Statistica Sinica Preprint No: SS-2023-0285	
Title	Global Group Testing and Screening with Dynamic Effects
Manuscript ID	SS-2023-0285
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202023.0285
Complete List of Authors	Ying Cui and Limin Peng
Corresponding Authors	Limin Peng
E-mails	lpeng@emory.edu

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0285

Statistica Sinica

Global Group Testing and Screening With Dynamic Effects

Ying Cui and Limin Peng [†]

Department of Biostatistics and Bioinformatics, Emory University

Abstract: Identifying outcome-related variables is of general research interest in biomedical research. This task can be complicated by the presence of dynamic (or varying) variable effects that often manifest meaningful scientific mechanisms. Appropriately accounting for possible dynamic effects is crucial to avoid depreciating some important variables. In this work, we propose a model-free testing and screening framework by adopting a global view pertaining to the concept of interval quantile independence. The new framework not only permits robust identification of variables dynamically associated with an outcome, but also offers the flexibility to perform group testing that simultaneously evaluates multiple continuous or discrete covariates. We show that the key testing strategy can naturally evolve into unconditional and conditional screening procedures for ultra-high dimensional settings that enjoys the desirable sure screening property. We demonstrate good practical utility of the proposed methods via extensive simulation studies and a real application to a microarray data set.

Key words and phrases: Dynamic effects; Hypothesis testing; Interval quantile independence; Variable screening.

[†]Corresponding author: Limin Peng, Department of Biostatistics and Bioinformatics, Emory University, Atlanta, GA, U.S.A. E-mail: lpeng@emory.edu.

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0285

1. Introduction

A general question arising from many biomedical studies is to determine whether some covariates are relevant to a study outcome. For example, in a genetics study, it is often of interest to identify a group of genes that contribute to the variations of a known disease marker or symptom (Subramanian et al., 2005; Efron and Tibshirani, 2007; Newton et al., 2007, for example). Addressing such an interest, however, may be complicated by the presence of dynamic (or varying) covariate effects. The key issue relates to the way how the relevance or importance of a covariate is defined. For instance, in the context of variable screening, the importance of a covariate was ranked by marginal correlation (Fan and Lv, 2008), maximum marginal likelihood estimate of a generalized linear model (Fan et al., 2010) or a generalized marginal utility function (Fan et al., 2009), and generalized correlation (Hall and Miller, 2009). These approaches involve an assumed linear or generalized linear relationship between the outcome and covariates or transformation thereof, which implicitly asserts a location-shift (or constant) effect for each covariate. Such a restriction was relaxed in model-free screening procedures through adopting nonparametric regression modeling (Fan et al., 2011; He et al., 2013, for example). However, there was a subtle limitation that the adopted nonparametric modeling only examines the local influence of a covariate on the mean or a pre-specified quantile of the outcome. A relevant covariate can be missed if its impact on the outcome is not manifested on the mean or the targeted quantile level. Such phenomena are illustrated by our simulation studies and real data example; for example, see Table 1 and

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0285

Table 3.

Addressing these caveats, a viable option is to measure a covariate's outcome-relevance pertaining to the concept of interval quantile independence (Zhu et al., 2018). Specifically, let Y denote a continuous outcome and denote the vector of the observed covariates as $\boldsymbol{X} = (X^{(1)}, \ldots, X^{(p)})^T$. Define $H_{0,j}: \ Q_Y(\tau \mid X^{(j)}) = Q_Y(\tau), \text{ a.s. for } \tau \in \Delta \subseteq (0,1)$. Here and hereafter, for a general random vector $\boldsymbol{V}$, define the τ-th conditional quantile function of Y given $\boldsymbol{V}$ as $Q_Y(\tau \mid \boldsymbol{V}) = \inf\{y: \ \text{pr}(Y \leq y|\boldsymbol{V}) \geq \tau\}$, and define the τ-th unconditional quantile function of Y as $Q_Y(\tau) = \inf\{y: \ \text{pr}(Y \leq y) \geq \tau\}$. When $X^{(j)}$ is continuous, $H_{0,j}$ refers to the interval quantile independence between Y and $X^{(j)}$ on quantile level intervals Δ and $[0,1]$ respectively for Y and $X^{(j)}$, as termed by Zhu et al. (2018). The consideration of $H_{0,j}$ confers a flexible view for defining relevant variables. In the multivariate setting, a covariate $X^{(j)}$ is considered as relevant or active if $Q_Y(\tau \mid \boldsymbol{X})$ functionally depends on $X^{(j)}$ for some $\tau \in \Delta \subseteq (0,1)$, where Δ is a pre-specified set of quantile levels. Under this view, the set of relevant variables is defined as $\mathcal{M}_\Delta = \{1 \leq r \leq p: \ \text{there exists } \tau \in \Delta \text{ such that } Q_Y(\tau \mid \boldsymbol{X}) \text{ depends on } X^{(r)}\}$.

The formulations of $H_{0,j}$ and $\mathcal{M}_\Delta$ take a global perspective to assess covariate effects throughout the range of the outcome distribution indexed by the quantile level interval Δ. Covariates in $\mathcal{M}_\Delta$ are permitted to have dynamic and non-additive effects across different ranges of the outcome. Several authors (Székely et al., 2007; Zhu et al., 2011; Mai and Zou, 2015; Pan et al., 2019; Zhou and Zhu, 2018; Liu et al., 2022, for example) considered a

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0285

similar general framework for defining covariate relevance, which utilizes the functional dependence of the whole conditional cumulative distribution of the outcome upon the covariate or its distribution. Compared to this alternative, employing the conditional quantile function allows one to naturally pinpoint one part of the outcome distribution (for covariate effect assessment) with a proper choice of Δ to align with particular scientific interests, for example, in average or abnormal outcomes. The flexibility in specifying Δ may also help mitigate potential identifiability concern. For example, when data are limited, say due to censoring to Y, simply setting $\Delta = (0, 1)$ may necessitate extrapolation with additional model assumptions.

To address $H_{0,j}$ and $\mathcal{M}_\Delta$, one available approach is to utilizing the novel interval quantile index proposed by Zhu et al. (2018), which is designed to measure the departure from the interval quantile independence between a pair of continuous variables. Zhu et al. (2018)'s nonparametric index estimator and the associated asymptotic theory naturally render a testing procedure for $H_{0,j}$ when $X^{(j)}$ is continuous. Zhu et al. (2018) also developed a model-free variable screening procedure that ranks the estimated interval quantile index for the relationship between the outcome Y and each continuous covariate $X^{(j)}$. While enjoying desirable theoretical properties (e.g., sure screening property) and appealing empirical performance, Zhu et al. (2018)'s procedures would encounter difficulties when some covariates are discrete. In addition, as the interval quantile index is oriented to study the relationship between two variables, it is not straightforward to adapt Zhu et al. (2018)'s procedures to simultaneously evaluate multiple covariates in terms of their relevance to the outcome. This task is

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0285

often needed in practice in order to sensibly account for inherent data hierarchy structure due to biological, spatial, or temporal factors.

In this work, we propose a new model-free strategy for tackling a generalized version of $H_{0,j}$ that concerns the outcome relevance of one or multiple covariates, which can be either continuous or discrete. Specifically, for an index set for J covariates, $G = \{r_1, \ldots, r_J\} \subseteq \{1, \ldots, p\}$, define $\boldsymbol{X}_G = (X^{(r_1)}, \ldots, X^{(r_J)})^{\mathsf{T}}$. A null hypothesis of our interest takes the form

$$H_{0,G}: \; Q_Y(\tau \mid \boldsymbol{X}_G) = Q_Y(\tau), \text{ a.s., } \text{ for } \tau \in \Delta \subseteq (0,1).$$

To address $H_{0,G}$, we uncover a useful connection between $H_{0,G}$ and a "working" linear quantile regression model, which suggests a nonparametric measure to quantify the departure from $H_{0,G}$. We construct an omnibus test statistic for $H_{0,G}$ from adapting the spirit of the classic Cramér-Von-Mises (C-V) type test statistics under the "working" linear quantile regression model. We establish the asymptotic behaviors of the proposed test statistic without assuming the working model holds.

We further utilize the proposed test statistic as the utility function to develop a new model-free variable screening procedure for ultra-high dimensional data. Given the flexibility of our test statistic in handling multiple covariates simultaneously, the new screening procedure can be performed with covariates pre-grouped by scientific needs or in a random manner for the benefit of saving computational time. We establish the desirable sure screening property for the new screening procedure. As a useful by-product, we can read-

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0285

ily transform the new screening procedure to perform conditional variable screening given some known relevant covariates under mild additional assumptions. We also prove the corresponding conditional sure screening property. As suggested by our numerical studies, in the presence of dynamic effects, the proposed global testing and screening procedures clearly outperform existing approaches that assume constant effects or locally focus on the covariate effects on the mean or a pre-specified quantile of the outcome.

2. The Proposed Global Testing Framework

2.1 Formulation of the proposed test statistic

Without loss of generality, let $G = \{1, \ldots, J\}$ and express the quantile level interval Δ as $[\tau_L, \tau_U]$ with $0 < \tau_L < \tau_U < 1$. Define $\boldsymbol{Z} = (1, \boldsymbol{X}_G^{\mathsf{T}})^{\mathsf{T}}$. As introduced in Section 1, the null hypothesis of interest is

$$H_{0,G}: \; Q_Y(\tau \mid \boldsymbol{X}_G) = Q_Y(\tau), \text{ a.s., for } \tau \in [\tau_L, \tau_U]. \tag{2.1}$$

The observed data consist of n independently identically distributed (i.i.d.) replicates of $(Y, \boldsymbol{Z})$, denoted as $\{(Y_i, \boldsymbol{Z}_i), i = 1, \ldots, n\}$. We assume that the conditional distribution of Y given $\boldsymbol{X}_G$ is continuous and strictly monotone and $E(\boldsymbol{Z}\boldsymbol{Z}^{\mathsf{T}})$ is positive definite.

To address $H_{0,G}$, we uncover a useful connection between $H_{0,G}$ and a "working" linear quantile regression model:

$$Q_Y(\tau \mid \boldsymbol{Z}) = \boldsymbol{Z}^{\mathsf{T}}\boldsymbol{\theta}_0(\tau), \quad \tau \in [\tau_L, \tau_U], \tag{2.2}$$

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0285

where $\boldsymbol{\theta}_0(\tau) = \{\alpha_0(\tau), \beta_0^{(1)}(\tau), \ldots, \beta_0^{(J)}(\tau)\}^{\mathsf{T}}$ is a vector of regression coefficients. A key fact is that $H_{0,G}$ holds if and only if model (2.2) holds with $\beta_0^{(j)}(\tau) = 0$ for $\tau \in [\tau_L, \tau_U]$ for $j = 1, \ldots, J$; see Lemma S1 and its proof in the Supplementary Materials.

To utilize this connection, we consider an estimator of $\boldsymbol{\theta}_0(\tau)$ defined as the minimizer of the quantile loss function $\arg\min_{\boldsymbol{b}\in R^{p+1}} \sum_{i=1}^n \rho_\tau(Y_i - \boldsymbol{Z}_i^{\mathsf{T}}\boldsymbol{b})$, where $\rho_\tau(u) \doteq u\{\tau - I(u \le 0)\}$ is the so-called "check" function (Koenker and Bassett, 1978). Solving this minimization problem is equivalent to solving the corresponding score estimating equation

$$\boldsymbol{S}_n(\boldsymbol{b}, \tau) \doteq n^{-1/2} \sum_{i=1}^n \boldsymbol{Z}_i [I(Y_i \le \boldsymbol{Z}_i^{\mathsf{T}}\boldsymbol{b}) - \tau] = 0, \tag{2.3}$$

with respect to $\boldsymbol{b}$. Denote the solution by $\widehat{\boldsymbol{\theta}}(\tau) = (\widehat{\alpha}_0(\tau), \widehat{\beta}^{(1)}(\tau), \ldots, \widehat{\beta}^{(J)}(\tau))^{\mathsf{T}}$. Without assuming the working model (2.2), we can show that $\widehat{\boldsymbol{\theta}}(\tau)$ may uniformly converge to $\widetilde{\boldsymbol{\theta}}(\tau) = (\widetilde{\alpha}_0(\tau), \widetilde{\beta}^{(1)}, \ldots, \widetilde{\beta}^{(J)}(\tau))^{\mathsf{T}}$ over $\tau \in [\tau_L, \tau_U]$, where $\widetilde{\boldsymbol{\theta}}(\tau)$ is the solution to the equation, $\boldsymbol{\mu}(\boldsymbol{b}, \tau) \doteq E[\boldsymbol{Z}\{I(Y \le \boldsymbol{Z}^{\mathsf{T}}\boldsymbol{b}) - \tau\}] = 0$, with respect to $\boldsymbol{b} \in R^{J+1}$; see Theorem S1 in the Supplementary Materials. By Lemma S2 in the Supplementary Materials, the solution to $\boldsymbol{\mu}(\boldsymbol{b}, \tau) = 0$ uniquely exists and $H_{0,G}$ implies $(\widetilde{\beta}^{(1)}(\tau), \ldots, \widetilde{\beta}^{(J)}(\tau))^{\mathsf{T}} = \mathbf{0}$.

Remark 1: Note that $\widetilde{\boldsymbol{\theta}}(\tau) = \boldsymbol{\theta}_0(\tau)$ when the "working" model (2.2) is the true model. When model (2.2) does not hold, the consideration of $\boldsymbol{\theta}_0(\tau)$ is no longer meaningful but $\widetilde{\boldsymbol{\theta}}(\tau)$ is still well defined as the solution to the deterministic equation $\boldsymbol{\mu}(\boldsymbol{b}, \tau) = 0$.

Motivated by these results, we propose to test data departure from $H_{0,G}$ by using the deviation of $(\widetilde{\beta}^{(1)}(\tau), \ldots, \widetilde{\beta}^{(J)}(\tau))^{\mathsf{T}}$ from $\mathbf{0} \in R^J$ for $\tau \in [\tau_L, \tau_U]$. Employing the connection between $\widetilde{\boldsymbol{\theta}}(\cdot)$ and the working model (2.2) permits leveraging existing inferential tools and

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0285

software for quantile regression to facilitate the task of testing $H_{0,G}$ based on $\widetilde{\boldsymbol{\theta}}(\cdot)$. It also provides an intuitive way to interpret $\widetilde{\boldsymbol{\theta}}(\cdot)$, which would capture covariate effects on the τ-th quantile of the outcome when the working model holds.

Specifically, we propose to construct the test statistic for $H_{0,G}$ as

$$\widehat{T}_{UC} = \max_{j \in G=\{1,\ldots,J\}} \widehat{T}^{(j)}_{inte},$$

where $\widehat{T}^{(j)}_{inte} = \int_{\tau_L}^{\tau_U} \left| n^{1/2} \widehat{\beta}^{(j)}(\tau) / \widehat{\sigma}^{(j)}_n(\tau) \right|^2 d\tau$ and $\widehat{\sigma}^{(j)2}_n(\tau)$ is the variance estimate for $n^{1/2}\{\widehat{\beta}^{(j)}(\tau) - \widetilde{\beta}^{(j)}(\tau)\}$ elaborated later. The construction of $\widehat{T}_{UC}$ reflects the idea of first utilizing the squared $\widetilde{\beta}^{(j)}(\tau)$ to capture the local influence of $X^{(j)}$ at the single τ, integrating the local effect over $\tau \in [\tau_L, \tau_U]$ to assess the global effect of $X^{(j)}$, and then taking the maximum global effect across all covariates. Such a test statistic shares a similar spirit of the Cramér-Von-Mises (C-V) test statistic and is expected to be sensitive to any departure of $(\widetilde{\beta}^{(1)}(\tau), \ldots, \widetilde{\beta}^{(J)}(\tau))^{\mathsf{T}}$ from the constant zero function.

In Theorem 1, we establish the limit null distribution of $\widehat{T}_{UC}$. The proof is provided in Section S1.4 of the Supplementary Materials.

Theorem 1 *Suppose the regularity conditions S3-S4 in the Supplementary Materials hold. Under the null hypothesis $H_{0,G}$, we have*

$$\widehat{T}_{UC} \quad \rightarrow_d \max_{j \in \{1,\ldots,J\}} \left[\int_{\tau_L}^{\tau_U} \{\mathcal{X}^{(j)}(\tau)\}^2 d\tau \right],$$

where $\mathcal{X}^{(j)}(\tau)$ is a mean zero Gaussian process defined in Section S1.4 of Supplementary Materials, $j = 1, \ldots, J$.

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0285

2.2 The proposed global testing procedure

Given the connection between $\widetilde{\boldsymbol{\theta}}(\tau)$ and the working model (2.2), we can readily obtain $\widehat{\beta}^{(j)}(\tau)$ by using the `rq()` function in the R package `quantreg`. As detailed in Theorem S2 in the Supplementary Materials, under certain regularity conditions, $n^{1/2}\{\widehat{\boldsymbol{\theta}}(\tau) - \widetilde{\boldsymbol{\theta}}(\tau)\}$ converges weakly to a mean zero Gaussian process for $\tau \in [\tau_L, \tau_U]$ with covariance $\boldsymbol{\Phi}(\tau', \tau) = E\{\boldsymbol{\xi}_i(\tau')\boldsymbol{\xi}_i(\tau)^{\mathsf{T}}\}$, where the influence function $\boldsymbol{\xi}_i(\tau)$ is defined in Theorem S2. The asymptotic result allows us to obtain the variance estimate $\widehat{\sigma}_n^{(j)}(\tau)$ from adapting Peng and Fine (2009)'s sample-based inference procedure outlined below with additional algorithmic details provided in Section S3 of the Supplementary Materials.

(1.a) Compute $\widehat{\boldsymbol{\Sigma}}(\tau, \tau) = n^{-1}\sum_{i=1}^{n} \boldsymbol{Z}_i\boldsymbol{Z}_i^{\mathsf{T}}[I\{Y_i \leq \boldsymbol{Z}_i^{\mathsf{T}}\widehat{\boldsymbol{\theta}}(\tau)\} - \tau]^2$.

(1.b) Conduct eigenvalue eigenvector decomposition for $\widehat{\boldsymbol{\Sigma}}(\tau, \tau)$ using `eigen()` function in R to find the matrix $\boldsymbol{E}_n(\tau) = \{\boldsymbol{e}_{n,0}(\tau), \ldots, \boldsymbol{e}_{n,J}(\tau)\}$ such that $\{\boldsymbol{E}_n(\tau)\}^2 = \widehat{\boldsymbol{\Sigma}}(\tau, \tau)$.

(1.c) Solve the perturbed estimating equation $\boldsymbol{S}_n(\boldsymbol{c}, \tau) = \boldsymbol{e}_{n,j}(\tau)$ for $j = 1, \ldots, J$ and denote the solution as $\boldsymbol{S}_n^{-1}\{\boldsymbol{e}_{n,j}(\tau), \tau\}$.

(1.d) Calculate $\boldsymbol{D}_n(\tau) = \left(\boldsymbol{S}_n^{-1}(\boldsymbol{e}_{n,0}(\tau), \tau) - \widehat{\boldsymbol{\theta}}(\tau), \ldots, \boldsymbol{S}_n^{-1}(\boldsymbol{e}_{n,J}(\tau), \tau) - \widehat{\boldsymbol{\theta}}(\tau)\right)$. Compute an estimate for the asymptotic variance of $n^{1/2}\{\widehat{\boldsymbol{\theta}}(\tau) - \widetilde{\boldsymbol{\theta}}(\tau)\}$ as $\boldsymbol{V}_n(\tau) \equiv n\boldsymbol{D}_n^{\otimes 2}(\tau)$. Obtain $\widehat{\sigma}_n^{(j)2}(\tau)$ as the $j + 1^{\text{th}}$ diagonal component of $\boldsymbol{V}_n(\tau)$.

The result in Theorem 1 indicates that the asymptotic null distribution of the proposed test statistic is non-standard. We develop a perturbation resampling procedure to obtain the

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0285

p value from testing $H_{0,G}$ based on the proposed test statistic. The resampling procedure is described as follow.

(2.a) Generate B independent sets of $\{\iota_i^b\}_{i=1}^n$, where $\{\iota_i^b\}_{i=1}^n$ are independent random variables from a standard normal distribution for $b = 1, \ldots, B$.

(2.b) Calculate $\widehat{\boldsymbol{\xi}}_i(\tau) = \{\widehat{A}(\widehat{\boldsymbol{\theta}}(\tau))\}^{-1}\boldsymbol{Z}_i\{I(Y_i \leq \boldsymbol{Z}_i^{\mathsf{T}}\widetilde{\boldsymbol{\theta}}(\tau)) - \tau\}$, where $\{\widehat{A}(\widehat{\boldsymbol{\theta}}(\tau))\}^{-1}$ is obtained from $\{\widehat{A}(\widehat{\boldsymbol{\theta}}(\tau))\}^{-1} = n^{1/2}\boldsymbol{D}_n(\tau)\boldsymbol{E}_n(\tau)^{-1}$.

(2.c) For $b = 1, \ldots, B$, calculate

$$\widehat{T}_{UC,b} = \max_{j\in\{1,\ldots,J\}}\left\{\int_{\tau_L}^{\tau_U}\left|n^{-1/2}\sum_{i=1}^{n}\widehat{\xi}_i^{(j)}(\tau)\iota_i^b/\widehat{\sigma}_n^{(j)}(\tau)\right|^2 d\tau\right\},$$

where $\widehat{\xi}_i^{(j)}(\tau)$ is the $j+1^{\text{th}}$ component of $\widehat{\boldsymbol{\xi}}_i(\tau)$.

(2.d) The p value is calculated by $p_{UC} = \sum_{b=1}^{B} I(\widehat{T}_{UC,b} > \widehat{T}_{UC})/B$.

Similar resampling procedures were used in other settings, such as Lin et al. (1993), Li and Peng (2014), and Cui and Peng (2022). The key idea is to approximate the limit null distribution through perturbing the influence function $\boldsymbol{\xi}_i(\tau)$. The above resampling procedure is easy to implement without involving smoothing. Justification for this procedure is provided in Section S1.6 of the Supplementary Materials.

In Theorem S3 in the Supplementary Materials, we further show that the proposed test statistic $\widehat{T}_{UC}$ is consistent (i.e., power approaching 1 as $n \to \infty$) against the alternative

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0285

hypothesis,

$$H_{a,G}: \text{For some } j_1 \in \{1,\ldots,J\}, \text{there exists } \tau \in [\tau_L, \tau_U] \text{ such that } |\widetilde{\beta}^{(j_1)}(\tau)| > 0.$$

This result suggests promising power of the proposed procedure for detecting departures from $H_{0,G}$.

Remark 2: Given a bijective map from R^J to R^J, $\boldsymbol{\Psi}(\cdot)$, it is easy to show that $H_{0,G}$ is equivalent to $Q_Y(\tau \mid \boldsymbol{\Psi}(\boldsymbol{X}_G)) = Q_Y(\tau)$, a.s. Consequently, carrying out the proposed testing procedure based on the transformed covariates in $\boldsymbol{\Psi}(\boldsymbol{X}_G)$ would still confer valid inference for $H_{0,G}$.

3. Variable Screening in Ultra-high Dimensional Setting

3.1 The proposed unconditional screening framework

Consider the ultra-high dimensional setting, where $p = O(\exp(n^c))$ for a positive $c < 1$. Suppose the observed covariates are grouped as $(\boldsymbol{X}_{G_1}^{\mathsf{T}}, \ldots, \boldsymbol{X}_{G_L}^{\mathsf{T}})^{\mathsf{T}}$, where $G_1, \ldots, G_L$ are non-overlapping index sets and $\cup_{l=1}^{L} G_l = \{1, \ldots, p\}$. While both p and L may depend on the sample size n, we omit n from their notation for presentation simplicity. Assume that the sizes of G_l's ($l = 1, \ldots, L$) are finite and uniformly bounded. This implies that p and L are of the same asymptotic order.

In practice, the grouping of covariates may be motivated by scientific needs, for example, grouping genes according to biological pathways. The special case with $p = L$ corresponds

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0285

to the regular scenario where no grouping is imposed to covariates. Thus, a unified definition of the set of relevant covariates, with or without grouping, is $\mathcal{M}_{[\tau_L,\tau_U]} = \{G_l : 1 \leq l \leq L$ and there exists $\tau \in [\tau_L, \tau_U]$ such that $Q_\tau(Y|\boldsymbol{X})$ depends on $X_{G_l}\}$. Assume that the cardinality of $\mathcal{M}_{[\tau_L,\tau_U]}$ is smaller than the sample size n. Let $\lfloor\cdot\rfloor$ and $\lceil\cdot\rceil$ denote the floor and ceiling operators respectively.

We propose the following variable screening procedure:

(3.a) Normalize $\boldsymbol{X}$ and Y.

(3.b) For each index set G_l, compute $\widehat{T}_{UC}$ for H_{0,G_l} and denote it by $w_{1,l},\ l = 1, \ldots, L$.

(3.c) Sort $\{\boldsymbol{X}_{G_1}, \ldots, \boldsymbol{X}_{G_L}\}$ according to $w_{1,l}$ in a decreasing order.

(3.d) Keep $\boldsymbol{X}_{G_l}$'s with $w_{1,l}$ greater than a pre-defined threshold ν_n or keep a pre-specified number (e.g., $\lfloor n/\log n \rfloor$) of covariates on the top of the list obtained from (3.c).

By the above procedure with some pre-determined threshold value ν_n, the set of remaining variables is defined as $\widehat{\mathcal{M}}_{[\tau_L,\tau_U]} = \{G_l : 1 \leq l \leq L,\ w_{1,l} \geq \nu_n\}$.

When there is no particular scientific reasons to group covariates, we have the variable screening problem with $L = p$. In this case, we may consider an alternative two-step screening procedure that first filters covariates by randomly formed groups and then conducts a second-step single covariate screening. Our numerical investigation shows that such a two-step procedure may preserve similar screening performance while saving computational time. Specifically, the two-step screening procedure includes the following steps:

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0285

(4.a) Normalize $\boldsymbol{X}$ and Y.

(4.b) Perform the first-step group-level screening:

(i) Shuffle the index set of the covariates $\{1, \ldots, p\}$ to $\{r_1, \ldots, r_p\}$.

(ii) With a pre-determined group size S_G, compute $L = \lceil p/S_G \rceil$. Divide the first $(L-1)\cdot S_G$ covariates into $L-1$ groups of equal size S_G. The Lth group includes the last $p - S_G \cdot (L-1)$ elements. Denote the resulting grouped covariates as $\{\boldsymbol{X}_{G_1}, \ldots, \boldsymbol{X}_{G_L}\}$.

(iii) Apply steps (3.b)–(3.d) to the grouped covariates $\{\boldsymbol{X}_{G_1}, \ldots, \boldsymbol{X}_{G_L}\}$ with a pre-sepecified threshold $\nu_{n,1}$ or a pre-specified number (e.g. $[n/\log n]$) of groups to keep.

(4.c) Express the set of remaining variables from (4.b) in terms of individual covariates, $\{X^{(\tilde{r}_1)}, \ldots, X^{(\tilde{r}_M)}\}$, and then perform the second-step individual-level screening:

(i) Obtain $\widehat{T}_{UC}$ for $H_{0,\{\tilde{r}_m\}}$, denoted by $w_{2,m}$, for $m = 1, \ldots, M$.

(ii) Sort $\{X^{(\tilde{r}_1)}, \ldots, X^{(\tilde{r}_M)}\}$ according to $w_{2,m}$ in a decreasing order.

(iii) Keep the covariates with $w_{2,m}$ greater than a pre-specified threshold $\nu_{n,2}$ or keep a pre-specified number (e.g. $[n/\log n]$) of covariates on the top of the list obtained from (4.c) (ii).

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0285

With this two-step screening procedure, the set of remaining variables is defined as $\widehat{\mathcal{M}}^{G}_{[\tau_L,\tau_U]} = \{r_m : 1 \leq m \leq M,\ w_{2,m} \geq \nu_{n,2}\}$.

We establish the sure screening property for the proposed unconditional screening procedures. Let $\widehat{T}^{(G_l)}_{UC}$ denote the proposed test statistic $\widehat{T}_{UC}$ for H_{0,G_l} and define

$$T^{(G_l)}_{UC} = \max_{j \in G_l} \int_{\tau_L}^{\tau_U} \left| n^{1/2} \widetilde{\beta}^{(j)}(\tau) / \sigma^{(j)}(\tau) \right|^2 d\tau,$$

where $\{\sigma^{(j)}(\tau)\}^2$ is the $j+1$th diagonal element of $\mathbf{\Phi}(\tau,\tau)$ defined in Theorem S2 of the Supplementary Materials. In Theorem 2, we establish the exponential probability bounds for $|n^{-1}\widehat{T}^{(G_l)}_{UC} - n^{-1}T^{(G_l)}_{UC}|$. This result serves as the key step to prove Corollary 1 and Corollary 2, which state the sure screening property of the proposed one-step screening procedure outlined in (3.a)–(3.d) and that of the proposed two-step screening procedure outlined in (4.a)–(4.c), respectively.

Theorem 2 *Suppose that the regularity conditions S1-S4 in the Supplementary Material hold. For any $c > 0$ and $1/4 < \zeta \leq 1/2$, there exists positive constant υ and η such that*

$$pr(\max_{1 \leq l \leq L} |n^{-1}\widehat{T}^{(G_l)}_{UC} - n^{-1}T^{(G_l)}_{UC}| \geq cn^{\zeta - 1/2}) \leq p\upsilon \exp\{-\eta n^{4\zeta - 1} - \log(n^{\zeta - 1/2})\}$$

for sufficiently large n.

Corollary 1 (Sure screening property of the one-step screening procedures) *Suppose that the regularity conditions S1-S5 in the Supplementary Material hold. If we take the threshold*

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0285

value $\nu_n = \delta^* n^{\zeta-1/2}$ *with* $\delta^* \leq \alpha_0/2$*, then there exists positive constants,* a_1 *and* b_1*, such that*

$$pr(\mathcal{M}_{[\tau_L,\tau_U]} \subseteq \widehat{\mathcal{M}}_{[\tau_L,\tau_U]}) \geq 1 - S_{[\tau_L,\tau_U]} \cdot a_1 \exp\{-b_1 n^{4\zeta-1} - \log(n^{\zeta-1/2})\}$$

for sufficiently large n*, where* $S_{\tau_L,\tau_U} = |\mathcal{M}_{[\tau_L,\tau_U]}|$ *is the cardinality of* $\mathcal{M}_{[\tau_L,\tau_U]}$*. In particular,* $pr(\mathcal{M}_{[\tau_L,\tau_U]} \subseteq \widehat{\mathcal{M}}_{[\tau_L,\tau_U]}) \to 1$ *as* $n \to \infty$*.*

Corollary 2 (Sure screening property of the two-step screening procedure) *Suppose that the regularity conditions S1-S4 and S6 in the Supplementary Material hold. If we take the threshold value* $\nu_{n,1} = \delta^* n^{\zeta-1/2}$ *and* $\nu_{n,2} = \delta^{**} n^{\zeta-1/2}$ *with* $0 < \delta^* \leq \alpha_0/2$ *and* $0 < \delta^{**} \leq \alpha_0/2$*, respectively, then there exists positive constant* a_2 *and* b_2*, such that*

$$pr(\mathcal{M}_{[\tau_L,\tau_U]} \subseteq \widehat{\mathcal{M}}^G_{[\tau_L,\tau_U]}) \geq 1 - S_{[\tau_L,\tau_U]} \cdot a_2 \exp\{-b_2 n^{4\zeta-1} - \log(n^{\zeta-1/2})\}$$

for sufficiently large n*, where* $S_{[\tau_L,\tau_U]} = |\mathcal{M}_{[\tau_L,\tau_U]}|$ *is the cardinality of* $\mathcal{M}_{[\tau_L,\tau_U]}$*. In particular,* $pr(\mathcal{M}_{[\tau_L,\tau_U]} \subseteq \widehat{\mathcal{M}}^G_{[\tau_L,\tau_U]}) \to 1$ *as* $n \to \infty$*.*

The proofs of Theorem 2, Corollary 1 and Corollary 2 are provided in Sections S2.2 and S2.3 of the Supplementary Materials.

3.2 A generalization to conditional screening

In practice, a set of covariates may be known to relate to the outcome by existing knowledge. In many studies, assessing the relative importance of other covariates in the presence of the known relevant covariates is of interest. This confers a conditional screening problem

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0285

(Barut et al., 2016). By the proposed testing strategy, we can readily generalize the screening procedures presented in Section 3.1 to conduct conditional variable screening.

Denote $\boldsymbol{X}_{\mathcal{C}}$ as the set of relevant covariates known from prior knowledge, and the rest covariates as $\boldsymbol{X}_{-\mathcal{C}}$. Suppose $\boldsymbol{X}_{-\mathcal{C}}$ is grouped as $\{\boldsymbol{X}_{G_{c,1}}, \ldots, \boldsymbol{X}_{G_{c,L_c}}\}$. When L_c equals the length of $\boldsymbol{X}_{-\mathcal{C}}$, no grouping is imposed to $\boldsymbol{X}_{-\mathcal{C}}$. Adapting the global perspective taken in the proposed unconditional screening framework, we consider $\boldsymbol{X}_{G_{c,l}}$ as conditionally irrelevant to the outcome if $Q_Y(\tau \mid \boldsymbol{X}_{\mathcal{C}}, \boldsymbol{X}_{-\mathcal{C}})$ does not depend on $\boldsymbol{X}_{G_{c,l}}$ for $\tau \in [\tau_L, \tau_U]$. Under this view, screening out conditionally irrelevant covariates is naturally linked to the problem of testing $H_{c,G_{c,l}}: \; Q_Y(\tau \mid \boldsymbol{X}_{\mathcal{C}}, \boldsymbol{X}_{G_{c,l}}) = Q_Y(\tau \mid \boldsymbol{X}_{\mathcal{C}})$ for $\tau \in [\tau_L, \tau_U]$. We assume that $\boldsymbol{X}_{\mathcal{C}}$ has a known type of relationship with the outcome. For simplicity, we assume that $Q_Y(\tau \mid \boldsymbol{X}_{\mathcal{C}})$ is linearly related to $\boldsymbol{X}_{\mathcal{C}}$ for $\tau \in [\tau_L, \tau_U]$. Similar to the finding in the unconditional setting, $H_{c,G_{c,l}}$ holds if and only if the working linear quantile regression model

$$Q_Y(\tau \mid \boldsymbol{X}_{\mathcal{C}}, \boldsymbol{X}_{G_{c,l}}) = \alpha_c(\tau) + \boldsymbol{X}_{\mathcal{C}}{}^{\mathsf{T}}\boldsymbol{\beta}_{c,1} + \boldsymbol{X}_{G_{c,l}}^{\mathsf{T}}\boldsymbol{\beta}_{c,2}, \quad \tau \in [\tau_L, \tau_U], \tag{3.1}$$

holds with $\boldsymbol{\beta}_{c,2} = \mathbf{0}$. This fact naturally motivates the following conditional variable screening procedure:

(5.a) Normalize $\boldsymbol{X}$ and Y.

(5.b) For each index set $G_{c,l}$, compute a conditional test statistic $\widehat{T}_C$ for $H_{c,G_{c,l}}$, which is obtained in the same manner as that for $\widehat{T}_{UC}$ except that the working linear quantile regression model includes $\boldsymbol{X}_{\mathcal{C}}$ in addition to $\boldsymbol{X}_{G_{c,l}}$. Denote the resulting $\widehat{T}_C$ by

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0285

$w_{c,l},\ l = 1, \ldots, L_c$.

(5.c) Sort $\{\boldsymbol{X}_{G_{c,1}}, \ldots, \boldsymbol{X}_{G_{c,L_c}}\}$ according to $w_{c,l}$ in a decreasing order.

(5.d) Keep $\boldsymbol{X}_{c,G_l}$'s with $w_{c,l}$ greater than a pre-defined threshold $\nu_{c,n}$ or keep a pre-specified number (e.g., $\lfloor n/\log n \rfloor$) of covariates on the top of the list obtained from (5.c).

By the above procedure with some pre-determined threshold value $\nu_{c,n}$, the set of remaining variables is defined as $\widehat{\mathcal{M}}^{\mathcal{C}}_{[\tau_L,\tau_U]} = \{G_{c,l} : 1 \le l \le L_c,\ w_{c,l} \ge \nu_{c,n}\}$.

We establish the sure screening property for the proposed conditional screening procedure. Denote the conditional test statistic $\widehat{T}_C$ for $H_{c,G_{c,l}}$ by $\widehat{T}_C^{(G_{c,l})}$. Let $T_C^{(G_{c,l})}$ be $\widehat{T}_C^{(G_{c,l})}$ with the coefficient estimate and variance estimate replaced by their population analogues. Define the set of conditionally relevant covariates as $\mathcal{M}^{(\mathcal{C})}_{\tau_L,\tau_U} = \{G_{c,l} :\ 1 \le l \le L_c,$ there exists $\tau \in [\tau_L, \tau_U]$ such that $Q_Y(\tau \mid \boldsymbol{X})$ depends on $\boldsymbol{X}_{G_{c,l}}\}$. The results of the exponential tail probability bound for $|n^{-1}\widehat{T}_C^{(G_{c,l})} - n^{-1}T_C^{(G_{c,l})}|$ and the sure screening property are respectively summarized in Theorem S4 and Corollary S1 in Section S2.1 of the Supplementary Materials. Their proofs are provided in Sections S2.2 and S2.3 of the Supplementary Materials.

4. Numerical Studies

4.1 Simulation studies for evaluating the proposed testing procedure

We first evaluate the proposed testing procedure for $H_{0,G}$ in univariate settings where $\boldsymbol{X}_G$ contains one covariate X. The specific set-ups for generating $(X, Y)^{\mathsf{T}}$ are presented in Table

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0285

S7 in the Supplementary Materials. In set-ups U1-U4, the working model (2.2) holds for $\tau \in [0.2, 0.8]$, which is the τ-interval of interest. Set-up U5 gives a scenario where the working linear quantile regression model does not hold. It is easy to see that U1 is a null case, where X has no effect on Y. In U2, a standard linear model holds and X has a constant effect on Y over $\tau \in [0.2, 0.8]$. U3 and U4 are two set-ups with dynamic effects varying across different τ's. The coefficient functions involved in set-ups U3 and U4 are presented in Figure S2 in the Supplementary Materials. In set-up U5, X takes a non-linear functional form to influence Y and thus the working model is not satisfied. We compare the following testing procedures:

GIT: the proposed test based on $\widehat{T}_{UC}$ with $[\tau_L, \tau_U] = [0.2, 0.8]$;

AQI: the test proposed in Zhu et al. (2018), with the quantile interval set as $[0.2, 0.8]$ for Y and $[0, 1]$ for X;

Q_S: rank score test (Gutenbrunner et al., 1993)

Q_W: Wald test (Koenker and Bassett, 1982)

L_W: Wald test based on linear regression.

In each setting, the significance level is set as 0.05. We consider sample sizes, 200 and 400.

Table S8 of the Supplementary Materials presents the empirical rejection rates in cases U1–U5 based on 1000 simulations. In the null case U1, all methods yield empirical sizes close to the nominal level of 0.05. In set-ups U3–U5, where dynamic covariate effects are present, we observe that the proposed method and Zhu et al. (2018)'s method, which are

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0285

designed to capture global effects throughout $\tau \in [0.2, 0.8]$, yield much higher power than tests which target the local effect on a single τ or the mean when dynamic effects are present, for example set-ups U3 and U4. These demonstrate substantial power gains resulted from integrating information across quantiles in the presence of dynamic covariate effects. In addition, we observe that the proposed method and Zhu et al. (2018)'s method have comparable performance in the univariate settings.

We also evaluate the proposed testing procedure in multivariate settings, where $\boldsymbol{X}_G$ includes two covariates X_1 and X_2. To illustrate the utility of our method for handling both continuous and discrete variables, we generate X_1 as a continuous variable and X_2 as a discrete variable. We consider five settings M1–M5 with configuration details shown in Table S7 in the Supplementary Materials. M1 is the null case, where both X_1 and X_2 have no effects on Y. M2 corresponds to the case where only X_1 influences Y and its effect is constant. In M3, both X_1 and X_2 have constant covariate effects on Y. In M4, X_1 and X_2 have partial effects on Y. The true coefficient functions, $q_{m_1}(\tau)$ and $q_{m_2}(\tau)$, are shown in Figure S2 in the Supplementary Materials. M5 is a set-up where X_1 and X_2 influence Y in a non-standard way and the working model (2.2) does not hold. In all multivariate settings, Zhu et al. (2018) is no longer applicable. We compare the proposed GIT to Q_S and Q_W with $\tau = 0.4$, 0.5, or 0.6 and the analysis of variance test for overall significance based on linear regression (ANOVA).

Table 1 reports the empirical rejection rates of these tests based on 1000 simulations.

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0285

All methods have empirical sizes close to the nominal level 0.05 in the null case M1. The empirical power of all tests grows as the sample size increases. When there are varying covariate effects, such as in set-ups M4–M5, the proposed method can yield much higher power than tests, Q_S, Q_W, and ANOVA, which target local covariate effects on a single τ or the mean. All these results suggest that good utility of the proposed tests to detect the existence of either constant or dynamic covariate effects, no matter the covariates are continuous or discrete.

We conduct additional simulation studies to investigate the performance of the proposed tests with different choices of $[\tau_L, \tau_U]$, different sets of candidate adjusting constants $\mathcal{U}$, and different cardinality of G. The results are summarized and discussed in Section S4.1 of the Supplementary Materials.

4.2 Simulation studies for evaluating the proposed screening procedures

We conduct simulation studies to evaluate the performance of the proposed one-step screening procedure in (3.a)–(3.d), denoted by GIT, and the proposed two-step procedure in (4.a)–(4.c) with $S_G = 2$, denoted by GOT. For comparisons, we consider existing approaches, including Fan and Lv (2008)'s method, denoted by SIS, He et al. (2013)'s method at $\tau = 0.25, 0.5$ or 0.75, denoted by QaSIS(τ), as well as Zhu et al. (2018)'s method with quantile interval sets, $[0.2, 0.8]$ for Y and $[0, 1]$ for X, denoted by AQI. When implementing He et al. (2013)'s method, we set the number of basis as 3. To assess the performance of these screen-

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0285

Table 1: Empirical rejection rates with 1000 replicates for $\boldsymbol{X}_G = (X_1, X_2)^\intercal$.

Set-up	n	$\tau \in [0.2, 0.8]$	τ=0.4		τ=0.5		τ=0.6		ANOVA
		GIT	Q_S	Q_W	Q_S	Q_W	Q_S	Q_W	
M1	200	0.058	0.050	0.055	0.053	0.048	0.050	0.047	0.044
	400	0.045	0.042	0.041	0.041	0.038	0.042	0.047	0.040
M2	200	0.933	0.805	0.791	0.827	0.799	0.823	0.801	0.952
	400	1.000	0.987	0.982	0.988	0.986	0.989	0.985	0.998
M3	200	0.900	0.809	0.797	0.831	0.812	0.811	0.802	0.959
	400	0.999	0.991	0.987	0.992	0.991	0.988	0.980	0.999
M4	200	0.409	0.169	0.131	0.052	0.036	0.116	0.111	0.051
	400	0.821	0.228	0.210	0.040	0.028	0.231	0.203	0.045
M5	200	0.415	0.452	0.325	0.384	0.345	0.268	0.264	0.025
	400	0.710	0.758	0.716	0.690	0.698	0.543	0.572	0.039

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0285

ing methods, we use the median minimum model size of the selected models required for sure screening, and the robust standard deviation, defined as the interquartile range of minimum model size, and the probability of selecting each $X^{(j)}$, and the probability of selecting all covariates in $\mathcal{A}$ when top $[n/\log(n)]$ covariates are maintained.

The simulation set-ups are described as follows:

S1 ($n = 200$, $p = 2000$): $Y = 0.2(X^{(1)} + 0.8X^{(2)} + 0.6X^{(3)} + 0.4X^{(4)} + 0.2X^{(5)}) + \varepsilon$, where ε follow the standard normal distribution.

S2 ($n = 200$, $p = 2000$): $Y = 0.2(X^{(1)} + 0.8X^{(2)} + 0.6X^{(3)} + 0.4X^{(4)} + 0.2X^{(5)}) + \exp(Z) \cdot \varepsilon$, where Z and ε follow the standard normal distribution.

S3 ($n = 400$, $p = 5000$): $Y = X^{(1)}I(X^{(1)} > 0) + X^{(2)}I(X^{(1)} \leq 0) + \exp(X^{(19)} + X^{(20)}) + \exp(X^{(3)}) \cdot \varepsilon$, where ε follows the standard normal distribution.

S4 ($n = 400$, $p = 5000$): $Q_Y(\tau|\boldsymbol{X}) = 3X^{(1)}I(X^{(1)} > 0) + 3X^{(3)}I(X^{(1)} \leq 0) + l_S(\tau) \cdot X^{(4)} + u_S(\tau) \cdot X^{(5)} + (s(X^{(2)}) + 1)^2 \cdot Q_\varepsilon(\tau)$, where $s(a) = (a - E(a))/sd(a)$ with $sd(a)$ denoting the standard deviation of a, $l_S(\tau)$ and $u_S(\tau)$ are plotted in Figure S2 of the Supplementary Materials, and ε follows standard Cauchy distribution.

In the above set-ups, the covariates $\boldsymbol{X} = \{X^{(1)}, \ldots, X^{(p)}\}^T$ are generated from multivariate normal distribution with mean zero and covariance matrix $\boldsymbol{\Sigma} = (0.9^{|k-k'|})_{p\times p}$. The error terms Z and ε are independent of $\boldsymbol{X}$. It is easy to see that the relevant covariate set is $\mathcal{A} =$

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0285

$\{X^{(1)}, X^{(2)}, X^{(3)}, X^{(4)}, X^{(5)}\}$ for set-ups S1, S2 and S4. In set-up S3, the relevant/active covariates are more separated from each other with $\mathcal{A} = \{X^{(1)}, X^{(2)}, X^{(3)}, X^{(19)}, X^{(20)}\}$. We further consider an additional set-up S4*, which is the same as S4 except that we transform half of the covariates to discrete covariates. Specifically, in S4*, we first use the same way to generate $(X^{(1)}, X^{(2)}, \ldots, X^{(p)})^T$, and then dichotomize $(X^{(2)}, X^{(4)}, \ldots, X^{(2\lfloor p/2 \rfloor)})^T$ at 0 to generate binary covariates defined as $I(X^{(j)} < 0)$ ($j = 2, 4, \ldots, 2\lfloor p/2 \rfloor$). In this case, Zhu et al. (2018)'s method can not be applied. To implement He et al. (2013)'s method, we use the linear option due to singular issues. We transform each normally distributed covariate by $\Phi(\cdot)$ before applying the proposed testing procedure.

In Table 2, we summarize the screening results based on 500 simulations. In set-up S1, where the error term follows the normal distribution and the relevant covariates are highly correlated with each other, we observe that all the methods perform quite well. In set-up S2, which differs from S1 only by the error distribution, we notice that there is substantial deterioration with the performance of SIS. The number of covariates needed for sure screening along with its variability inflates substantially from MMMS(RSD)$= 5(0)$ to $50(264)$, and the probability of retaining all relevant covariates drops significantly from 1.00 to 0.47. In the other three set-ups, S3, S4 and S4*, we have similar observations regarding the under-performance of Fan and Lv (2008). Such observations are not surprising and are likely caused by the fact that the normal error assumption is no longer valid in these settings. Also, we notice that He et al. (2013) has varying performance for different τ's. For example, in set-up S2, He et al.

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0285

(2013)'s method with $\tau = 0.5$ may select the relevant covariates over 80% of times; while by He et al. (2013)'s method with $\tau = 0.25$ or 0.75 the probability of keeping all relevant variables reduces to be below 20%. Compared to He et al. (2013)'s method, which focuses on local effects, the screening procedures that examine global effects, such as Zhu et al. (2018)'s method and the proposed methods, GIT and GOT, demonstrate better performance, as reflected by larger selection probabilities, pr($\mathcal{A}$), and smaller model sizes measured by MMMS and RSD. A reasonable interpretation is that the global testing procedures leverage information across different τ's, thereby producing higher detection power.

In set-ups S1, S2, and S4, where relevant covariates are strongly correlated, the proposed methods, GIT and GOT, and Zhu et al. (2018)'s method, AQI, have similar performance. In set-up S3, the relevant covariates are separated into two clusters with one cluster including $X^{(1)}$, $X^{(2)}$, and $X^{(3)}$ and the other including $X^{(19)}$ and $X^{(20)}$. In addition, $X^{(19)}$ and $X^{(20)}$ have stronger covariate effects than $X^{(1)}$, $X^{(2)}$, and $X^{(3)}$. In this case, though the proposed methods and Zhu et al. (2018)'s method all have high probabilities of selecting relevant covariates, Zhu et al. (2018)'s method yields relatively larger model size as compared to the proposed methods. This is caused by the tendency of Zhu et al. (2018)'s method to select "neighboring" covariates around $X^{(19)}$ and $X^{(20)}$, such as $X^{(18)}$ or $X^{(21)}$. Since these covariates are highly corrected with $X^{(19)}$ and $X^{(20)}$ (which have strong effects on the outcome), Zhu et al. (2018)'s method may catch the trails of these neighboring covariates by producing interval quantile independence indices comparable to or even higher than those

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0285

for $X^{(1)}, X^{(2)}$, and $X^{(3)}$. Despite this discrepancy, we think that the proposed methods and Zhu et al. (2018)'s method have quite comparable performance in variable screening when all covariates are continuous, while the proposed methods offer flexibility to naturally accommodate discrete covariates.

We also conduct additional simulation studies to compare the proposed screening method with the screening methods of Székely et al. (2007) and Zhou and Zhu (2018) to evaluate the potential benefits from utilizing covariate grouping information in the proposed screening methods. Details are provided in Section S4.3 of the Supplementary Materials.

Table 2: The simulation results for unconditional procedures.

Set-up	Method	MMS		pr($X^{(j)}$)					pr($\mathcal{A}$)
		MMMS	RSD	$X^{(1)}$	$X^{(2)}$	$X^{(3)}$	$X^{(4)}$	$X^{(5)}$	
S1	GIT	5	1	1.000	1.000	1.000	1.000	1.000	1.000
	SIS	5	0	1.000	1.000	1.000	1.000	0.998	0.998
	QaSIS(0.25)	8	10	0.976	0.990	0.988	0.982	0.910	0.884
	QaSIS(0.5)	6	3	0.990	0.998	0.998	0.992	0.972	0.958
	QaSIS(0.75)	8	9	0.994	0.992	0.988	0.968	0.912	0.896
	AQI	5	0	1.000	1.000	1.000	1.000	1.000	1.000
	GOT($S_G = 2$)	5	1	0.998	0.996	1.000	0.998	0.992	0.990
S2	GIT	5	1	1.000	1.000	1.000	0.998	1.000	0.998

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0285

Table 2 – continued from previous page

Set-up	Method	MMS		pr($X^{(j)}$)					pr($\mathcal{A}$)
		MMMS	RSD	$X^{(1)}$	$X^{(2)}$	$X^{(3)}$	$X^{(4)}$	$X^{(5)}$	
	SIS	50	264	0.694	0.694	0.688	0.628	0.554	0.466
	QaSIS(0.25)	97	130	0.482	0.544	0.518	0.392	0.268	0.142
	QaSIS(0.5)	20	20	0.978	0.988	0.972	0.956	0.854	0.822
	QaSIS(0.75)	98	128	0.448	0.536	0.490	0.410	0.286	0.178
	AQI	5	1	1.000	1.000	1.000	0.998	0.998	0.996
	GOT($S_G = 2$)	5	1	0.998	0.996	1.000	0.998	0.996	0.992
S3	GIT	13	7	1.000	1.000	1.000	1.000	1.000	1.000
	SIS	214	798	0.358	0.402	0.446	1.000	1.000	0.300
	QaSIS(0.25)	25	10	0.992	1.000	0.982	1.000	1.000	0.976
	QaSIS(0.5)	27	9	0.996	1.000	0.998	1.000	1.000	0.994
	QaSIS(0.75)	90	77	0.406	0.562	0.640	1.000	1.000	0.340
	AQI	20	6	1.000	1.000	1.000	1.000	1.000	1.000
	GOT($S_G = 2$)	13	7	1.000	1.000	1.000	1.000	1.000	1.000
S4	GIT	5	0	1.000	1.000	1.000	1.000	1.000	1.000
	SIS	2194	3860	0.260	0.266	0.256	0.214	0.196	0.140
	QaSIS(0.25)	33	37	0.912	0.950	0.992	0.954	0.888	0.794
	QaSIS(0.5)	6	2	1.000	1.000	1.000	1.000	1.000	1.000

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0285

Table 2 – continued from previous page

Set-up	Method	MMS		pr($X^{(j)}$)					pr($\mathcal{A}$)
		MMMS	RSD	$X^{(1)}$	$X^{(2)}$	$X^{(3)}$	$X^{(4)}$	$X^{(5)}$	
	QaSIS(0.75)	10	9	1.000	1.000	1.000	1.000	1.000	1.000
	AQI	5	0	1.000	1.000	1.000	1.000	1.000	1.000
	GOT($S_G = 2$)	5	0	1.000	1.000	1.000	1.000	1.000	1.000
S4*	GIT	5	1	1.000	1.000	1.000	1.000	1.000	1.000
	SIS	2180	4139	0.332	0.264	0.376	0.306	0.346	0.186
	QaSIS(0.25)	5	0	1.000	1.000	1.000	1.000	1.000	1.000
	QaSIS(0.5)	5	0	1.000	1.000	1.000	1.000	1.000	1.000
	QaSIS(0.75)	8	4	1.000	0.912	1.000	1.000	1.000	0.912
	AQI	–	–	–	–	–	–	–	–
	GOT($S_G = 2$)	5	1	1.000	1.000	1.000	1.000	1.000	1.000

In addition, we conduct simulation studies to investigate the performance of the proposed conditional screening procedure outlined in (5.a)–(5.d). The details are included in Section S4.3 of the Supplementary Materials. The results strongly support the advantage of taking a global view for assessing covariate effects, particularly in the presence of dynamic covariate effects.

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0285

5. An Application to a Gene Microarray Dataset

We apply the proposed methods to a microarray dataset (Scheetz et al., 2006), which contains the gene expression levels of 31,098 probe sets on 120 12-week old male offsprings of rats. With this dataset, one interest is to identify the set of genes related to gene *TRIM32*, which is a known predictor for genetically heterogeneous diseases including Muscular Dystrophy, Limb-Girdle, Autosomal Recessive 8 and Bardet-Biedl Syndrome 11. The probe id corresponding to gene *TRIM32* is "1389163_at".

We first illustrate the utility of the proposed global testing procedure through evaluating the marginal relevance of six example genes to the expression level of *TRIM32*. To test the effect of each of these genes, we apply the proposed global test, Wald tests for linear quantile regression with $\tau = 0.25, 0.5$ and 0.75, Wald tests for linear regression with outliers and after the removal of outliers based on Cook's Distance, and Wald test based on the robust linear regression (Hampel et al., 1986). Table 3 presents the p values obtained from these different tests. For the genes with probe id "1367462_at" and "1372996_at", their effects are captured by linear regression after removing outliers, robust linear regression, quantile regression based tests with most choices of τ, as well as the proposed test. The test based on standard linear regression does not detect the effect of either of these two genes, likely due to the "diluting" influence from the outliers. We have opposite findings regarding the effects of the genes with probe id "1367479_at" and "1367525_at". As hinted by the scatter plots in Figure S4 in the Supplementary Materials, these discrepant results are likely caused by the

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0285

presence of a few outliers, which are not appropriately handled by standard linear regression and thus leads to spurious effect estimates. These demonstrate the robustness of the proposed testing procedure against outliers.

Table 3: Summary for the p values of six example genes.

Probe id	GIT	$Q_W(0.25)$	$Q_W(0.5)$	$Q_W(0.75)$	L_W	L_W(rm)	RL_W
1367462_at	0.0004	0.0530	0.0018	0.0007	0.4673	0.0005	0.0005
1372996_at	0.0024	0.0012	0.0257	0.0111	0.9714	0.0049	0.0078
1367479_at	0.7616	0.2396	0.5638	0.6088	0.0176	0.1663	0.4939
1367525_at	0.9640	0.5331	0.9326	0.8139	0.0082	0.9819	0.7016
1379467_at	0.0040	0.1097	0.3106	0.0293	0.4370	0.5645	0.5439
1381314_at	0.0184	0.0951	0.0352	0.7773	0.1069	0.4604	0.1371

As suggested by exploratory marginal linear quantile regression analyses (see the third column of Figure S4 in the Supplementary Materials), constant location-shift effects may not be adequate for the genes with probe id, "1379467_at" and "1381314_at", but are presumed by linear regression based tests. In this case, the local quantile regression based tests separately examine the effects of these genes at different quantile levels; thus it is not surprising that the resulting p values suggest significant effects at some τ's but not at the other τ's. All linear regression based tests fail to capture the effects of these two genes. This may reflect effect attenuation resulted from assuming a varying effect as constant. The proposed test, by taking a global perspective for assessing effects, sensibly support the relevance of these two

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0285

genes to the outcome.

We apply the proposed screening procedures to help identify outcome-relevant genes out of 31,097 genes. In our analyses, we first perform the proposed unconditional screening procedures to filter out most irrelevant genes. Specifically, we keep the genes ranked top $[2n/\log n] = 50$. With the remaining genes, we perform Zheng et al. (2015)'s globally adaptive quantile regression method with $\tau \in [0.2, 0.8]$ for further variable selection. We also analyze the same data by alternative combinations of screening and variable selection approaches, including Fan and Lv (2008) coupled with adaptive Lasso for linear regression (Zou, 2006), He et al. (2013) coupled with locally concerned quantile regression with adaptive Lasso penalty (Belloni and Chernozhukov, 2011) for $\tau = 0.25$, 0.5, or 0.75, and Zhu et al. (2018) coupled with Zheng et al. (2015)'s globally adaptive quantile regression method with $\tau \in [0.2, 0.8]$. When applying each approach, we determine the tuning parameter in the variable selection step by cross validation.

The heatmap presented in Figure 1 informs the sets of genes selected by different approaches and also displays the Pearson's correlation in expression level between the genes selected by the proposed one-step approach and the genes selected by the other approaches. With the same variable selection procedure, using the proposed global tests for variable screening leads to more parsimonious selection of genes as compared to adopting Zhu et al. (2018) which also takes a global view for variable screening. We observe that the gene with probe id "1393510_at" selected from using the proposed methods is also selected from using

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0285

Fan and Lv (2008), He et al. (2013) and Zhu et al. (2018). All genes selected based on the proposed one-step approach have moderate or high correlations with at least one gene selected by the other approaches. This observation may help endorse the sensible gene selection by the proposed approach based on the results from several benchmark approaches.

For each approach, we further assess the quantile prediction performance. To compare across different approaches, we adjust the tuning parameter in the variable selection step so that all approaches select the same number of genes. For a given number of selected genes, denoted by g, following the approach developed by Li and Peng (2017), we measure the quantile prediction error as

$$\widehat{PE}^{(g)} = n^{-1} \sum_{i=1}^{n} \int_{\tau_L}^{\tau_U} \rho_\tau [Y_i - \boldsymbol{X}_{S,i}^{\mathsf{T}} \widehat{\boldsymbol{\theta}}_S(\tau)] d\tau,$$

where $\rho_\tau(u) = u\{\tau - I(u < 0)\}$, $\boldsymbol{X}_S$ represents the express levels of the selected genes, and $\widehat{\boldsymbol{\theta}}_S(\tau)$ represents the estimated regression quantiles derived from the final model fitting at the variable selection step. Under a linear regression model or a local quantile regression model, the $\widehat{\boldsymbol{\theta}}_S(\tau)$ is extrapolated as a constant function over τ equal to the regression coefficient estimate.

In Table 4, we report $\widehat{PE}^{(g)}$ with $g = 1, \ldots, 10$ resulted from all the approaches considered. We see that the estimated prediction errors associated with the proposed methods are always comparable or smaller than those associated with the other methods. For example, with $g = 5$, the estimated prediction errors associated with the proposed methods are both around 0.11 and are smaller than the other approaches.

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0285

We also apply conditional screening procedures with the conditioning covariates representing the two important genes suggested in Scheetz et al. (2006), *Abca4* and *Opn1sw* with probe ID "1384603_at" and "1388025_at", respectively. We pair the proposed method and Barut et al. (2016)'s method respectively with globally adaptive quantile regression method (Zheng et al., 2015) and linear regression with adaptive LASSO for variable selection. The heatmap presented in Figure S3 in the Supplementary Materials indicates that the proposed method, CGIT, yields much more sparse gene selection results as compared to Barut et al. (2016)'s method, CSIS, and the expression level of genes selected by the proposed method are well correlated with those of the genes selected by Barut et al. (2016)'s method. The results in Table 4 show that the estimated prediction error is 0.19 based on linear regression with only *Abca4* and *Opn1sw* as covariates. The prediction errors decrease when the conditionally relevant covariate set includes additional covariates identified from conditional variable screening and variable selection. The prediction errors associated with the proposed method are smaller than those associated with Barut et al. (2016)'s method in most cases. The prediction error reduction from using the proposed method instead of Barut et al. (2016)'s method, is more apparent when there are fewer selected genes. This may indirectly imply that the proposed method, as compared to Barut et al. (2016)'s method, may give higher priority to genes with more predictive power and thus leads to larger gains in prediction when the "model size" is smaller.

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0285

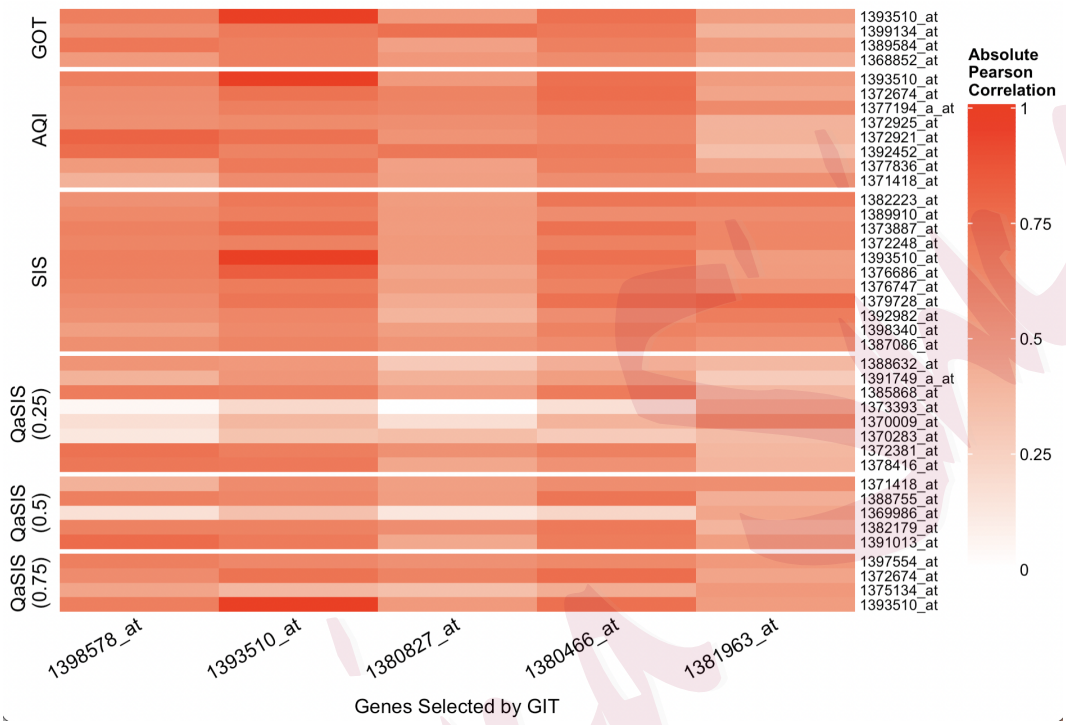

Figure 1: Heat map for the correlation between the proposed unconditional screening procedure and the other methods.

6. Remarks

In this work, we develop a new testing and screening framework that can help determine outcome-relevant covariates in classic univariate and multivariate settings and ultra-high dimensional settings. The proposed methods sensibly adopt a global perspective that examines covariate effects over a continuum of outcome quantiles. Such a global perspective shares a similar spirit with the concept of globally concerned quantile regression proposed by Zheng et al. (2015). Nevertheless, Zheng et al. (2015)'s work is hinged upon the assumption of

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0285

Table 4: The estimated prediction error $\widehat{PE}^{(g)}(\Delta)$ for different model size.

Method	Size:1	Size:2	Size:3	Size:4	Size:5	Size:6	Size:7	Size:8	Size:9	Size:10
A. Unconditional screening										
GIT	0.178	0.171	0.118	0.117	0.116	0.115	0.114	0.113	0.109	0.108
GOT	0.177	0.172	0.169	0.127	0.113	0.112	0.111	0.111	0.111	0.111
AQI	0.171	0.157	0.142	0.141	0.140	0.134	0.133	0.124	0.118	0.108
SIS	0.185	0.172	0.159	0.144	0.137	0.123	0.113	0.109	0.109	0.108
QaSIS ($\tau = 0.25$)	0.195	0.179	0.179	0.167	0.166	0.162	0.156	0.155	0.144	0.143
QaSIS ($\tau = 0.5$)	0.150	0.148	0.138	0.129	0.122	0.120	0.119	0.119	0.117	0.116
QaSIS ($\tau = 0.75$)	0.180	0.173	0.158	0.154	0.155	0.135	0.131	0.130	0.125	0.124
B. Conditional screening (*Abca4*, *Opn1sw*)										
CGIT	–	–	0.161	0.163	0.163	0.156	0.113	0.104	0.103	0.103
CSIS	–	–	0.178	0.169	0.142	0.130	0.120	0.117	0.112	0.105

a global linear quantile regression model, while our testing procedures tackle a non-model-based null hypothesis and the corresponding screening procedure is model-free. Our numerical studies strongly support the advantages of the proposed methods over existing locally concerned methods, particularly in data settings with dynamic covariate effects.

It is worth mentioning that Wang et al. (2018) also investigated the null hypothesis $H_{0,G}$ in a scenario where $[\tau_L, \tau_U]$ reduces a singleton set or becomes a discrete set consisting of

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0285

multiple specified quantile levels. In this special case, Wang et al. (2018)'s method and ours are different n several aspects. First, Wang et al. (2018) assumed a multivariate linear quantile regression model as the true model, while in our framework, the multivariate quantile regression model (2.2) is only treated as a working model. Secondly, in terms of the test statistic construction, Wang et al. (2018) employed a maximum-type statistic defined based on the working univariate quantile regression models separately assumed for each covariate, while our strategy is to utilize a novel quantity, $\tilde{\boldsymbol{\theta}}(\tau)$, which is closely connected to $H_{0,G}$ and can be conveniently estimated by adopting a working multivariate quantile regression. Thirdly, both approaches involve variance estimation for regression quantiles but different procedures are used. Specifically, Wang et al. (2018) utilized a kernel-based estimator. In contrast, we adopt a sample-based variance estimation procedure circumventing the need for smoothing. Similar to Wang et al. (2018), we consider a fixed number J of covariates in establishing the asymptotic properties of the test statistic for $H_{0,G}$. However, accommodating diverging J's remains a challenge that warrants further investigation to enhance the applicability of our methodology.

Under the proposed testing framework, we capture the covariate relevance through the quantity, $\tilde{\boldsymbol{\theta}}(\tau)$, the definition of which does not rely on any model specification. It is reasonable to expect that the finite-sample power of the proposed tests is low when $\tilde{\boldsymbol{\theta}}(\tau)$'s magnitude is small across $\tau \in \Delta$. In practice, this type of scenarios may be diagnosed through empirical examination of the covariate-response relationship, and the limitations of the proposed tests

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0285

pertaining to the inadequate sample size or the small effect size may be acknowledged.

Supplementary Material

Supplementary Material available online includes technical proofs and additional results.

Acknowledgments

This work was supported by NIH grant R01 HL113548.

References

Barut, E., J. Fan, and A. Verhasselt (2016). Conditional sure independence screening. *Journal of the American Statistical Association 111*(515), 1266–1277.

Belloni, A. and V. Chernozhukov (2011). L1-penalized quantile regression in high-dimensional sparse models. *The Annals of Statistics 39*(1), 82–130.

Cui, Y. and L. Peng (2022). Assessing dynamic covariate effects with survival data. *Lifetime data analysis 28*(4), 675–699.

Efron, B. and R. Tibshirani (2007). On testing the significance of sets of genes. *The annals of applied statistics 1*(1), 107–129.

Fan, J., Y. Feng, and R. Song (2011). Nonparametric independence screening in sparse

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0285

ultra-high-dimensional additive models. *Journal of the American Statistical Association 106*(494), 544–557.

Fan, J. and J. Lv (2008). Sure independence screening for ultrahigh dimensional feature space. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 70*(5), 849–911.

Fan, J., R. Samworth, and Y. Wu (2009). Ultrahigh dimensional feature selection: beyond the linear model. *The Journal of Machine Learning Research 10*, 2013–2038.

Fan, J., R. Song, et al. (2010). Sure independence screening in generalized linear models with np-dimensionality. *The Annals of Statistics 38*(6), 3567–3604.

Gutenbrunner, C., J. Jureˆcková, R. Koenker, and S. Portnoy (1993). Tests of linear hypotheses based on regression rank scores. *Journaltitle of Nonparametric Statistics 2*(4), 307–331.

Hall, P. and H. Miller (2009). Using generalized correlation to effect variable selection in very high dimensional problems. *Journal of Computational and Graphical Statistics 18*(3), 533–550.

Hampel, F. R., E. M. Ronchetti, P. Rousseeuw, and W. A. Stahel (1986). *Robust statistics: the approach based on influence functions*. Wiley-Interscience; New York.

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0285

He, X., L. Wang, H. G. Hong, et al. (2013). Quantile-adaptive model-free variable screening for high-dimensional heterogeneous data. *The Annals of Statistics 41*(1), 342–369.

Koenker, R. and G. Bassett (1978). Regression quantiles. *Econometrica 46*(1), 33–50.

Koenker, R. and G. Bassett (1982). Robust tests for heteroscedasticity based on regression quantiles. *Econometrica 50*(1), 43–61.

Li, R. and L. Peng (2014). Varying coefficient subdistribution regression for left-truncated semi-competing risks data. *Journal of Multivariate Analysis 131*, 65–78.

Li, R. and L. Peng (2017). Assessing quantile prediction with censored quantile regression models. *Biometrics 73*(2), 517–528.

Lin, D. Y., L.-J. Wei, and Z. Ying (1993). Checking the cox model with cumulative sums of martingale-based residuals. *Biometrika 80*(3), 557–572.

Liu, W., Y. Ke, J. Liu, and R. Li (2022). Model-free feature screening and fdr control with knockoff features. *Journal of the American Statistical Association 117*(537), 428–443.

Mai, Q. and H. Zou (2015). The fused kolmogorov filter: A nonparametric model-free screening method. *The Annals of Statistics 43*(4), 1471–1497.

Newton, M. A., F. A. Quintana, J. A. Den Boon, S. Sengupta, and P. Ahlquist (2007). Random-set methods identify distinct aspects of the enrichment signal in gene-set analysis. *The Annals of Applied Statistics 1*(1), 85–106.

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0285

Pan, W., X. Wang, W. Xiao, and H. Zhu (2019). A generic sure independence screening procedure. *Journal of the American Statistical Association 114*(526), 928–937. PMID: 31692981.

Peng, L. and J. P. Fine (2009). Competing risks quantile regression. *Journal of the American Statistical Association 104*(488), 1440–1453.

Scheetz, T. E., K.-Y. A. Kim, R. E. Swiderski, A. R. Philp, T. A. Braun, K. L. Knudtson, A. M. Dorrance, G. F. DiBona, J. Huang, T. L. Casavant, et al. (2006). Regulation of gene expression in the mammalian eye and its relevance to eye disease. *Proceedings of the National Academy of Sciences 103*(39), 14429–14434.

Subramanian, A., P. Tamayo, V. K. Mootha, S. Mukherjee, B. L. Ebert, M. A. Gillette, A. Paulovich, S. L. Pomeroy, T. R. Golub, E. S. Lander, et al. (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences 102*(43), 15545–15550.

Székely, G. J., M. L. Rizzo, N. K. Bakirov, et al. (2007). Measuring and testing dependence by correlation of distances. *The Annals of Statistics 35*(6), 2769–2794.

Wang, H. J., I. W. McKeague, and M. Qian (2018). Testing for marginal linear effects in quantile regression. *Journal of the Royal Statistical Society. Series B, Statistical methodology 80*(2), 433.

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0285

Zheng, Q., L. Peng, and X. He (2015). Globally adaptive quantile regression with ultra-high dimensional data. *Annals of statistics 43*(5), 2225.

Zhou, Y. and L. Zhu (2018). Model-free feature screening for ultrahigh dimensional datathrough a modified blum-kiefer-rosenblatt correlation. *Statistica Sinica 28*(3), 1351–1370.

Zhu, L., Y. Zhang, K. Xu, et al. (2018). Measuring and testing for interval quantile dependence. *The Annals of Statistics 46*(6A), 2683–2710.

Zhu, L.-P., L. Li, R. Li, and L.-X. Zhu (2011). Model-free feature screening for ultrahigh-dimensional data. *Journal of the American Statistical Association 106*(496), 1464–1475.

Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association 101*(476), 1418–1429.

Emory University

E-mail: ying.cui@emory.edu

Emory University

E-mail: lpeng@emory.edu