Statistica Sinica Preprint No: SS-2016-0402.R2	
Title	Conditional quantile correlation learning for ultrahigh dimensional varying coefficient models and its application in survival analysis
Manuscript ID	SS-2016-0402.R2
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202016.0402
Complete List of Authors	Xiaochao Xia Jialiang Li and Bo Fu
Corresponding Author	Xiaochao Xia
E-mail	xia_xiao_chao@126.com
Notice: Accepted version subject to English editing.	

Statistica Sinica: Newly accepted Paper
(accepted version subject to English editing)

Conditional quantile correlation learning for ultrahigh dimensional varying coefficient models and its application in survival analysis

Xiaochao Xia[1*], Jialiang Li[2,3,4], Bo Fu[5,6]

[1]*College of Science, Huazhong Agricultural University, Wuhan, China*

[2]*Department of Statistics and Applied Probability, National University of Singapore*

[3]*Duke-NUS Graduate Medical School*

[4]*Singapore Eye Research Institute*

[5]*School of Data Science, Fudan University, Shanghai, China*

[6]*Administrative Data Research Centre for England, University College London, UK*

Abstract

In this paper, we consider a robust approach to the ultrahigh dimensional variable screening under varying coefficient models. Different from the existing works focusing on the mean regression function, we propose a novel procedure based on the conditional quantile correlation sure independent screening (CQCSIS). This new proposal is applicable to heterogeneous or heavy-tailed data in general and is invariant to monotone transformation of the response. Furthermore, we generalize such a screening procedure to address censored lifetime data through inverse probability weighting. The CQCSIS can be easily implemented due to an application of nonparametric B-spline approxi-

mation, and computed much faster than the kernel based screening method. Under some regularity conditions, we establish sure screening properties including screening consistency and ranking consistency for proposed approaches. In this paper we also attempt to construct a two-stage variable selection procedure for a further improvement of performance of CQCSIS based on a group SCAD penalization. Extensive simulation examples and real data applications are presented for illustration.

Keywords: Robust ultrahigh dimensional screening; Conditional quantile correlation; Survival data analysis.

1 Introduction

We consider the varying coefficient model

$$Y = \beta_0(T) + \beta_1(T)X_1 + \cdots + \beta_p(T)X_p + \varepsilon, \tag{1}$$

where Y is the response variable, $X_j, j = 1, \ldots, p$ are the centred predictors, $\beta_j(\cdot), j = 0, 1, \ldots, p$ are unknown coefficient functions, T is an index variable, and ε is the error. Over the past two decades, model (1) has been systematically studied and extensively applied in economics, finance, health sciences, social sciences, among others, because it enjoys appealing flexibility of nonparametric models to capture the dynamic impacts of the response on relevant covariates, inherits good interpretability of linear models as well as avoids the curse of dimensionality. We refer to [1] for a comprehensive review of the methodology and theory of varying coefficient models via local polynomial smoothing.

With the rapid development of information technology and data science, much attention has been paid upon identifying the truly significant features or signals. Variable selection plays a vital role to this end. Specifically, under model (1), many penalized variable selection procedures have been documented in the last two decades, including for example the

adaptive Lasso [2] and the SCAD [3, 4]. However, these methods may be challenging in terms of estimation accuracy and computational stability when the dimension of the feature space is extremely large. For example, in real data analysis in Section 5, the number of predictors is as high as hundreds of thousands while the number of observations is only hundreds. Extracting most predictive information from such an extremely large number of candidate variables is a common research goal. Following the pioneering research work of [5], a sure independent screening (SIS) step is now commonly adopted as a necessary preliminary learning for ultrahigh dimensional data prior to the penalized step.

Recently, many excellent variable screening methods for nonparametric models especially for varying coefficient models were presented in the literature [6–10, 31]. We highlight a few relevant works for the marginal varying coefficient model

$$Y = b_{0j}(T) + b_{1j}(T)X_j + \eta, \quad j = 1, \cdots, p, \tag{2}$$

where b_0 and b_1 are intercept and slope functions. [9] used the norm of the slope function defined as

$$u_j = \mathbb{E}[(b_{1j}(T))^2], \tag{3}$$

to screen variables for longitudinal data. [7] proposed the following quantity

$$u_j = \mathbb{E}[(b_{0j}(T) + b_{1j}(T)X_j)^2] - \mathbb{E}[(b_0(T))^2] \tag{4}$$

as a screener, where $b_0(T) = \mathbb{E}[Y|T]$, and they showed that (4) is equivalent to

$$u_j = \mathbb{E}\left[\frac{(\mathrm{Cov}(Y, X_j|T))^2}{\mathrm{Var}(X_j|T)}\right]. \tag{5}$$

Slightly different from (5), [10] proposed

$$u_j = \mathbb{E}\left[\frac{(\mathrm{Cov}(Y, X_j|T))^2}{\mathrm{Var}(X_j|T)\mathrm{Var}(Y|T)}\right] \tag{6}$$

based on the conditional correlation learning (CC-SIS). Note that both [9] and [7] considered B-spline approximation for the coefficient functions, while [10] used kernel smoothing technique. In these works, the SIS properties were rigorously established. In another word, all the aforementioned approaches can successfully pick out a small subset of variables that contains all truly active variables with an overwhelming probability.

The above screening approaches for varying coefficient models, however, might perform unsatisfactorily when the data is heteroscedastic or heavy-tailed, because the spirit of their methods is oriented from mean regression and they are not robust in the presence of outliers. Heterogeneous data may be quite common in many scientific investigations. A well-known solution is the quantile regression technique [11]. For ultrahigh dimensional data, [12] considered the feature screening problem based on quantile regression and developed a non-parametric screening procedure. [13] proposed a conditional quantile screening procedure. Instead of using quantile method, [15] proposed to use a fused Kolmogorov filter for variable screening, which incorporates continuous, discrete and categorical variables. Moreover, we have noted that [15] recently studied the screening utility based on the quantile correlation originally introduced by [16]. However, none of these robust approaches took into account the varying effects of covariates on the response. Partially motivated by this, the current paper aims to work out a robust screening procedure for the varying coefficient model.

We notice that there are some recent works on ultrahigh dimensional survival analysis. For example, [17] proposed a principled sure independence screening procedure under Cox models. In order to deal with ultrahigh dimensional and heterogeneous survival data, [18] proposed rank-based independent screening method for survival data via weighted

rank correlation. Using quantile regression technique, [12] proposed the inverse probability weighted approach to deal with censoring data and [13] proposed censored conditional quantile screening, which concentrated on the technique of redistribution of mass for censored observations. Our proposal in this paper can provide a new solution to survival screening and the performance is shown to be competitive with the existing approaches in our numerical analysis.

Different from earlier works, this paper has made several innovative contributions, summarized as follows. Firstly, we propose a screening method for ultrahigh dimensional varying coefficient models, which can be applicable to dealing with the heteroscedastic or heavy-tailed data. Our screening procedure can be easily implemented and fast since (i) the estimated utility merely involves fitting four univariate nonparametric regression functions based on B-spline approximation, which can be easily implemented in statistic software `R` using `bs()` function, and (ii) our B-spline based approach has computational cost of $O(pnL_n)$ operations, where L_n is the number of spline basis functions and n is sample size, much lower than $O(pn^2)$ operations by the kernel based approach (see [10]). Our proposed utility is invariant to transformation of the response because of the nature of conditional quantile correlation. Secondly, we extended our approach to handle ultrahigh dimensional survival data. This might be quite appealing in survival data analysis to allow for the presence of varying coefficient effect. For example, in the breast cancer data set analyzed in this paper, it could be more reasonable to examine genetic effects as a function of patients' age. Thirdly, under mild technical conditions, we provided the theoretical justification for the screening approaches, that is, our approaches can achieve SIS property. Compared to the nonparametric independent screening (NIS) method by [7], our method can handle data with higher order of dimensionality. Fourthly, we presented a two-stage approaches to refine proposed screening methods, where group penalized variable selection procedures based on quantile regression models are adapted. Such additional step enhances the practical performance of

our program and immediately leads to a broader range of applications.

The remaining of the paper is organized as follows. In Section 2, a general screening approach based on conditional quantile correlation learning is introduced. Technical conditions needed are listed and asymptotic properties are established. In Section 3, an extension to censored response data is developed and related theoretical properties are established. Section 4 provides two-stage variable selection procedures. Numerical studies and empirical analysis of real datasets are carried out in Section 5. Concluding remarks are given in Section 6. All the proofs of main results are relegated to the Appendix.

2 Varying-coefficient Conditional Quantile Correlation Screening

2.1 Screening Methods

In this section, we introduce an SIS procedure based on conditional quantile correlation. [16] proposed a quantile correlation for autoregression modeling, which is defined by

$$\mathrm{qcor}_\tau(Y,X)=\frac{\mathrm{qcov}_\tau(Y,X)}{\sqrt{\mathrm{Var}(I(Y-Q_{\tau,Y}>0))\mathrm{Var}(X)}}, \tag{7}$$

where $\mathrm{qcov}_\tau(Y,X)=\mathrm{Cov}(I(Y-Q_{\tau,Y}>0),X)=\mathbb{E}[\psi_\tau(Y-Q_{\tau,Y})(X-\mathbb{E}(X))]$, $Q_{\tau,Y}$ is the τ quantile of Y, and $\psi_\tau(u)=\tau-I(u<0)$ for $\tau\in(0,1)$. The above correlation takes a value between -1 and 1 and it is asymmetric with respect to Y and X. This definition is different from the classic correlation and possesses the property of monotone invariance in variable Y. Also, as shown by [16], $\mathrm{qcor}_\tau(Y,X)$ is closely related to the slope of the τth quantile regression of Y on X. Specifically, denote by $(a^*_{0\tau},a^*_{1\tau})$ the minimizer of $\mathbb{E}\{\rho_\tau(Y-a_{0\tau}-a_{1\tau}X)\}$ with respect to $a_{0\tau}$ and $a_{1\tau}$. Then we can show that $\mathrm{qcor}_\tau(Y,X)=\varphi(a^*_{1\tau})$, where $\varphi(\cdot)$ is a

continuous and increasing function, and $\varphi(a_{1\tau}^*) = 0$ if and only if $a_{1\tau}^* = 0$.

Following equation (7), we define a conditional quantile correlation (CQC) for Y and X_j given T by

$$\text{cqcor}_\tau(Y, X_j|T) = \frac{\text{qcov}_\tau(Y, X_j|T)}{\sqrt{\text{Var}(I(Y - Q_{\tau,Y} > 0)|T)\text{Var}(X_j|T)}}, \tag{8}$$

where $\text{qcov}_\tau(Y, X_j|T) = \text{Cov}(I(Y - Q_{\tau,Y} > 0), X_j|T)$. Next, we propose the following utility as a new screener

$$u_j = \mathbb{E}\{[\text{cqcor}_\tau(Y, X_j|T)]^2\}. \tag{9}$$

In the following presentation, we denote $m_{1j}(t) = \mathbb{E}\{I(Y - Q_{\tau,Y} > 0)X_j|T = t\}, m_{2j}(t) = \mathbb{E}\{I(Y - Q_{\tau,Y} > 0)|T = t\}, m_{3j} = \mathbb{E}\{X_j^2|T = t\}$ and $m_{4j}(t) = \mathbb{E}\{X_j|T = t\}$ and let $\rho_j(t) = \text{cqcor}_\tau(Y, X_j|T = t)$. Thus, (9) becomes $u_j = \mathbb{E}\{\rho_j^2(T)\}$, where

$$\rho_j(t) = \frac{m_{1j}(t) - m_{2j}(t)m_{4j}(t)}{\sqrt{\{m_{2j}(t) - m_{2j}^2(t)\}\{m_{3j}(t) - m_{4j}^2(t)\}}}.$$

We may now construct a counterpart of u_j based on a sample consisting of observations $\{Y_i, \mathbf{X}_i, T_i, i = 1, \ldots, n\}$. An empirical utility is naturally constructed as

$$\widehat{u}_j = \frac{1}{n}\sum_{i=1}^{n} \widehat{\rho}_j^2(T_i), \tag{10}$$

where

$$\widehat{\rho}_j(t) = \frac{\widehat{m}_{1j}(t) - \widehat{m}_{2j}(t)\widehat{m}_{4j}(t)}{\sqrt{\{\widehat{m}_{2j}(t) - \widehat{m}_{2j}^2(t)\}\{\widehat{m}_{3j}(t) - \widehat{m}_{4j}^2(t)\}}}, \tag{11}$$

where $\widehat{m}_{kj}(t)$'s are nonparametric estimators of $m_{kj}(t)$ for $k = 1, 2, 3, 4$. In practice these

functions can be estimated via local kernel smoothing or other nonparametric approximation methods. In what follows, we consider the B-spline basis approximation to obtain $\widehat{m}_{kj}(t)$. Note that due to the existence of B-spline approximation error, the estimate of CQC, $\widehat{\rho}_j(t)$, does not enjoy asymptotic normality, which is different from [16].

Specifically, let $\widehat{Q}_{\tau,Y}$ be the sample τth quantile of Y and denote $m(t) = \mathbb{E}\{g(X)|T = t\}$ for any generic function g. Suppose that $\{B_k(\cdot), k = 1, \cdots, L_n\}$ with $\|B_k\|_\infty \leq 1$ are a sequence of normalized B-spline basis functions, where L_n is the number of knots. Then, according to the theory of B-spline approximation, there exists a vector $\boldsymbol{\gamma} \in \mathrm{R}^{L_n}$ such that $m(t) \approx \mathbf{B}(t)'\boldsymbol{\gamma}$, where $\mathbf{B}(\cdot) = (B_1(\cdot), \cdots, B_{L_n}(\cdot))'$. Then, based on sample observations $\{(T_i, g(X_i)), i = 1, \ldots, n\}$, one can obtain an estimator for $\boldsymbol{\gamma}$ to be

$$\widehat{\boldsymbol{\gamma}} = (\mathbf{B}'\mathbf{B})^{-1}\mathbf{B}'\mathbf{f}, \tag{12}$$

using the least squares method, where $\mathbf{B} = (\mathbf{B}(T_1), \cdots, \mathbf{B}(T_n))'$ and $\mathbf{f} = (g(X_1), \cdots, g(X_n))'$. Hence, the estimator for $m(t)$ is defined by

$$\widehat{m}(t) = \widehat{\mathbb{E}}\{g(X)|T = t\} = \mathbf{B}(t)'(\mathbf{B}'\mathbf{B})^{-1}\mathbf{B}'\mathbf{f} \tag{13}$$

Using such an idea, we can obtain a simple estimator for $\widehat{u}_j$ in (10) with

$$\widehat{m}_{kj}(t) = \mathbf{B}(t)'(\mathbf{B}'\mathbf{B})^{-1}\mathbf{B}'\mathbf{f}_{kj}, \qquad k = 1, \cdots, 4,$$

where $\mathbf{f}_{1j} = (I(Y_1 - \widehat{Q}_{\tau,Y} > 0)X_{1j}, \cdots, I(Y_n - \widehat{Q}_{\tau,Y} > 0)X_{nj})'$, $\mathbf{f}_{2j} = (I(Y_1 - \widehat{Q}_{\tau,Y} > 0), \cdots, I(Y_n - \widehat{Q}_{\tau,Y} > 0))'$, $\mathbf{f}_{3j} = (X_{1j}^2, \cdots, X_{nj}^2)'$ and $\mathbf{f}_{4j} = (X_{1j}, \cdots, X_{nj})'$. Then, we select the following set of variables

$$\widehat{\mathcal{M}} = \{j : \widehat{u}_j > \nu_n, 1 \leq j \leq p\}, \tag{14}$$

where ν_n is a user-specified threshold parameter.

2.2 Theoretical Properties

In order to study the theoretical properties of proposed screening procedure, we denote by $\mathcal{M}_* = \{j : \beta_j(t) \neq 0 \text{ for some } t \in \mathcal{T}\}$ the set of truly active variables, with nonsparsity size $s_n = |\mathcal{M}_*|$. We impose the following regularity conditions, which might not be the weakest but facilitate to establish the screening consistency of the proposed CQC screener.

(C1) Assume that the support of index variable T is bounded and denoted by $\mathcal{T} = [a, b]$ with finite constants a and b, on which its density f is bounded away from zero and infinity, that is there exist two positive constants M_1, M_2 such that $0 < M_1 \leq \inf_{t\in\mathcal{T}} f(t) \leq \sup_{t\in\mathcal{T}} f(t) \leq M_2 < \infty$.

(C2) For all $j = 1, \ldots, p$, there exist positive constants K_1, K_2 such that $P(|X_j| > x|T) \leq K_1 \exp(-K_2^{-1}x)$ almost surely.

(C3) The functions $m_{kj}, k = 1, 2, 3, 4, j = 1, \ldots, p$ belong to a class of functions $\mathcal{B}$, where the rth derivative $m^{(r)}$ of any class member m exists and is Lipschitz of order α. That is,

$$\mathcal{B} = \{m(\cdot) : |m^{(r)}(s) - m^{(r)}(t)| \leq M|s - t|^\alpha \text{ for } s, t \in \mathcal{T}\},$$

for some positive constant M, where r is a nonnegative integer and $\alpha \in (0, 1]$ such that $d \equiv r + \alpha > 0.5$.

(C4) In a neighbourhood of $Q_{\tau,Y}$, conditional densities $f_{Y|(X_j,T)}(y)$ of Y given (X_j, T) and $f_{Y|T}(y)$ of Y given T are uniformly bounded away from zero and infinity and their derivatives $f'_{Y|(X_j,T)}(y)$ and $f'_{Y|T}(y)$ are bounded.

(C5) There exist positive constants K_3, K_4 such that $\inf_{t\in\mathcal{T}} \mathrm{Var}(I(Y > Q_{\tau,Y})|t) \geq K_3 > 0$ and $\inf_{t\in\mathcal{T}} \mathrm{Var}(X_j|t) \geq K_4 > 0$.

(C6) $\min_{j\in\mathcal{M}_*} u_j \geq 2CL_n n^{-2\kappa}$ for some $\kappa > 0$ and $C > 0$.

(C7) $\lim_{n\to\infty} n^{2\kappa} L_n^{-1/2-d} = 0$ and $\lim_{n\to\infty} n^{2\kappa-\iota} L_n^{-1/2} = 0$ for some $0 < \iota < 1/2$, where d is defined in (C3) and κ is given in (C6).

Remark 1: Conditions (C1), (C2) and (C4) are mild distribution assumptions. Condition (C2) requires the conditional sub-exponential tail probability for covariates X_j given T uniformly in j, which guarantees that $m_{kj}(t), k = 1, 2, 3, 4$ are finite uniformly in $t \in \mathcal{T}$. This condition can be weakened by adding more constraints on the dimensionality p. Nevertheless, the sure screening property established below still holds and can be proved with slightly different technical arguments. Condition (C3) is the regularity condition for the smoothness of coefficient functions, which facilitates the B-spline approximation. Condition (C5) requires that the CQC is well defined. Condition (C6) reflects that the significant covariates is identifiable by marginal models, which can be revealed by partial orthogonality condition [19]. That is, $\{X_j : j \in \mathcal{M}_*\}$ is independent of $\{X_j : j \notin \mathcal{M}_*\}$. Condition (C7) bounds the number of basis functions L_n from below, which implies that L_n should not be chosen too small in order to ensure that the approximation error is negligible. Similar requirements can be found in [6, 7, 9] for screening in ultra-high dimensional varying coefficient models.

Theorem 2.1. *(Sure Screening Property) Under conditions (C1)-(C5) and (C7),*
(i) if $L_n^{-3} n \to \infty$ and $L_n^{-2} n^{1-4\kappa} \to \infty$ as $n \to \infty$, then there exist positive constants δ_1, δ_2 such that

$$P\Big(\max_{1\leq j\leq p} |\widehat{u}_j - u_j| > CL_n n^{-2\kappa}\Big)$$
$$\leq O\big(pn\{L_n^2 \exp(-\delta_1 L_n^{-3} n) + L_n \exp(-\delta_2 L_n^{-2} n^{1-4\kappa})\}\big);$$

(ii) if condition (C6) is further satisfied, then by taking $\nu_n = CL_n n^{-2\kappa}$, *we have*

$$P\Big(\mathcal{M}_* \subset \widehat{\mathcal{M}}\Big) \geq 1 - O\big(s_n n\{L_n^2 \exp(-\delta_1 L_n^{-3} n) + L_n \exp(-\delta_2 L_n^{-2} n^{1-4\kappa})\}\big);$$

and (iii) under conditions of (ii) and the condition that $\max_{j\notin\mathcal{M}_*} u_j = o(L_n n^{-2\kappa})$, *we have*

$$P(\widehat{\mathcal{M}} = \mathcal{M}_*) = 1 - o(1).$$

Remark 2: Theorem 2.1 suggests that we can handle NP dimensionality of order

$$\log p = o\big(L_n^{-3} n + L_n^{-2} n^{1-4\kappa}\big).$$

In comparison with [7], our proposed CQC screening procedure achieves a higher exponential rate for the dimensionality under similar conditions. This can be partly explained by the use of indicator function in the proposed utility. Note that if we take $L_n = O(n^{1/(2d+1)})$, i.e. the optimal convergence rate in nonparametric regression [20], then condition (C7) reduces to $\kappa < \min(\frac{\iota}{2} + \frac{1}{4(2d+1)}, \frac{1}{4})$. Accordingly, if $\frac{1}{4(2d+1)} < \kappa < \min(\frac{\iota}{2} + \frac{1}{4(2d+1)}, \frac{1}{4})$, the dimensionality we can handle is as high as $\log p = o\big(n^{\frac{2(d-1)}{2d+1}}\big)$, which is the same order as in [7]. Moreover, if $\kappa \leq \frac{1}{4(2d+1)}$, then we can deal with the dimensionality of order $\log p = o\big(n^{\frac{2d-1}{2d+1}-4\kappa}\big)$, provided that $d > \max(\frac{1+4\kappa}{2(1-4\kappa)}, 1)$ in order to guarantee the consistency of screening procedure.

Remark 3: Theorem 2.1(i) together with conditions of Theorem 2.1(iii) implies that with probability tending to one, $\max_{j\notin\mathcal{M}_*} \widehat{u}_j < cL_n n^{-2\kappa}$ for any $c > 0$. Thus, by choosing $\nu_n = cL_n n^{-2\kappa}$, we can prove model selection consistency.

Theorem 2.2. *(Ranking Consistency Property) Under conditions (C1)-(C7), suppose that*

$$\liminf_{n\to\infty} \Big\{ \min_{j\in\mathcal{M}_*} u_j - \max_{j\notin\mathcal{M}_*} u_j \Big\} > 0 \tag{15}$$

and that $\log p < C_{11}\delta_0^2 L_n^{-4} n - 2\log L_n - \log n$, *where* C_{11}, δ_0 *are constants defined in the Appendix, then we have*

$$\liminf_{n\to\infty} \left\{ \min_{j\in\mathcal{M}_*} \widehat{u}_j - \max_{j\notin\mathcal{M}_*} \widehat{u}_j \right\} > 0$$

in probability.

Remark 4: Theorem 2.2 indicates that the true significant variables have an overwhelming probability of greater $\widehat{u}_j$ than non-informative variables, and hence implies that all important predictors are ranked in the top.

Remark 5: Assumption (15) requires a clear separation between the CQC of signal predictors and noisy predictors. It seems that such a condition may not be easily satisfied for all high dimensional models. When this data assumption is not available, the results from Theorem 2.2 may not hold and the theorem may be of only theoretical interest.

Let $\mathbf{b} = (\mathrm{Cov}\{I(Y > Q_{\tau,Y}), X_1|T\}, \cdots, \mathrm{Cov}\{I(Y > Q_{\tau,Y}), X_p|T\})'$. The following property says that if $\mathbb{E}\{\|\mathbf{b}\|^2\} = O(n^\gamma)$ for some $\gamma > 0$, the model after screening is of polynomial size with probability tending to one. When the predictors are weakly correlated or independent and the number of active predictors, s_n, is of polynomial size, the vector is quite sparse with s_n nonzero entries. Under such a setting, the condition imposed on $\mathbf{b}$ is valid. Otherwise the following theorem may fail for highly correlated regressors.

Theorem 2.3. *(False Selection Rate) Under conditions of Theorem 2.1, there exist positive constants* $\delta_3, \delta_4, \widetilde{C}$ *such that*

$$P\big(|\widehat{\mathcal{M}}| \le \widetilde{C} n^{2\kappa} L_n^{-1} \mathbb{E}\{\|\mathbf{b}\|^2\}\big) \ge 1 - O\big(pn\{L_n^2 \exp(-\delta_3 L_n^{-3} n) + L_n \exp(-\delta_4 L_n^{-2} n^{1-4\kappa})\}\big).$$

3 CQC Screening for Survival Data

In this section, we extend the CQC screening procedure to handle ultrahigh dimensional survival data under varying coefficient model. Suppose that we observe the data $\{\widetilde{Y}_i, \Delta_i, \mathbf{X}_i = (X_{i1}, \cdots, X_{ip})', T_i; i = 1, \ldots, n\}$, consisting of n independent copies of $(\widetilde{Y}, \Delta, \mathbf{X}, T)$, where $\widetilde{Y} = \min(Y, Z)$ and $\Delta = I(Y \leq Z)$, in which Y represents the failure time variable and Z stands for the censoring time. In this paper, we assume that censoring variable Z is independent of covariates.

From equation (8), it is easy to see that

$$\text{cqcor}_\tau(Y, X_j|T) = \frac{\mathbb{E}\{\psi_\tau(Y - Q_{\tau,Y})X_j|T\} - \mathbb{E}\{\psi_\tau(Y - Q_{\tau,Y})|T\}\mathbb{E}\{X_j|T\}}{\sqrt{\text{Var}\{\psi_\tau(Y - Q_{\tau,Y})|T\}\text{Var}(X_j|T)}}$$

Then, motivated by [21], we define a novel weight-nested version of CQC as

$$\text{cqcor}_{\tau,w}(\widetilde{Y}, X_j|T) = \frac{\mathbb{E}\{\psi_{\tau,w}(\widetilde{Y} - Q_{\tau,Y})X_j|T\} - \mathbb{E}\{\psi_{\tau,w}(\widetilde{Y} - Q_{\tau,Y})|T\}\mathbb{E}\{X_j|T\}}{\sqrt{\text{Var}(\psi_{\tau,w}(\widetilde{Y} - Q_{\tau,Y})|T)\text{Var}(X_j|T)}} \tag{16}$$

where $\psi_{\tau,w}(v) = \tau - w(F)I(v < 0)$ with $1 - F(y) = P(Y > y)$ being survival distribution, and

$$w(F) = \begin{cases} 1, & \Delta = 1 \text{ or } F(Z) > \tau, \\ \frac{\tau - F(Z)}{1 - F(Z)}, & \Delta = 0 \text{ and } F(Z) < \tau, \end{cases}$$

which is a weight function that redistributes the masses of censored observations to the right. To understand the construction of $w(F)$, we remark that $I(Y_i - Q_{\tau,Y} < 0)$ is observed if the observation is uncensored and equals 0 if $\widetilde{Y}_i = Z_i > Q_{\tau,Y}$. The uncertain situation is that $\Delta_i = 0$ and $Z_i < Q_{\tau,Y}$, in which case, $E\{I(Y - Q_{\tau,Y} < 0)|Y_i > Z_i\} = \frac{\tau - F(Z)}{1 - F(Z)}$. Thus, we assign the weight 1 to the observed data, while, in the ambiguous case, we distribute the weight $\frac{\tau - F(Z)}{1 - F(Z)}$ to the "pseudo" observation at Z_i. This weight function does not affect the

quantile fit. The redistribution-of-mass idea was first introduced by [22] and incorporated for quantile regression by [21]. Clearly, when data are completely observed, i.e., $\Delta_i = 1$ for all i, the above correlation reduces to that defined in (8). Hence, we can introduce the following utility for CQC screening

$$u_{j,w} = \mathbb{E}\{[\text{cqcor}_{\tau,w}(\widetilde{Y}, X_j|T)]^2\}.$$

Let $m_{1j,w}(t) = \mathbb{E}\{w(F)I(\widetilde{Y} < Q_{\tau,Y})X_j|T = t\}$, $m_{2j,w}(t) = \mathbb{E}\{w(F)I(\widetilde{Y} < Q_{\tau,Y})|T = t\}$ and $m_{3j,w}(t) = \mathbb{E}\{w^2(F)I(\widetilde{Y} < Q_{\tau,Y})|T = t\}$ and denote $\rho_{j,w}(t) = -\text{cqcor}_{\tau,w}(Y, X_j|T = t)$. Then, we have $u_{j,w} = \mathbb{E}\{\rho^2_{j,w}(T)\}$, where

$$\rho_{j,w}(t) = \frac{m_{1j,w}(t) - m_{2j,w}(t)m_{4j}(t)}{\sqrt{\{m_{3j,w}(t) - m^2_{2j,w}(t)\}\{m_{3j}(t) - m^2_{4j}(t)\}}}.$$

Let $\widehat{F}(y)$ be the Kaplan-Meier estimator of $F(y)$ based on $\{(\widetilde{Y}_i, \Delta_i), i = 1, \ldots, n\}$ and $\widehat{Q}_{\tau,Y}$ be the sample τth quantile $\widehat{F}^{-1}(\tau)$, that is an estimator of $Q_{\tau,Y}$ when Y is subject to right censoring. We define an empirical version of $u_{j,w}$ as

$$\widehat{u}_{j,w} = \frac{1}{n}\sum_{i=1}^{n} \widehat{\rho}^2_{j,w}(T_i), \tag{17}$$

with

$$\widehat{\rho}_{j,w}(T_i) = \frac{\widehat{m}_{1j,w}(T_i) - \widehat{m}_{2j,w}(T_i)\widehat{m}_{4j}(T_i)}{\sqrt{[\widehat{m}_{3j,w}(T_i) - \widehat{m}^2_{2j,w}(T_i)][\widehat{m}_{3j}(T_i) - \widehat{m}^2_{4j}(T_i)]}},$$

where

$$\widehat{m}_{kj,w}(T_i) = \mathbf{B}(t)'(\mathbf{B}'\mathbf{B})^{-1}\mathbf{B}'\mathbf{f}_{kj,w}, \qquad k = 1, 2, 3,$$

and $\mathbf{f}_{1j,w} = (w_1(\widehat{F})I(\widetilde{Y}_1 < \widehat{Q}_{\tau,Y})X_{1j}, \cdots, w_n(\widehat{F})I(\widetilde{Y}_n < \widehat{Q}_{\tau,Y})X_{nj})'$, $\mathbf{f}_{2j,w} = (w_1(\widehat{F})I(\widetilde{Y}_1 < \widehat{Q}_{\tau,Y}), \cdots, w_n(\widehat{F})I(\widetilde{Y}_n < \widehat{Q}_{\tau,Y}))'$ and $\mathbf{f}_{3j,w} = (w_1^2(\widehat{F})I(\widetilde{Y}_1 < \widehat{Q}_{\tau,Y}), \cdots, w_n^2(\widehat{F})I(\widetilde{Y}_n < \widehat{Q}_{\tau,Y}))'$. Then, we select a subset of variables

$$\widehat{\mathcal{N}} = \{j : \widehat{u}_{j,w} \geq \varsigma_n, 1 \leq j \leq p\}, \tag{18}$$

where ς_n is a pre-specified threshold parameter.

To establish the sure independent screening properties, we further introduce the following regularity conditions, which are standard in censored quantile regression (e.g. see [13, 21]).

(D1) In a neighbourhood of $Q_{\tau,Y}$, $F(y)$ is twice differentiable, the density $f_Y(y)$ and the conditional densities $f_{Y|(X_j,T)}(y)$ and $f_{Y|T}(y)$ are uniformly bounded away from zero and infinity, and their first derivatives $f'_{Y|(X_j,T)}(y)$ and $f'_{Y|T}(y)$ are bounded uniformly.

(D2) In a neighbourhood of $Q_{\tau,Y}$, the conditional densities $h_{Z|(X_j,T)}(z)$ and $h_{Z|T}(z)$ are uniformly bounded away from zero and infinity, and their first derivatives $h'_{Z|(X_j,T)}(z)$ and $h'_{Z|T}(z)$ are bounded uniformly.

(D3) Assume that $P(Y \leq \Lambda_s) > \tau > 0$, where Λ_s represents the end time of the study.

(D4) There exist positive constants K_5, K_6 such that $\inf_{t\in\mathcal{T}} \mathrm{Var}(w(F)I(Y < Q_{\tau,Y})|t) \geq K_5 > 0$ and $\inf_{t\in\mathcal{T}} \mathrm{Var}(X_j|t) \geq K_6 > 0$.

(D5) $\min_{j\in\mathcal{M}_*} u_{j,w} \geq 2C_w L_n n^{-2\kappa}$ for some $\kappa > 0$ and $C_w > 0$.

Theorem 3.1. *(Sure Screening Property) Under conditions (C1)-(C3), (C7) and (D1)-(D4), (i) if $L_n^{-3}n \to \infty$ and $L_n^{-2}n^{1-4\kappa} \to \infty$ as $n \to \infty$, then there exist positive constants δ_5, δ_6 such that*

$$\begin{aligned} &P\Big(\max_{1\leq j\leq p} \big|\widehat{u}_{j,w} - u_{j,w}\big| > C_w L_n n^{-2\kappa}\Big) \\ &\leq O\big(pn\{L_n^2 \exp(-\delta_5 L_n^{-3}n) + L_n \exp(-\delta_6 L_n^{-2}n^{1-4\kappa})\}\big); \end{aligned}$$

(ii) if condition (D5) is further satisfied, then by taking $\varsigma_n = C_w L_n n^{-2\kappa}$, *we have*

$$P\Big(\mathcal{M}_* \subset \widehat{\mathcal{N}}\Big) \geq 1 - O\big(s_n n\{L_n^2 \exp(-\delta_5 L_n^{-3} n) + L_n \exp(-\delta_6 L_n^{-2} n^{1-4\kappa})\}\big).$$

Theorem 3.2. *(Ranking Consistency Property) Under conditions (C1)-(C3), (C7) and (D1)-(D5), suppose that*

$$\liminf_{n\to\infty} \Big\{ \min_{j\in\mathcal{M}_*} u_{j,w} - \max_{j\notin\mathcal{M}_*} u_{j,w} \Big\} > 0 \tag{19}$$

and that $\log p = o(L_n^{-4} n)$, *then we have*

$$\liminf_{n\to\infty} \Big\{ \min_{j\in\mathcal{M}_*} \widehat{u}_{j,w} - \max_{j\notin\mathcal{M}_*} \widehat{u}_{j,w} \Big\} > 0$$

in probability.

Let $\mathbf{b}_w = (\mathrm{Cov}\{w(F)I(\widetilde{Y} < Q_{\tau,Y}), X_1|T\}, \cdots, \mathrm{Cov}\{w(F)I(\widetilde{Y} < Q_{\tau,Y}), X_p|T\})'$. The following property indicates that if $\mathbb{E}\{\|\mathbf{b}_w\|^2\} = O(n^{\gamma'})$ for some $\gamma' > 0$, the model after screening is of polynomial size with probability tending to one.

Theorem 3.3. *(False Selection Rate) Under conditions of Theorem 3.1, there exist positive constants* $\delta_7, \delta_8, \widetilde{C}$ *such that*

$$P\big(|\widehat{\mathcal{N}}| \leq \widetilde{C} n^{2\kappa} L_n^{-1} \mathbb{E}\{\|\mathbf{b}_w\|^2\}\big) \geq 1 - O\big(pn\{L_n^2 \exp(-\delta_7 L_n^{-3} n) + L_n \exp(-\delta_8 L_n^{-2} n^{1-4\kappa})\}\big).$$

4 Two-Stage Approaches

It has been well known that results from a single SIS procedure are rather crude (see [5, 7, 9, 10]). In implementation, we don't directly determine threshold parameters ν_n in (14) and ς_n in (18) when carrying out an SIS procedure. Instead, we usually select the first d_n

predictors in the top ranked list as important variables after screening. By this, we can see that a large d_n corresponds to small ν_n and ς_n and vice verse. From now on, with a bit abuse of notation, we use d_n, slightly different from d defined in Condition (C3), to denote the size of the screened model, where the subscript n in d_n is used to stress the dependence on sample size. In practice, despite carrying out a SIS procedure can substantially reduce the ultrahigh dimensionality with a specified d_n, a large d_n would yield a large model that inevitably includes some irrelevant variables. Many papers have proposed efficient data-driven approaches to select d_n. For example, [17] proposed a principled selection method by controlling the false positive rate; [18] developed a technique based on multiple testing. However, these approaches cannot guarantee that the selected set is exactly the same as the truly active set. To this end, in what follows, we propose a two-stage approach for variable selection. Similar application has appeared in the research paper of [10]. Simply speaking, we conduct a CQCSIS in the first stage and continue with a group penalized variable selection in the second stage.

Denote by $\mathbf{X} = (X_1, \ldots, X_q)'$ the vector consisting of the retained variables after screening. Still applying the B-spline basis approximation, we write $\boldsymbol{\gamma} = (\gamma_1', \ldots, \gamma_q')'$, $\boldsymbol{\Pi}_i = (\Pi_{i1}', \ldots, \Pi_{iq}')'$ and $\Pi_{ij} = (X_{ij}B_1(T_i), \ldots, X_{ij}B_{L_n}(T_i))'$. For fully observed data, we consider the following group penalized quantile regression:

$$\min_{\boldsymbol{\gamma}} \frac{1}{n} \sum_{i=1}^{n} \rho_\tau(Y_i - \boldsymbol{\Pi}_i'\boldsymbol{\gamma}) + \sum_{j=1}^{q} p_\lambda(\|\gamma_j\|_{\mathrm{B}}), \tag{20}$$

where $\|\gamma_j\|_{\mathrm{B}} = (\gamma_j' \int_0^1 \mathbf{B}(t)\mathbf{B}'(t)\gamma_j)^{1/2}$, $\rho_\tau(u) = u[\tau - I(u < 0)]$, $p_\lambda(\cdot)$ is a nonnegative and nonconcave penalty function such as the SCAD [23] or the MCP [24].

Denote $\widehat{\boldsymbol{\gamma}}$ by the minimizer of (20). Without loss of generality, we let $\beta_j(t), j = 1, \cdots, s$ be the nonzero coefficient function and $\beta_j(t) \equiv 0, j = s+1, \cdots, q$, where q may depend on n. In order to derive the asymptotic theory for $\widehat{\boldsymbol{\gamma}}$, we make the following assumptions, which

are commonly used in quantile regression and similar to those in [3, 4] and [27].

(E1) The conditional density $f_{U|\mathbf{X}}(u|\mathbf{x})$ of U given $\mathbf{X} = \mathbf{x}$ is bounded away from zero and infinity uniformly in u and $\mathbf{x}$.

(E2) There exists a positive constant $\overline{M}$ such that $|X_k| \leq \overline{M}$ for all $1 \leq k \leq q$. The eigenvalues of the matrix $\mathbb{E}\{\mathbf{X}\mathbf{X}'|U = u\}$ are uniformly bounded away from zero and infinity for all u.

(E3) The density $f_\epsilon(\cdot)$ of random error $\epsilon = Y - \mathbf{X}'\boldsymbol{\beta}(T)$ is continuous at 0 and bounded away from zero and infinity.

Proposition 4.1. *Under conditions of Theorem 2.1 and Assumptions (E1)-(E3) and if* $\lambda \to 0$ *and* $\frac{\lambda}{(q/n)^{1/2}L_n} \to \infty$ *as* $n \to \infty$*, then we have*
(i) $\widehat{\beta}_j, j = 1, \cdots, s$ *are nonzero and* $\widehat{\beta}_j = 0, j = s+1, \cdots, q$ *with probability approaching one;*
(ii) $\|\widehat{\beta}_j - \beta_j\|_{L_2} = O_p(\sqrt{L_n/n} + L_n^{-d}), j = 1, \cdots, s.$

Remark 6: The part (i) of this property says that the proposed group penalization is consistent in variable selection, that is it selects relevant covariates and identifies irrelevant covariates with probability tending to one. The part (ii) provides the convergence rate for the estimated nonzero coefficient functions. From part (ii), we can see that $L_n \asymp n^{\frac{1}{2d+1}}$ is the optimal convergence rate. In this case, we have $\|\widehat{\beta}_j - \beta_j\|_{L_2} = O_p(n^{-\frac{d}{2d+1}})$ for penalized varying coefficient quantile regression, which is the same as that for penalized varying coefficient mean regression [3]. The proof of Proposition 4.1 can be finished by following the arguments in [4] and the details are omitted.

It is usually difficult to directly solve the optimization problem (20) because of the nonconvexity. In the sequel, we propose to implement such a nonconvex optimization via a first order approximation. Our algorithm can be viewed as a combination of the local linear

approximation (LLA) by [25] and the algorithm by [26]. We use the Bayesian Information Criterion (BIC) proposed by [27] to obtain the best regularized parameter. The computational details are presented as follows.

Let $\mathbf{H} = \int_0^1 \mathbf{B}(t)\mathbf{B}(t)'\mathrm{d}t$ and decompose $\mathbf{H} = \mathbf{A}'\mathbf{A}$. Denote $\gamma_j^* = \mathbf{A}\gamma_j$, i.e., $\mathbf{A} = \mathbf{H}^{1/2}$ and $\Pi_{ij}^* = \mathbf{A}^{-1}\Pi_{ij}$, an L_n-vector. Thus, $\|\gamma_j\|_{\mathrm{B}} = \|\gamma_j^*\|_2$ and problem (20) becomes

$$\min_{\boldsymbol{\gamma}^*} \frac{1}{n}\sum_{i=1}^{n} \rho_\tau(Y_i - (\boldsymbol{\Pi}_i^*)'\boldsymbol{\gamma}^*) + \sum_{j=1}^{p} p_\lambda(\|\gamma_j^*\|_2). \tag{21}$$

Suppose we have appropriate initial estimates $\gamma_j^{*,init} = \mathbf{A}\gamma_j^{init}, j = 1, \ldots, q$, where γ_j^{init}'s are the initial estimates for the original problem (20). Then we apply Taylor's expansion to the penalty function, $p_\lambda(\|\gamma_j^*\|_2)$, at the point $\gamma_j^{*,init}$, that is,

$$\begin{aligned} p_\lambda(\|\gamma_j^*\|_2) &\approx p_\lambda(\|\gamma_j^{*,init}\|_2) + \sum_{k=1}^{L_n} \frac{p'_\lambda(\|\gamma_j^{*,init}\|_2)}{\|\gamma_j^{*,init}\|_2} |\gamma_{kj}^{*,init}|(|\gamma_{kj}^*| - |\gamma_{kj}^{*,init}|) \\ &= p_\lambda(\|\gamma_j^{*,init}\|_2) - p'_\lambda(\|\gamma_j^{*,init}\|_2)\|\gamma_j^{*,init}\|_2 + \sum_{k=1}^{L_n} \frac{p'_\lambda(\|\gamma_j^{*,init}\|_2)}{\|\gamma_j^{*,init}\|_2} |\gamma_{kj}^{*,init}| \cdot |\gamma_{kj}^*|, \end{aligned}$$

where $p'_\lambda(\cdot)$ represents the derivative of $p_\lambda(\cdot)$. Note that such an approximation may be regarded as a two-step approximation where we first apply the LLA on the penalty function, $p_\lambda(\cdot)$, yielding an ℓ_2 group regularization that can be solved by a second order cone programming [4], and then apply a further approximation for $\|\gamma_j^*\|_2$ as in [26]. Consequently, we convert problem (21) to the following minimization problem

$$\min_{\boldsymbol{\gamma}^*} \sum_{i=1}^{n} \rho_\tau(Y_i - (\boldsymbol{\Pi}_i^*)'\boldsymbol{\gamma}^*) + n\sum_{j=1}^{p}\sum_{k=1}^{L_n} \omega_{\lambda,kj}|\gamma_{kj}^*| \tag{22}$$

where $\omega_{\lambda,kj} = \frac{p'_\lambda(\|\gamma_j^{*,init}\|_2)}{\|\gamma_j^{*,init}\|_2}|\gamma_{kj}^{*,init}|$.

We have to remark that, apparently, (22) is a weighted ℓ_1 regularization for quantile

regression, which encourages sparsity of individual coefficients. The procedure does not yield sparsity of groups of coefficients because the weights assigned to the coefficients within the same group are different, leading to unequal shrinkage for the coefficients within one common group. To address this issue, we modify the above problem (22) as

$$\min_{\boldsymbol{\gamma}^*} \sum_{i=1}^{n} \rho_\tau(Y_i - (\boldsymbol{\Pi}_i^*)'\boldsymbol{\gamma}^*) + n \sum_{j=1}^{p} \widetilde{\omega}_{\lambda,j} \sum_{k=1}^{L_n} |\gamma_{kj}^*| \tag{23}$$

where $\widetilde{\omega}_{\lambda,j} = \frac{p'_\lambda(\|\gamma_j^{*,init}\|_2)}{\|\gamma_j^{*,init}\|_2} \max_{1\le k\le L_n} |\gamma_{kj}^{*,init}|$. The minimization problem (23) can be solved by the following linear programming

$$\begin{aligned} \min_{\{\eta_i^+,\eta_i^-,\gamma_{kj}^{*+},\gamma_{kj}^{*-}\}} & \tau \sum_{i=1}^{n} \eta_i^+ + (1-\tau) \sum_{i=1}^{n} \eta_i^- + n \sum_{j=1}^{p} \sum_{k=1}^{L_n} \widetilde{\omega}_{\lambda,j} \cdot (\gamma_{kj}^{*+} + \gamma_{kj}^{*-}), \\ \text{such that} \quad & \eta_i^+ - \eta_i^- = Y_i - (\boldsymbol{\Pi}_i^*)'(\boldsymbol{\gamma}^{*+} - \boldsymbol{\gamma}^{*-}), i = 1,\ldots,n \\ & \gamma_{kj}^{*+} \ge 0, \gamma_{kj}^{*-} \ge 0, j = 1,\ldots,p; k = 1,\ldots,L_n \\ & \eta_i^+ \ge 0, \eta_i^- \ge 0, i = 1,\ldots,n, \end{aligned} \tag{24}$$

where $\boldsymbol{\gamma}^{*+} = (\gamma_{11}^{*+},\ldots,\gamma_{1L_n}^{*+},\ldots,\gamma_{q1}^{*+},\ldots,\gamma_{qL_n}^{*+})'$, $\boldsymbol{\gamma}^{*-} = (\gamma_{11}^{*-},\ldots,\gamma_{1L_n}^{*-},\ldots,\gamma_{q1}^{*-},\ldots,\gamma_{qL_n}^{*-})'$, and $z^+ = zI(z>0)$ and $z^- = -zI(z<0)$ for any variable z. Denote by $\widehat{\boldsymbol{\gamma}}_\lambda^*$ the solution of above problem. For the selection of tuning parameter, λ, we use the following BIC function:

$$\text{BIC}(\lambda) = \log\Big\{ \sum_{i=1}^{n} \rho_\tau(Y_i - (\boldsymbol{\Pi}_i^*)'\widehat{\boldsymbol{\gamma}}_\lambda^*) \Big\} + df \frac{\log n}{2n} C_n, \tag{25}$$

where df is the number of nonzero entries of $\widehat{\boldsymbol{\gamma}}_\lambda^*$ and C_n is a diverging number, say $\log p$. Such a BIC selector has been demonstrated to be consistent in variable selection for the quantile varying coefficient models [27].

5 Numerical Studies

5.1 Monte Carlo Studies

In this subsection, we conduct simulations to examine the finite sample performance of the proposed CQCSIS. Following [12], we consider two criteria for evaluating the performance, where the first criterion is the minimum model size (MMS), i.e., the smallest number of covariates needed to include all the active variables, and the second is the proportion of all the active variables selected (PS) with the screening threshold parameter being specified as $d_n = \lfloor n/\log n \rfloor$. Throughout this subsection, we adopt the following simulation setup: the sample size $n = 400$, the number of basis $L_n = \lfloor n^{1/5} \rfloor + 1$, the covariate dimension $p = 1000$, and the number of simulations $N = 200$ for each example. With code written in R and run on a PC with Intel(R) Core i5 3.30 GHz processor, an implementation of CQCSIS with 200 sampling for each example takes about 10 minutes. This computation does not represent a substantial cost compared to the costs of data collection and an analysts time. Due to space limit, we here merely provide one simulation example and other examples are given in the online Supplementary Materials.

Example 1. Let $\mathbf{X} = (X_1, \ldots, X_p)'$ be a p dimensional random vector following the multivariate normal distribution with mean zero and covariance matrix $\Sigma = (\sigma_{j,k})_{1 \le j,k \le p}$ where $\sigma_{j,k} = \varrho^{|j-k|}$. Simulate the index variable T from the unit uniform distribution and generate the response as

$$Y = 5TX_1 + 3(2T-1)^2 X_2 + 4\sin(2\pi T)X_3 + \varepsilon,$$

where the error ε is considered to be one of the following cases:

- Case (1a): the error follows the standard normal distribution, i.e., $\varepsilon \sim N(0,1)$;
- Case (1b): the error follows the Cauchy distribution with location zero and scale one,

i.e., $\varepsilon \sim C(0,1)$;

- Case (1c): the error follows the normal distribution with varying variance, that is, $0.5(\frac{\exp(T)}{1+\exp(T)}X_2 + 3(T-1)^2X_4 + \sin(2\pi T)X_5) \cdot (\varepsilon - Q_{\varepsilon,\tau})$ with $\varepsilon \sim N(0,1)$;

- Case (1d): the error follows the scale-varied Cauchy distribution, that is $0.5(\frac{\exp(T)}{1+\exp(T)}X_2 + 3(T-1)^2X_4 + \sin(2\pi T)X_5) \cdot C(0,1)$.

Among the four cases, Cases (1a) and (1b) are thin-tailed and heavy-tailed homoscedastic models respectively, while Cases (1c) and (1d) are heteroscedastic models. In Case (1c), the number of active covariates s_n is 3 at the τth quantile but 5 elsewhere. In Case (1d), the number of active covariates is 3 at the median but 5 elsewhere.

The results including the median of MMS, its robust standard deviation (RSD) and the average of PS out of N simulations for the CQCSIS method proposed in current paper and the nonparametric independent screening (NIS) method proposed by [7] are summarized in Table 1. From Table 1, we can see that our CQCSIS method performs substantially better than the NIS method, especially when the data are heavy-tailed or heteroscedastic. When error is normal, the NIS performs slightly better than the CQCSIS for homoscedastic data, yet both methods have comparable performance. Increasing the correlation among covariates improves the screening performance for all methods. This is evidenced in other studies as well and can be explained by the fact that the sets of jointly correlated markers may be relatively more distinguishable than uncorrelated ones.

5.2 Real Data Analysis

5.2.1 Hospital Episode Statistics Data

We applied the CQCSIS method proposed in Section 2 to Hospital Episode Statistics (HES) data, which is a statistical database of demographic, medical and administrative information

covering all admissions to National Health Service (NHS) hospitals in England. Although not originally collected for research, large-scale administrative data have been increasingly used for population health research because they usually cover large populations and are relatively inexpensive to acquire and amenable to computerized data extraction [28]. For instance, epidemiological studies using HES data have driven significant service changes in health-care delivery in England [29]. In health service research, in-hospital death is often used to assess and improve hospital quality. However, a comparison of in-hospital death rate between hospitals is not a good standard for monitoring hospital performance directly, because the number of in-hospital death is likely to be influenced by the characteristics of admitted patients. These characteristics may be distributed differently across hospitals. For example, the mortality rate in a hospital that treats more severe patients is more likely to be higher than that in another hospital having less severe patients, even if their clinical performance are the same. Our scientific interest is then to define an indicator of quality of care in hospitals by taking account of hospital death variation explained by the characteristics of their admitted patients. Such characteristics-adjusted death rates reflect differences in quality of care that is related to hospital performance and it would be useful for managers of hospitals and health policy makers to motivate quality improvements and to influence outcomes of health care by informing consumer choices and setting professional standards.

We used an extract of admitted patient care HES data for the 2010/2011 financial year, including over 14 million records for each episode of admitted patient care delivered by NHS hospitals in England. In HES, *episode* refers to an uninterrupted period of care under a particular hospital consultant. A single inpatient admission in one hospital trust in HES is named as a *spell*, which may include more than one episode. We obtained aggregated hospital-level data from 254 NHS hospitals whose mean number of admitted patients is 20890 (SD=11932) and mean number of death is 59.7 (SD=50.1; 10% percentile=17; 25% percentile=26; median=44; 75% percentile=76; 90% percentile=117) in the 2010/2011 fi-

nancial year. We intend to predict the number of in-hospital death for each hospital using the aggregated characteristics of admitted patients. There are a very large number of HES variables on patient characteristics that are described in a 309-page HES Admitted Patient Care Data Dictionary and are available to use subject to spending a significant amount of time to clean the raw data. To illustrate the proposed method, we considered 315 aggregated characteristics of admitted patients.

In the varying-coefficient model, we considered the number of admitted patients in each hospital as the index variable T, which is actually an indicator of the hospital size. Without loss of generality, we re-scaled T to [0, 1]. For the highly skewed death outcome, we applied the proposed CQC SIS with $\tau = 0.5$ to select $d_n = \lfloor n/\log n \rfloor = 46$ covariates in the first stage and then conducted a group SCAD penalization based on median regression in the second stage.

The selected predictors for the median number of in-hospital death are IMD decile groups ('most deprived 10%'), age on admission groups (AGE = 12 years), method of admission group ('the birth of a baby is in this Health Care Provider'), intended management group ('patient not to stay in hospital overnight'), diagnosis groups ('disease of the genitourinary system; 'symptoms, signs, abnormal findings, ill-defined causes'), source of admission group ('babies born in or on the way to hospital') and treatment specialty group ('accident & emergency'). Their estimated functional coefficients are presented in Figure 1, suggesting that the effects of selected predictors are all varying with the change of the hospital size. By comparison, we also applied [7]'s screening method and a group SCAD penalization based on mean regression, which selected 2 predictors (IMD decile and treatment specialty) overlapped with those from our method, and 7 different predictors (1 in method of admission group, 2 in treatment specialty groups and 4 in diagnosis groups).

We then applied our model and the model obtained under [7] to predict the actual number of in-hospital death for individual hospitals. The prediction error of our model is smaller

than that of the model obtained under [7]. In particular, there are 252 (86%) hospitals with a predicted extra death rate within the range [-0.1%, 0.1%] from our method comparing to 248 (84%) hospitals from the method based on mean regression. It is well understood that the distribution of in-hospital death is quite different from a bell-shaped curve and often may contain outlying cases. By using our CQC based learning approaches we can safeguard the estimation accuracy and reduce the influence from a small portion of extreme medical records. `The data analysis for this example was conducted within University College London (UCL) Data Safe Haven - Identifiable Data Handling Solution (IDHS).`

5.2.2 Lung Cancer Data

In this subsection, we illustrate the performance of the censored CQCSIS method proposed in Section 3 using a familiar microarray data set. The data set is extracted from a large retrospective, multi-site, blinded study [30] and involves 442 lung adenocarcinomas, the specific type of lung cancer that is increasing in incidence. Gene expression data were generated by four different laboratories under a common protocol. The same data set was examined by various authors [32, 33], among others). The data consists of measurements of 22,283 gene expressions. A total of 440 subjects after removing the subjects with missing measurements in overall survival time are included in the downstream analysis. The median follow-up time is 46.5 months. The overall censoring proportion is about 46.4%. A primary goal of studying this dataset is to identify those genes that are associated with the overall survival of lung cancer patients. To evaluate the gene effects we consider functional coefficients using patient age as an index variable. Before applying our proposed method, we standardize the expression measurements for each gene to have mean zero and variance one. Because of the high censoring rate of this data, we concentrate on two quantile levels, $\tau = 0.25$ and 0.5 for the analysis. We compare our CQCSIS with existing approaches examined in section 5.1 and the SIS based on Cox proportional hazards model [34].

Table 2 reports the results on the overlaps of selected genes by various screening procedures. When the screening parameter takes $d = \lfloor n/\log n \rfloor = 72$, our proposed screening at quantile level $\tau = 0.25$ has only one overlap (Gene ID 265) with SIS, two overlaps (Gene ID 12834 and 265) with the conditional quantile screening, and zero overlap with the remaining screening procedures. SIS may not be appropriate when proportional hazards assumption is violated. QaSIS and CQSIS do not account for varying-coefficients and are lack of sufficient model flexibility. Thus the genes selected from those methods may not be as important as results from CQCSIS. The low agreement between CQCSIS and other existing approaches suggest that using our CQCSIS may lead to completely new discovery that is unavailable in the previous literature. The results on the top 20 genes by various methods are listed in Table 3, where there are eight genes overlapped between those by CQCSIS(0.25) and those by CQCSIS(0.25), whose IDs are `5596 12834 1310 17193 5719 12234 9896 7466`. Subsequently, we apply a group SCAD penalization upon to these overlapped gene expressions and then result in only one significant gene (ID `12834`) being retained in the final model.

6 Concluding Remarks

We studied variable screening problem for ultrahigh dimensional varying coefficient models via conditional quantile correlation. Our CQCSIS approach is more suitable for heavy-tailed high-dimensional data sets than the traditional correlation-based screening approaches. At the same time we require stronger technical conditions which may not always be satisfied. In practice, we may adopt the following guideline for choosing between existing screening approaches and CQCSIS: In addition to exploratory graphical examination, we can formally apply the model selection test (eg. [36]) to determine whether the data follow heavy-tailed distribution for complete or incomplete censored cases. After the heavy tail distribution status is confirmed, we can use our proposed CQCSIS as well as CQCSIScens; otherwise,

we may use the conventional correlation based screening procedures such as NSIS by [7] and CC-SIS by [10].

Like the existing methods, our proposal focuses on marginal models and therefore might suffer the false selection problem. That is, in the final model after screening, the covariates that are marginal correlated but jointly non-informative may be recruited as redundant members and those that are marginal uncorrelated but jointly informative could be mistakenly screened out. An iterative screening or joint screening as a supplementary procedure is usually needed. In ultrahigh dimensional varying coefficient models, [35] considered forward variable selection procedure to address this issue. However the residual sum of squares based method is not robust to outliers. Further development of iterated or joint screening under the CQC framework is left as a future investigation. In addition, we briefly discuss the influence of taking different quantile levels on the performance of both CQCSIS and an integrated version over a range of quantile level by an additional simulation example in the supplementary materials.

Acknowledgement

Xia's research is partly supported by the Fundamental Research Funds for the Central Universities (Grant No. 2662016QD006). Li's research was partly supported by Academic Research Funding R-155-000-174-114 in Singapore. We thank the Editor, the Associate Editor and two referees for very helpful comments.

Supplementary Materials

The supplementary materials consist of more additional simulation studies as well as technical details for the proof of Theorems 2.1-2.3 and 3.1-3.3.

References

[1] Fan J., Zhang W. Statistical methods with varying coefficient models. *Statistics and Its Interface*, 2008, **1**, 179-195.

[2] Wang H., Xia Y. Shrinkage estimation of the varying coefficient models. *Journal of the American Statistical Association*, 2009, **104**, 747-757.

[3] Wang L., Li H., Huang J.Z. Variable selection in nonparametric varying-coefficient models for analysis of repeated measurements. *Journal of American Statistical Association*, 2008, **103**, 1556-1569.

[4] Noh H., Chung K., Van Keilegom I. Variable selection of varying coefficient models in quantile regression. *Electronic Journal of Statistics*, 2012, **6**, 1220-1238.

[5] Fan J., Lv J. Sure independence screening for ultrahigh dimensional feature space. *Journal of the Royal Statistical Society, Series B*, 2008, **70**, 849-911.

[6] Fan J., Feng Y., Song R. Nonparametric independence screening in sparse ultra-high-dimensional additive models. *Journal of the American Statistical Association*, 2011, **106**, 544-557.

[7] Fan J., Ma Y., Dai W. Nonparametric independent screening in sparse ultra-high dimensional varying coefficient models. *Journal of the American Statistical Association*, 2014, **109**, 1270-1284.

[8] Song R., Yi F., Zou H. On varying-coefficient independence screening for high-dimensional varying-coefficient models. *Statistica Sinica*, 2014b, **24**, 1735-1752.

[9] Cheng M., Honda T., Li J., Peng H. Nonparametric independence screening and structure identification for ultra-high dimensional longitudinal data. *Annals of Statistics*, 2014, **42**, 1819-1849.

[10] Liu J., Li R., Wu R. Feature selection for varying coefficient models with ultrahigh-dimensional covariates. *Journal of the American Statistical Association*, 2014, **109**, 266-274.

[11] Koenker R. *Quantile Regression*. 2005, Cambridge: Cambridge University Press.

[12] He X., Wang, L., Hong H. Quantile-adaptive model-free variable screening for high-dimensional heterogeneous data. *Annals of Statistics*, 2013, **41**, 342-369.

[13] Wu Y., Yin G. Conditional qunatile screening in ultrahigh-dimensional heterogeneous data. *Biometrika*, 2015, **102**, 65-76.

[14] Mai Q., Zou H. The fused Kolmogorov filter: A nonparametric model-free screening method. *Annals of Statistics*, 2015, **43**, 1471-1497.

[15] Ma X., Zhang J. Robust model-free feature screening via quantile correlation. *Journal of Multivariate Analysis*, 2016, **143**, 472-480

[16] Li G., Li Y., Tsai C.L. Quantile correlations and quantile autoregressive modeling. *Journal of the American Statistical Association*, 2015, **110**, 246-261.

[17] Zhao D.S., Li Y. Principled sure independence screening for Cox models with ultra-high-dimensional covariates. *Journal of Multivariate Analysis*, 2012, **105**, 397-411.

[18] Song R., Lu W., Ma S., Jeng X.J. Censored rank independence screening for high-dimensional survival data. *Biometrika*, 2014a, **101**, 799-814.

[19] Hunag J., Horowitz J.L., Ma S. Asymptotic properties of bridge estimators in sparse high-dimensional regression models. *Annals of Statistics*, 2008, **36**, 587-613.

[20] Stone C.J. Optimal glaobal rates of convergence for nonparametric regression. *Annals of Statistics*, 1982, **10**, 1040-1053.

[21] Wang H. Judy, Wang L. Locally weighted censored quantile regression. *Journal of the American Statistical Association*, 2009, **104**, 1117-1128.

[22] Efron B. The Two-Sample Problem With Censored Data *in Proceedings Fifth Berkeley Symposium in Mathematical Statistics*, 1967, Vol. IV, eds. L. Le Cam and J. Neyman, New York: Prentice-Hall, pp. 831-853

[23] Fan J., Li R. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 2001, **96**, 1348-1360.

[24] Zhang C. Nearly unbiased variable selection under minimax concave penalty. *Annals of Statistics*, 2010, **38**, 894-942.

[25] Zou H., Li R. One-step sparse estimate in nonconcave penalized likelihood models. *Annals of Statistics*, 2008, **36**, 1509-1533.

[26] Tang Y.L., Wang H.J., Zhu Z.Y. Variable selection in quantile varying coefficient models with longitudinal data. *Computational Statistics & Data Analysis*, 2013, **57**, 435-449.

[27] Lee E.R., Noh H. and Park B.U. Model selection via Bayesian information criterion for quantile regression models. *Journal of the American Statistical Association*, 2014, **109**, 216-229.

[28] Iezzoni L.I. Assessing quality using administrative data. *Annals of Internal Medicine*, 1997, **127**, 666-674.

[29] Healthcare for London. (2010). *Cardiovascular project-The case for change*. London: NHS Commissioning Support for London.

[30] Shedden K., Talyor J. M., Enkemann S. A., Tsao M. S., Yeatman T. J., Gerald W. L. et al. Gene expression based survival prediction in lung adenocarcinoma: A multisite, blinded validation study. *Nature Medicine*, 2008, **14**, 822-827.

[31] Xia X., Yang H., Li J. Feature screening for generalized varying coefficient models with application to dichotomous responses. *Computational Statistics & Data Analysis*, 2016, **102**, 85-97.

[32] Xia X., Jiang B., Li J., Zhang W. Low-dimensional confounder adjustment and high-dimensional penalized estimation for survival analysis. *Lifetime Data Analysis*, 2016, **22**, 549-569.

[33] Li J., Zheng Q., Peng L., Huang Z. Survival impact index and ultrahigh-dimensional model-free screening with survival outcomes. *Biometrics*, 2016, **72**, 1145-1154.

[34] Fan J., Feng Y., Wu Y. High dimensional variable selection for Cox's proportional hazards model. *IMS Collections*, 2010, **6**, 70-86.

[35] Cheng M., Honda T., Zhang J. Forward variable selection for sparse ultra-high dimensional varying coefficient models. *Journal of the American Statistical Association*, 2016, **111**, 1209-1221.

[36] Panahi H., Model selection test for the heavy-tailed distributions under censored samples with application in financial data. *International Journal of Financial Studies*, 2016, **4**, 24.

			$\varrho = 0$			$\varrho = 0.4$			$\varrho = 0.8$		
Case	Method(τ)	s_n	MMS	RSD	PS	MMS	RSD	PS	MMS	RSD	PS
(1a)	CQCSIS(0.50)	3	4	3	0.975	3	0	1.000	3	0	1.000
	CQCSIS(0.75)	3	6	11	0.880	3	0	1.000	3	0	1.000
	NIS	3	3	0	0.990	3	0	1.000	3	0	1.000
(1b)	CQCSIS(0.50)	3	7	18	0.845	3	0	1.000	3	0	1.000
	CQCSIS(0.75)	3	20	75	0.695	3	0	1.000	3	0	1.000
	NIS	3	430	388	0.090	298	388	0.255	133	329	0.420
(1c)	CQCSIS(0.50)	5	551	327	0.005	540	321	0.045	5	1	1.000
	CQCSIS(0.75)	3	15	34	0.785	3	0	1.000	3	0	1.000
	NIS	5	498	338	0.055	530	378	0.055	5	0	1.000
(1d)	CQCSIS(0.50)	3	6	16	0.885	3	0	1.000	3	0	1.000
	CQCSIS(0.75)	5	693	269	0.015	367	332	0.095	5	1	0.995
	NIS	5	701	259	0.005	506	349	0.080	77	238	0.475

Table 1: Results of the median of minimum model size (MMS), its robust standard deviation (RSD) and the proportion of truly active covariates selected (PS) with a pre-specified threshold size $d_n = \lfloor n/\log n \rfloor$ for Example 1.

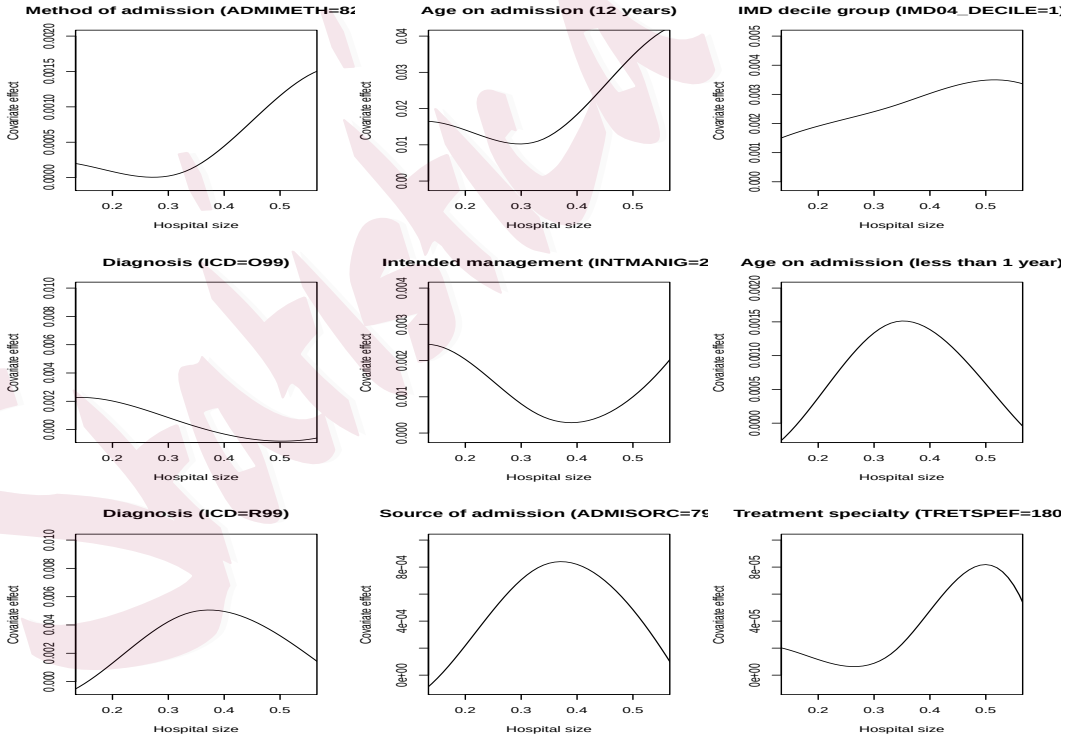

Figure 1: Estimated functional coefficients for the selected predictors.

	CQCSIScens(0.25)	CQCSIScens(0.5)
$d_n = 20$		
SIS	0	0
CRSIS	0	0
QaSIS(0.25)	0	0
QaSIS(0.5)	0	0
CQSIS(0.25)	1	1
CQSIS(0.5)	0	0
CQCSIScens(0.25)	20	8
CQCSIScens(0.5)	8	20
$d_n = 72$		
SIS	1	1
CRSIS	0	1
QaSIS(0.25)	0	1
QaSIS(0.5)	0	1
CQSIS(0.25)	2	2
CQSIS(0.5)	0	1
CQCSIScens(0.25)	72	28
CQCSIScens(0.5)	28	72

Table 2: The number of overlaps of the top d_n genes selected by various methods for Lung cancer data, where $d_n = 20$ and 72, respectively.

Rank	SIS	CRSIS	QaSIS		CQSIS		CQCSIScens	
			$\tau = 0.25$	$\tau = 0.5$	$\tau = 0.25$	$\tau = 0.5$	$\tau = 0.25$	$\tau = 0.5$
1	20612	13344	6253	7426	20022	20612	5596	**12834**
2	2875	12876	7426	6312	2031	4024	**12834**	7466
3	4051	5782	6974	6253	6691	13085	1310	1310
4	7951	10921	6312	16877	4920	8223	18786	16515
5	8236	16422	16877	3949	4921	18286	17193	12234
6	9847	1630	4078	9769	8620	2875	18256	5596
7	13085	14638	9769	5752	22233	4313	8543	17193
8	4313	436	16933	4078	17266	4835	5719	5719
9	14544	15885	5347	15402	4382	15746	12234	8804
10	149	7010	20336	9464	565	816	7995	1151
11	11626	752	5752	16933	9558	17369	20779	9896
12	17303	5184	6781	6361	17714	12334	14012	5604
13	12536	2732	3949	6974	20612	4051	16054	4660
14	17369	10150	5703	20336	16763	8466	12720	20921
15	4835	7512	6687	6781	17374	10238	9896	9172
16	8934	20723	16986	6687	10027	11626	9845	7479
17	3406	22246	5948	5347	2737	12818	8596	20418
18	5145	18471	15402	5398	**12834**	9847	8588	12757
19	9311	2675	9464	3977	21948	3212	7466	7618
20	265	363	5	10975	9197	14289	15455	5330

Table 3: Top 20 selected genes (ID) for Lung cancer data by five screening methods: SIS, screening based on Cox model; CRSIS, censored robust screening; QaSIS, quantile-adaptive screening; CQSIS, conditional quantile screening; CQCSIScens, proposed censored conditional quantile correlation screening.