z_{jk}\right) I\left(Z_{ik} < Z_{jk}\right) \Bigg\} \\ \stackrel{\text{def}}{=}& \frac{2}{n(n-1)} \sum_{i<j} h_2\left(Z_{ik}; X_i; Z_{jk}; X_j; Z_{lk}; X_l; F\right)\end{aligned}$$

with $h_1$ and $h_2$ being the kernels of the U-statistics. Likewise, we have

$$\widehat{\mathrm{cCCQC}}_T(Y, X_k) = \frac{(n-1)(n-2)}{n^2} \left( \frac{1}{n-2} \widehat{R}_{k1} + \widehat{R}_{k2} \right), \tag{A.2}$$

where $\widehat{R}_{k1}$ is obtained by replacing $F$ and $Q_\tau(Y)$ in $\widetilde{R}_{k1}$ with $\widehat{F}_n$ and $\widehat{F}_n^{-1}(\tau)$, respectively, and similarly for $\widehat{R}_{k2}$.

Due to the fact $I(\cdot)$ is uniformly bounded, simple calculations yield

$$\begin{aligned}\left|\widehat{R}_{k1} - \widetilde{R}_{k1}\right| \leqslant & \frac{2}{n} \sum_{i=1}^{n} \left| \int_{\delta^*}^{1-\delta^*} w_{i\tau}(\widehat{F}_n) I(X_i < \widehat{F}_n^{-1}(\tau)) \mathrm{d}\tau \right. \\ & \left. - \int_{\delta^*}^{1-\delta^*} w_{i\tau}(F) I(X_i < Q_\tau(Y)) \mathrm{d}\tau \right| \qquad \text{(A.3)} \\ \leqslant & \frac{2}{n} \sum_{i=1}^{n} \sup_{\tau \in [\delta^*, 1-\delta^*]} \left| w_{i\tau}(\widehat{F}_n) I(X_i < \widehat{F}_n^{-1}(\tau)) - w_{i\tau}(F) I(X_i < Q_\tau(Y)) \right| \qquad \text{(A.4)}\end{aligned}$$

and

$$\left|\widehat{R}_{k2} - \widetilde{R}_{k2}\right| \leqslant \frac{1}{n} \sum_{i=1}^{n} \sup_{\tau \in [\delta^*, 1-\delta^*]} \left| w_{i\tau}(\widehat{F}_n) I(X_i < \widehat{F}_n^{-1}(\tau)) - w_{i\tau}(F) I(X_i < Q_\tau(Y)) \right|. \tag{A.5}$$

Under Assumptions (A.2), (A.3) and (A.5), we have $\left|\left|\widehat{F}_n - F\right|\right|_\infty = O\left(n^{-1/2} \left(\log(n)\right)^{1/2}\right)$ and $\left|\left|\widehat{F}_n^{-1} - Q_\tau(Y)\right|\right|_\infty = O(n^{-1/2} \left(\log(n)\right)^{1/2})$ almost surely via invoking Lemma 8.4 in [23].

Employing arguments similar to those for dealing with (S11)–(S13) in [15] and combining Lemma 3, we have that there exists a positive constant $c_1, c_2$, $c_3$, $c_4$ and $c_5$ such

that

$$\begin{aligned}
&\mathsf{P}\left(\frac{1}{n}\sum_{i=1}^{n}\sup_{\tau\in[\delta^*,1-\delta^*]}\left|w_{i\tau}(\widehat{F}_n)I\left(X_i<\widehat{F}_n^{-1}(\tau)\right)-w_{i\tau}(F)I\left(X_i<Q_\tau(Y)\right)\right|\geqslant cn^{t-\frac{1}{2}}\right)\\
\leqslant&\mathsf{P}\left(\frac{1}{n}\sum_{i=1}^{n}\sup_{\tau\in[\delta^*,1-\delta^*]}I\left(Q_\tau(Y)-c_1cn^{t-\frac{1}{2}}<Y_i\leqslant Q_\tau(Y)+c_1cn^{t-\frac{1}{2}}\right)\geqslant\frac{1}{4}cn^{t-\frac{1}{2}}\right)\\
&+\mathsf{P}\left(\frac{3}{n}\sum_{i=1}^{n}\sup_{\tau\in[\delta^*,1-\delta^*]}I\left(Q_\tau(Y)-c_2cn^{t-\frac{1}{2}}<C_i\leqslant Q_\tau(Y)+c_2cn^{t-\frac{1}{2}}\right)\geqslant\frac{1}{4}cn^{t-\frac{1}{2}}\right)\\
&+\mathsf{P}\left(\frac{1}{n}\sum_{i=1}^{n}\sup_{\tau\in[\delta^*,1-\delta^*]}I\left(Q_\tau(Y)-c_3cn^{t-\frac{1}{2}}<C_i\leqslant Q_\tau(Y)+c_3cn^{t-\frac{1}{2}}\right)\geqslant\frac{1}{4}cn^{t-\frac{1}{2}}\right)\\
\leqslant&\exp\left(-c_4n^{2t}\right)+2\exp\left(-c_5n^{2t}\right).
\end{aligned}$$

Using the result above, we get

$$\begin{aligned}
&\mathsf{P}\left(\left|\widetilde{\mathrm{cCCQC}}_T(Y,X_k)-\widehat{\mathrm{cCCQC}}_T(Y,X_k)\right|\geqslant 2cn^{t-\frac{1}{2}}\right)\\
\leqslant&\mathsf{P}\left(\frac{n-1}{n^2}\left|\widetilde{R}_{k1}-\widehat{R}_{k1}\right|\geqslant cn^{t-\frac{1}{2}}\right)+\mathsf{P}\left(\frac{(n-1)(n-2)}{n^2}\left|\widetilde{R}_{k2}-\widehat{R}_{k2}\right|\geqslant cn^{t-\frac{1}{2}}\right)\\
\leqslant&\mathsf{P}\left(\left|\widetilde{R}_{k1}-\widehat{R}_{k1}\right|\geqslant cn^{t-\frac{1}{2}}\right)+\mathsf{P}\left(\left|\widetilde{R}_{k2}-\widehat{R}_{k2}\right|\geqslant cn^{t-\frac{1}{2}}\right)\\
\leqslant&\exp\left(-c_6n^{2t}\right).
\end{aligned}\tag{A.6}$$

On the other hand, invoking the proof of Theorem 2 in [32], we can directly use the theory of U-statistics to establish asymptotic property of $\widehat{\mathrm{cCCQC}}_T(Y,X_k)$. Our following arguments are exactly parallel to those used in the proof of Theorem 2 of [32] with a slight modification. Hence, the details are omitted here and a detailed technical report is available from the author. In other words, it is easy to show that there exists a sufficiently small constant $s_n\in(0,2n^{\frac{1}{2}-t})$ such that

$$\mathsf{P}\left(\left|\widetilde{\mathrm{CMDC}}_T(Y,X_k)-\mathrm{CMDC}_T(Y,X_k)\right|>cn^{t-\frac{1}{2}}\right)\leqslant O\left(\exp\left(c_7n\log\left(1-\frac{1}{2}s_nn^{t-\frac{1}{2}}\right)\right)\right).\tag{A.7}$$

Combining (A.6) and (A.7) leads to the desired result.

## Appdendix B: Proof of Theorem 2

Denote $w_k=\mathrm{cCCQC}(X,Z_k)$ and $\widehat{w}_k=\widehat{\mathrm{cCCQC}}(X,Z_k)$. The proof of Theorem 2 follows the proofs of Theorem 2.2 in [33].

$$\mathsf{P}\left(\min_{k\in\mathcal{A}}\widehat{w}_k-\max_{k\in\mathcal{A}^c}\widehat{w}_k<d_1/2\right)\leqslant\mathsf{P}\left(\min_{k\in\mathcal{A}}\widehat{w}_k-\max_{k\in\mathcal{A}^c}\widehat{w}_k-\left(\min_{k\in\mathcal{A}}w_k-\max_{k\in\mathcal{A}^c}w_k\right)<-d_1/2\right)$$

$$\leqslant \mathsf{P}\left(\left|\min_{k\in\mathcal{A}}\widehat{w}_k - \max_{k\in\mathcal{A}^c}\widehat{w}_k - \left(\min_{k\in\mathcal{A}} w_k - \max_{k\in\mathcal{A}^c} w_k\right)\right| > d_1/2\right)$$
$$\leqslant \mathsf{P}\left(2\max_{1\leqslant k\leqslant p}|\widehat{w}_k - w_k| > d_1/2\right)$$
$$\leqslant O\left[p\left\{n\exp\left(-c_1 n^{2t}\right) + \exp\left(c_2 n\log\left(1-\frac{1}{2}s_n n^{t-\frac{1}{2}}\right)\right)\right\}\right]$$

Noting that $p = o\left(\exp(an^{t+\frac{1}{2}})\right)$, then we have that for some sufficiently large $N$,

$$\sum_{n=N}^{\infty} p\left\{n\exp\left(-c_1 n^{2t}\right) + \exp\left(c_2 n\log\left(1-\frac{1}{2}s_n n^{t-\frac{1}{2}}\right)\right)\right\} < c\sum_{n=N}^{\infty} n^{-2} < \infty.$$

Hence, using Borel-Contelli Lemma leads to the desired result.

□

## References


[1] TIBSHIRANI R. Regression shrinkage and selection via the lasso [J]. *J R Stat Soc Ser B*, 1996, **58**(**1**): 267–288.

[2] FAN J Q, LI R Z. Variable selection via nonconcave penalized likelihood and it oracle properties [J]. *J Amer Statist Assoc*, 2001, **96**(**456**): 1348–1360.

[3] KIM Y, CHOI H, OH H S. Smoothly clipped absolute deviation on high dimensions [J]. *J Amer Statist Assoc*, 2008, **103**(**484**): 1665–1673.

[4] ZOU H, LI R Z. One-step sparse estimates in nonconcave penalized likelihood models [J]. *Ann Statist*, 2008, **36**(**4**): 1509–1533.

[5] EFRON B, HASTIE T, JOHNSTONE I, et al. Least angle regression (with discussion) [J]. *Ann Statist*, 2004, **32**(**2**): 409–499.

[6] ZOU H, HASTIE T. Regularization and variable selection via the elastic net [J]. *J R Stat Soc Ser B*, 2005, **67**(**2**): 301–320.

[7] ZOU H, ZHANG H H. On the adaptive elastic-net with a diverging number of parameters [J]. *Ann Statist*, 2009, **37**(**4**): 1733–1751.

[8] ZOU H. The adaptive lasso and its oracle properties [J]. *J Amer Statist Assoc*, 2006, **101**(**476**): 1418–1429.

[9] CANDES E, TAO T. The Dantzig selector: Statistical estimation when $p$ is much larger than $n$ (with discussion) [J]. *Ann Statist*, 2007, **35**(**6**): 2313–2404.

[10] FAN J Q, FENG Y, WU Y C. Ultrahigh dimensional variable selection for Cox's proportional hazards model [J]. *IMS Collections*, 2010, **6**: 70–86.

[11] Sihai-Dave-Zhao, LI Y. Principled sure independence screening for Cox models with ultrahigh-dimensional covariates [J]. *J Multivariate Anal*, 2012, **105**(**1**): 397–411.

[12] FAN J Q, SAMWORTH R, WU Y C. Ultrahigh dimensional feature selection: Beyond the linear model [J]. *J Mach Learn Res*, 2009, **10**: 2013–2038.

[13] FAN J Q, LV J C. Sure independence screening for ultrahigh dimensional feature space (with discussion) [J]. *J R Stat Soc Ser B*, 2008, **70**(**5**): 849–911.

[14] SONG R, LU W B, MA S G, et al. Censored rank independence screening for high-dimensional survival data [J]. *Biometrika*, 2014, **101**(**4**): 799–814.

[15] WU Y S, YIN G S. Conditional quantile screening in ultrahigh-dimensional heterogeneous data [J]. *Biometrika*, 2015, **102**(**1**): 65–76.

[16] ZHOU T Y, ZHU L P. Model-free feature screening for ultrahigh dimensional censored regression [J]. *Stat Comput*, 2017, **27**(**4**): 947–961.

[17] XU K, HUANG X D. Conditional-quantile screening for ultrahigh-dimensional survival data via martingale difference correlation [J]. *Sci China Math*, 2018, **61**(**10**): 1907–1922.

[18] ZHANG J, YIN G S, LIU Y Y, et al. Censored cumulative residual independent screening for ultrahigh-dimensional survival data [J]. *Lifetime Data Anal*, 2018, **24**(**2**): 273–292.

[19] PAN W L, WANG X Q, XIAO W N, et al. A generic sure independence screening procedure [J]. *J Amer Statist Assoc*, 2019, **114**(**526**): 928–937.

[20] XU K, HUANG X D. Feature screening for high-dimensional survival data via censored quantile correlation [J]. *J Sys Sci Complex*, 2021, **34**(**3**): 1207–1224.

[21] ZHANG J, LIU Y Y, CUI H J. Model-free feature screening via distance correlation for ultrahigh dimensional survival data [J]. *Stat Pap*, 2021, **62**(**6**): 2711–2738.

[22] XU K, SHEN Z, HUANG X D, et al. Projection correlation between scalar and vector variables and its use in feature screening with multi-response data [J]. *J Stat Computat Sim*, 2020, **90**(**11**): 1923–1942.

[23] HE X M, WANG L, HONG H G. Quantile-adaptive model-free variable screening for high-dimensional heterogeneous data [J]. *Ann Statist*, 2013, **41**(**1**): 342–369.

[24] ZOU H, YUAN M. Composite quantile regression and the oracle model selection theory [J]. *Ann Statist*, 2008, **36**(**3**): 1108–1126.

[25] FAN Y, TANG M L, TIAN M Z. Composite quantile regression for varying-coefficient single-index models [J]. *Commun Stat-theor M*, 2016, **45**(**10**): 3027–3047.

[26] ZHAO W H, LIAN H, SONG X Y. Composite quantile regression for correlated data [J]. *Comput Stat Data Anal*, 2017, **109**: 15–33.

[27] WANG H J, WANG L. Locally weighted censored quantile regression [J]. *J Amer Statist Assoc*, 2009, **104**(**487**): 1117–1128.

[28] KONG E, XIA Y C. An adaptive composite quantile approach to dimension reduction [J]. *Ann Statist*, 2014, **42**(**4**): 1657–1688.

[29] XU K. Model-free feature screening via a modified composite quantile correlation [J]. *J Stat Plan Infer*, 2017, **188**: 22–35.

[30] MA X J, ZHANG J X. Robust model-free feature screening via quantile correlation [J]. *J Multivariate Anal*, 2016, **143**: 472–480.

[31] ROSENWALD A, WRIGHT G, CHAN W C, et al. The use of molecular profiling to predict survival after chemotherapy for diffuse large-B-cell lymphoma [J]. *New Engl J Med*, 2002, **346**(**25**): 1937–1947.

[32] ZHU L P, LI L X, LI R Z, et al. Model-free feature screening for ultrahigh dimensional data [J]. *J Amer Statist Assoc*, 2011, **106**(**496**): 1464–1475.

[33] CUI H J, LI R Z, ZHONG W. Model free feature screening for ultrahigh dimensional discriminant analysis [J]. *J Amer Statist Assoc*, 2015, **110**(**510**): 630–641.

# 高维生存数据的删失复合条件分位数筛选

刘薇 [1] 李应求 [2]

[1] 湖南财政经济学院数学与统计学院, 长沙, 410205

[2] 长沙理工大学数学与统计学院, 长沙, 410114

**摘 要:** 本文提出了一种删失复合条件分位数系数 (cCCQC), 用于评估高维删失回归模型中各预测变量的相对重要性. cCCQC 利用了跨分位数的所有有用信息, 能够有效地检测非线性效应, 包括交互作用和异质性. 此外, 基于 cCCQC 的筛选方法对异常值具有鲁棒性, 并具有确定筛选性质. 模拟结果表明, 该方法在高维预测变量的生存数据集中表现良好, 尤其是在变量高度相关的情况下.

**关键词:** 高维生存数据; 删失复合条件分位数系数; 特征筛选; 确定筛选性质; 排序相合性

**中图分类号:** O212.1