Statistica Sinica Preprint No: SS-2020-0115	
Title	Sparse Composite Quantile Regression with Ultra-high Dimensional Heterogeneous Data
Manuscript ID	SS-2020-0115
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202020.0115
Complete List of Authors	Lianqiang Qu, Meiling Hao and Liuquan Sun
Corresponding Author	Meiling Hao
E-mail	meilinghao@uibe.edu.cn
Notice: Accepted version subject to English editing.	

Sparse Composite Quantile Regression with Ultra-high Dimensional Heterogeneous Data

Lianqiang Qu[1], Meiling Hao[2] and Liuquan Sun[3]

[1]School of Mathematics and Statistics, Central China Normal University, Wuhan, Hubei, 430079, P.R.China

[2]School of Statistics, University of International Business and Economics, Beijing 100029, P.R.China

[3]Institute of Applied Mathematics, Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing, 100190, P.R.China

Abstract. Quantile regression is widely employed in heterogeneous data, but to select covariates that globally affect the response and estimate coefficients simultaneously are very challenging. In this article, we introduce a novel sparse composite quantile regression screening method for the analysis of ultra-high dimensional heterogeneous data. The proposed method enjoys the sure screening property, provides a consistent selection path, and yields consistent estimates of coefficients simultaneously across a continuous range of quantile levels. An extended Bayesian information criterion is employed to select the "best" candidate from the path. Extensive simulation studies demonstrate the effectiveness of the proposed method, and an application to a gene expression dataset is provided.

Keywords: Quantile regression; Sparsity; Ultra-high dimensional data; Variable screening.

1 Introduction

Ultra-high dimensional data frequently arise in a wide variety of scientific fields, such as genomics, biomedical imaging, signal processing, finance, and so forth. For such data, the number of covariates p greatly exceeds the sample size n, and even grows at an exponential rate of n. One major feature of these datasets is heterogeneity. This poses challenges but also great opportunities for statistical analysis.

Quantile regression, as an important alternative to linear regression, is a technique to investigate the heterogeneity across quantiles (Koenker and Bassett (1978)). For high dimensional data, many penalized quantile regression methods have been well developed to inquire into covariate effects at a single or multiple prespecified quantile levels (Zou and Yuan (2008); Wang, Wu and Li (2012); Fan, Fan and Barut (2014)). However, these existing models are sensitive to the specific choices of quantile levels and may overlook some important covariates, which are undesirable for interpretation. To settle this issue, Belloni and Chernozhukov (2011) and Zheng, Peng and He (2015, 2018) extended quantile regression methods to examine regression quantiles over a continuous set of quantile levels. Such kind of quantile regression method enjoys two advantages: (1) it takes advantage of all useful information across quantiles and can draw a robust conclusion; (2) it grasps global sparsity more concisely. These offer a useful complement to regularized quantile regression, and make it much more flexible for variable selection, robust estimation, and heteroscedasticity detection. However, regularized quantile regression may not perform well under ultra-high dimensional scenarios, especially in the aspects of computational

expediency, statistical accuracy, and algorithmic stability (Fan and Lv (2010)). This inspires the development of screening methods.

The sure independence screening (SIS) method was proposed for sparse recovery in ultra-high dimensional linear regression models (Fan and Lv (2008)), and the idea is to rank all covariates by using the marginal correlation between each covariate and the response. This method enjoys the sure screening property and is widely applied in various models (Fan, Samworth and Wu (2009); Fan and Song (2010); Zhu et al. (2011); Fan, Feng and Song (2011); Liu, Li and Wu (2014); Song et al. (2014); Fan et al. (2017); Kong et al. (2017); Pan et al. (2019)). To derive robust statistics, He, Wang and Hong (2013) considered a quantile-adaptive model-free variable screening method. Wu and Yin (2015) developed a conditional quantile screening method via a goodness-of-fit-like marginal utility. Ma, Li and Tsai (2017) employed the quantile partial correlation and proposed three variable screening algorithms. For other related works, we refer the readers to Zhang and Zhou (2018); Li, Ma and Zhang (2018) and the references therein. Note that these screening methods only considered model sparsity at a single or multiple quantile levels. Recently, Ma and Zhang (2016) and Xu (2017) proposed composite quantile correlation via integrating quantile levels from 0 to 1, which enjoys the sure screening property and grasps global sparsity. However, these two works did not study the estimation of coefficients, and also did not consider an interval of quantile levels that well captures part or all of the conditional distributions.

Against this background, we aim to develop a variable screening method that can glob-

ally capture important features and estimate their coefficients simultaneously. Motivated by the work of Zheng, Peng and He (2015), we adopt a quantile regression model with an interval of quantile levels, denoted as $\Theta \subset (0, 1)$, and propose an approach called sparse composite quantile regression (SCQR) for variable screening. The SCQR naturally embeds the sparsity information about regression functions in composite quantile regression and identifies active covariates by the estimates of regression functions over a continuum of quantile levels. It utilizes the joint effects rather than the marginal effects of candidate covariates, in the spirit of Xu and Chen (2014) and Yang et al. (2018). Compared to that of Xu and Chen (2014) and Yang et al. (2018), our proposed method is robust in model selection, and the development of theory and algorithm is not a trivial extension of existing methods, due to a nonsmooth objective function.

The main contribution of this article is twofold. First, we establish the consistency properties of our method in terms of model selection and parameter estimation. Specifically, the SCQR method can create a solution path including the true model with probability approaching one, and can also yield a consistent estimate across a continuous range of quantile levels. To the best of our knowledge, this is new in the screening literature. An extended Bayesian information criterion (EBIC) (Lee, Noh and Park (2014)) is employed to identify the ideal model. Second, we employ a smoothing technique to develop an iterative groupwise-hard-thresholding method to approximate our proposed solution, establish the convergence of the proposed algorithm, and show the sure screening property of the approximation solution. The proposed algorithm overcomes two kinds of computational

challenges. One is that the objective function is not differentiable at zero point. The other comes from the ℓ_0 constraint, which results in a heavy computation burden of the existing programming for quantile regression.

The rest of the paper is organized as follows. Section 2 provides some preliminaries about high-dimensional sparse quantile regression models and describes the SCQR method. Section 3 presents a high efficient algorithm for the SCQR procedure. Section 4 establishes the theoretical properties of the SCQR procedure and the proposed algorithm. An application to a gene expression dataset is provided in Section 5, and some concluding remarks are given in Section 6. Simulation studies and all proofs are given in the online Supplementary Material.

2 Methodology

2.1 Some Preliminaries

Let $\mathbf{X} = (1, x_1, \ldots, x_p)^\top$ be a $(p+1)$-dimensional vector of covariates, and $Q_Y(\tau|\mathbf{X}) = \inf\{y|P(Y \leq y|\mathbf{X}) \geq \tau\}$ denote the τth conditional quantile of a response variable Y given $\mathbf{X}$. For the analysis, the following quantile regression model (Zheng, Peng and He (2015)) is considered:

$$Q_Y(\tau|\mathbf{X}) = \mathbf{X}^\top \boldsymbol{\beta}^\star_\tau \quad \text{for} \quad \tau \in \Theta, \tag{1}$$

where $\boldsymbol{\beta}_\tau^\star = (\beta_{\tau,0}^\star, \beta_{\tau,1}^\star, \ldots, \beta_{\tau,p}^\star)^\top$ is a $(p+1)$-dimensional vector of unknown coefficient functions of τ, $\Theta \subset (0,1)$ is a pre-specified continuous quantile index set of interest, and can be taken generally as the union of multiple disjoint intervals. In what follows, let $|A|$ denote the cardinality of a set A, $M^\star(\tau) = \{1 \le j \le p : \beta_{\tau,j}^\star \neq 0\}$, and $M^\star = \cup_{\tau\in\Theta} M^\star(\tau)$.

We consider ultra-high dimensional data here, namely $\log(p) = o(n^{\xi_0})$ with $\xi_0 > 0$, in which a large number of predictors are irrelevant to the response. Examples of such data include gene expression microarray data, single nucleotide polymorphism data and high-frequency financial data (Ma, Li and Tsai (2017)). Two common sparsity assumptions for $\tilde{\boldsymbol{\beta}}_\tau^\star \equiv (\beta_{\tau,1}^\star, \ldots, \beta_{\tau,p}^\star)^\top \in \mathbb{R}^p$ arise to ensure the model interpretability and identifiability: local sparsity (LS) condition (Belloni and Chernozhukov (2011)) and global sparsity (GS) condition (Zheng, Peng and He (2015)). The LS condition assumes that $|M^\star(\tau)| = o(n)$, which tends to cause over-fitting phenomenon by simply taking the union of active covariate sets selected separately for each $\tau \in \Theta$. The GS condition assumes $|M^\star| = o(n)$, which is indispensable to derive a parsimonious model. Thus, we employ the GS assumption for variable screening to identify all the significant covariates related to the interesting segment of the conditional distribution of the response.

2.2 Sparse Composite Quantile Regression

We approximate $\boldsymbol{\beta}_\tau$ by a piecewise constant function with respect to $\tau \in \Theta$. Specifically, denote τ_0 and τ_K be the infimum and supremum of Θ, respectively. Let $\tau_0 < \cdots < \tau_K$ be a partition of Θ, and define the approximate function as $\bar{\boldsymbol{\beta}}_\tau = \sum_{k=1}^K \boldsymbol{\beta}_{\tau_k} I(\tau_{k-1} < \tau \le$

$\tau_k) \equiv (\bar{\beta}_{\tau,0}, \bar{\beta}_{\tau,1}, \ldots, \bar{\beta}_{\tau,p})^\top$ for $\tau \in \Theta$, where $\boldsymbol{\beta}_{\tau_k} = (\beta_{\tau_k,0}, \beta_{\tau_k,1}, \ldots, \beta_{\tau_k,p})^\top \in \mathbb{R}^{p+1}$, and $I(\cdot)$ denotes an indicator function. Define $D = (\boldsymbol{\beta}_{\tau_1}, \ldots, \boldsymbol{\beta}_{\tau_K}) \equiv (\boldsymbol{d}_0, \ldots, \boldsymbol{d}_p)^\top \in \mathbb{R}^{(p+1)\times K}$. Thus, to determine whether $\beta_{\tau,j} \equiv 0$ over Θ reduces to identify whether $\boldsymbol{d}_j$ is a zero vector or not $(1 \le j \le p)$. The latter is a row-wise sparsity problem for the coefficient matrix D, and hence we can utilize the group learning method.

Suppose that the observed data consist of n independent and identically distributed replicates of $(Y, \mathbf{X}^\top)^\top$, denoted by $\{(Y_i, \mathbf{X}_i^\top)^\top, i = 1, \ldots, n\}$. We employ composite quantile regression (CQR) in Zou and Yuan (2008) to estimate D. Let $\mathcal{U}_n(D) = (nK)^{-1} \sum_{k=1}^{K} \sum_{i=1}^{n} \rho_{\tau_k}(Y_i - \mathbf{X}_i^\top \boldsymbol{\beta}_{\tau_k})$ be the objective function, where $\rho_\tau(u) = u\{\tau - I(u < 0)\}$ is the check function (Koenker (2005)). Based on the GS condition, we consider the following problem:

$$\min_{D} \ \mathcal{U}_n(D) \qquad \text{subject to} \qquad \sum_{j=1}^{p} I(\|\boldsymbol{d}_j\|_2 \neq 0) \le t, \tag{2}$$

where t is a positive integer. Note that t controls the sparse level in problem (2). If we take $t < n$, then there are at least $(p-t)$ covariates screened out from model (1). Let $\hat{D} = (\hat{\boldsymbol{\beta}}_{\tau_1}, \ldots, \hat{\boldsymbol{\beta}}_{\tau_K})$ be a minimizer of problem (2). An efficient algorithm is proposed to solve problem (2) in Section 3. Denote $\hat{\boldsymbol{\beta}}_\tau = \sum_{k=1}^{K} \hat{\boldsymbol{\beta}}_{\tau_k} I(\tau_{k-1} < \tau \le \tau_k) \equiv (\hat{\beta}_{\tau,0}, \hat{\beta}_{\tau,1}, \ldots, \hat{\beta}_{\tau,p})^\top$ as the estimate of $\bar{\boldsymbol{\beta}}_\tau$, which is the approximation of $\boldsymbol{\beta}_\tau^\star$, and define $\hat{M}_t$ as the selected model index by using $\hat{\boldsymbol{\beta}}_\tau$, that is, $\hat{M}_t = \cup_{\tau\in\Theta}\{1 \le j \le p : \hat{\beta}_{\tau,j} \not\equiv 0\}$.

Since our method is a group learning method with a sparsity constraint for composite quantile regression, we call it as sparse composite quantile regression (SCQR). The main difference between the CQR and SCQR is that the coefficients $\beta_{\tau,j}$ are deemed as con-

stants over $\tau \in \Theta$ in the CQR but they are a group of functions in the SCQR. Besides, the proposed procedure employs the joint effects of candidate variables, which makes it distinct from marginal screening methods.

Let $s = |M^\star|$ be the true mode size. As guaranteed by Theorem 3 in Section 4, one has that $\hat{M}_s = M^\star$ holds with probability tending to one under certain regularity conditions. However, s is unknown and needs to be estimated in practice. Motivated by Wang (2009), we can derive a solution path by problem (2) and adopt an EBIC to estimate s. Specifically, let $\tilde{t} < n$ be a prespecified positive integer. We solve problem (2) for given $t \in \{1, \ldots, \tilde{t}\}$, and get a solution path of candidate models: $\{\hat{M}_1, \ldots, \hat{M}_{\tilde{t}}\}$. Theorem 3 implies that for choosing $\tilde{t} \geq s$, one can always guarantee that $M^\star$ is contained in one of the candidate models $\{\hat{M}_1, \ldots, \hat{M}_{\tilde{t}}\}$ with an overwhelming probability. For $\mathbf{X}_i = (1, x_{i1}, \ldots, x_{ip})^\top$ and an arbitrary subset $M \subset \{1, \ldots, p\}$, let $\mathbf{X}_{i,M}$ be the subvector of $\mathbf{X}_i$ consisting of all x_{ij} with $j \in M$. Also $\hat{\boldsymbol{\beta}}_{\tau_k, M}$ is similarly defined for $1 \leq k \leq K$. The EBIC is defined as

$$\mathrm{EBIC}(\hat{M}_t) = \log\left\{\frac{1}{nK}\sum_{k=1}^{K}\sum_{i=1}^{n}\rho_{\tau_k}\left(Y_i - \mathbf{X}_{i,\hat{M}_t}^\top \hat{\boldsymbol{\beta}}_{\tau_k,\hat{M}_t}\right)\right\} + C_n\frac{t\log(n)}{n},$$

where C_n is a positive constant that diverges along with the sample size n. We determine a hard-thresholding parameter $\hat{t}$ by $\hat{t} = \arg\min_{1\leq t\leq \tilde{t}} \mathrm{EBIC}(\hat{M}_t)$. Then the final selected model is defined as $\hat{M} = \hat{M}_{\hat{t}}$.

Remark 1. Since there is a tradeoff between computation and model selection accuracy when choosing $\tilde{t}$ in practice, we set $\tilde{t} = [n^{1/5}\log(n)]$, where $[a]$ denotes the largest integer part of a. This empirical choice is analogous to the recommended $\tilde{t}$ values in Xu and Chen (2014), and it works well in both our simulation studies and real data analysis.

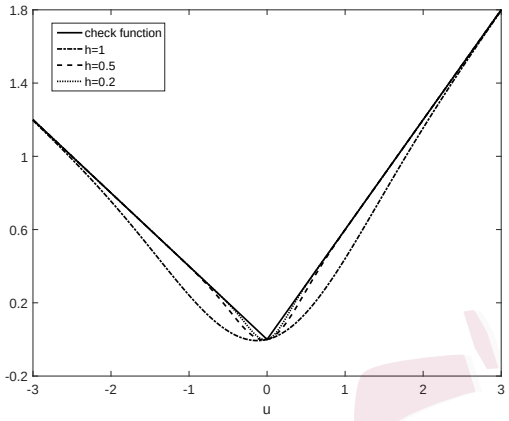

Figure 1: $\psi_{\tau,h}(u)$ is a smoothed approximation of $\rho_\tau(u)$

3 Computational Algorithm

Koenker and D'Orey (1987) developed parametric linear programming to compute a quantile regression function for all $\tau \in (0, 1)$. Many algorithms have been recently introduced for high-dimensional sparse penalized quantile regression approaches; see Gu et al. (2018) for an overview. For problem (2), there are C_p^t candidate submodels to fit the data for a given t, where C_p^t denotes the number of t-combinations from a given set of p elements. This will increase the computation burden of the existing algorithm. In addition, the check function $\rho_\tau(u)$ is not differentiable at point $u = 0$. To overcome these issues, we develop a high efficient algorithm to solve problem (2), which combines a smoothing technique and an iterative hard-thresholding algorithm.

First, we approximate the indicator function $I(u < 0)$ in $\rho_\tau(u)$ by a local distribution function $\Phi(-u/h)$, where $\Phi(\cdot)$ is the standard normal cumulative distribution function and h is a bandwidth that converges to 0 as $n \to \infty$. This method is originally devised

by Heller (2007) for rank regression. Define $\psi_{\tau,h}(u) = u\{\tau - \Phi(-u/h)\}$, which is smooth and differentiable at point $u = 0$. Note that if $u \geq 0$, $\psi_{\tau,h}(u) \to u\tau$ as $n \to \infty$, whereas if $u < 0$, $\psi_{\tau,h}(u) \to u(\tau - 1)$. Figure 1 illustrates that $\rho_\tau(u)$ can be approximated well by $\psi_{\tau,h}(u)$ with an appropriate h. Thus, a smoothed version of problem (2) is given as follows:

$$\min_{D} \tilde{\mathcal{U}}_n(D) \qquad \text{subject to} \qquad \sum_{j=1}^{p} I(\|\boldsymbol{d}_j\|_2 \neq 0) \leq t, \tag{3}$$

where $\tilde{\mathcal{U}}_n(D) = (nK)^{-1} \sum_{k=1}^{K} \sum_{i=1}^{n} \psi_{\tau_k,h}(Y_i - \mathbf{X}_i^\top \boldsymbol{\beta}_{\tau_k})$. If the bandwidth h satisfies $nh \to \infty$ and $nh^4 \to 0$ as $n \to \infty$, then Lemma 1 in the online Supplement Material indicates that the check function is equivalent to the smoothed version with probability tending to one. Thus, we can focus on solving problem (3). For the bandwidth, we used the rule of thumb bandwidth and chose $h = O(n^{-1/3})$. Let $\ell_\tau(\boldsymbol{\beta}) = n^{-1} \sum_{i=1}^{n} \psi_{\tau,h}(Y_i - \mathbf{X}_i^\top \boldsymbol{\beta})$, and denote $\dot{f}(\cdot)$ and $\ddot{f}(\cdot)$ as the first and second derivatives of any function $f(\cdot)$, respectively. Consider the following quadratic approximation to $\ell_\tau(\boldsymbol{v})$:

$$\varphi_\tau(\boldsymbol{u}|\boldsymbol{v}) = \ell_\tau(\boldsymbol{v}) + \langle \boldsymbol{u} - \boldsymbol{v}, \dot{\ell}_\tau(\boldsymbol{v}) \rangle + \frac{\lambda}{2} \|\boldsymbol{u} - \boldsymbol{v}\|_2^2, \tag{4}$$

where $\langle \cdot, \cdot \rangle$ denotes the inner product in the Euclidean space, and λ is a pre-specified positive constant. It can be seen that $\varphi_\tau(\boldsymbol{v}|\boldsymbol{v}) = \ell_\tau(\boldsymbol{v})$, and thus $\varphi_\tau(\boldsymbol{u}|\boldsymbol{v})$ nicely approximates $\ell_\tau(\boldsymbol{v})$ for $\boldsymbol{u}$ close to $\boldsymbol{v}$. Let $B = (\check{\boldsymbol{\beta}}_{\tau_1}, \ldots, \check{\boldsymbol{\beta}}_{\tau_K}) \equiv (\boldsymbol{b}_0, \ldots, \boldsymbol{b}_p)^\top \in \mathbb{R}^{(p+1)\times K}$. In view of equation (4), the smoothed composite quantile function $\tilde{\mathcal{U}}(\cdot)$ can be approximated by

$$\mathcal{Q}_\lambda(B|D) \equiv \frac{1}{K} \sum_{k=1}^{K} \varphi_{\tau_k}(\check{\boldsymbol{\beta}}_{\tau_k}|\boldsymbol{\beta}_{\tau_k}) = \tilde{\mathcal{U}}(D) + \frac{1}{K} \sum_{k=1}^{K} \langle \check{\boldsymbol{\beta}}_{\tau_k} - \boldsymbol{\beta}_{\tau_k}, \dot{\ell}_{\tau_k}(\boldsymbol{\beta}_{\tau_k}) \rangle + \frac{\lambda}{2K} \|B - D\|_F^2,$$

where $\|A\|_F$ is the Frobenius norm of an arbitrary matrix A. Using $\mathcal{Q}_\lambda(B|D)$, we can obtain an iterative solution to problem (3). Specifically, let $D^{[l]}$ be the estimate of D at the lth iteration. We update $D^{[l]}$ by $D^{[l+1]}$, where

$$D^{[l+1]} = \arg\min_B \ \mathcal{Q}_\lambda(B|D^{[l]}) \quad \text{subject to} \quad \sum_{j=1}^{p} I(\|\boldsymbol{b}_j\|_2 \neq 0) \leq t.$$

It is also equivalent to

$$D^{[l+1]} = \arg\min_B \ \left\|B - [D^{[l]} - \frac{1}{\lambda}\dot{\Psi}(D^{[l]})]\right\|_F^2 \quad \text{subject to} \quad \sum_{j=1}^{p} I(\|\boldsymbol{b}_j\|_2 \neq 0) \leq t, \tag{5}$$

where $\dot{\Psi}(D) = (\dot{\ell}_{\tau_1}(\boldsymbol{\beta}_{\tau_1}), \ldots, \dot{\ell}_{\tau_K}(\boldsymbol{\beta}_{\tau_K})) \in \mathbb{R}^{(p+1)\times K}$.

Proposition 1. *Let $D = (\boldsymbol{d}_0, \ldots, \boldsymbol{d}_p)^\top \in \mathbb{R}^{(p+1)\times K}$ be an arbitrary matrix. If $\hat{B} = (\hat{\boldsymbol{b}}_0, \ldots, \hat{\boldsymbol{b}}_p)^\top$ is an optimal solution to the following problem*

$$\min_{B\in\mathbb{R}^{(p+1)\times K}} \|B - D\|_F^2 \quad \textit{subject to} \quad \sum_{j=1}^{p} I(\|\boldsymbol{b}_j\|_2 \neq 0) \leq t,$$

then $\hat{B}$ has a closed form with the jth row defined as

$$\hat{\boldsymbol{b}}_0 = \boldsymbol{d}_0 \ \textit{ and } \ \hat{\boldsymbol{b}}_j = \boldsymbol{d}_j I(d_j^* \geq d_{(t)}^*) \ \textit{ for } \ 1 \leq j \leq p, \tag{6}$$

where $d_j^* = \|\boldsymbol{d}_j\|_2$, and $d_{(t)}^*$ is the t-th largest value of $d_1^*, \ldots, d_p^*$.

The proof is given in the online Supplement Material. Proposition 1 indicates that equation (6) is indeed a hard-thresholding rule. It first ranks the importance of covariates according to the estimates of $\|\boldsymbol{d}_j\|_2$ in decreasing order, and then filters out those having small effects over Θ.

Based on Proposition 1, we get that $D^{[l+1]}$ defined in (5) has the following form:

$$\boldsymbol{d}_j^{[l+1]} = \check{\boldsymbol{d}}_j^{[l]} I(\|\check{\boldsymbol{d}}_j^{[l]}\|_2 \geq \check{d}_{(t)}) \ \text{ for } \ 1 \leq j \leq p, \tag{7}$$

where $\check{\boldsymbol{d}}_j^{[l]}$ is the transposition of the jth row of $[D^{[l]} - \lambda^{-1}\dot{\Psi}(D^{[l]})]$, and $\check{d}_{(t)}$ is the t-th largest value of $\|\check{\boldsymbol{d}}_1^{[l]}\|_2, \ldots, \|\check{\boldsymbol{d}}_p^{[l]}\|_2$.

However, there still exists a step-size λ in updating rule (7), which plays an important role in the convergence of the algorithm. Our empirical studies indicate that a large value of λ often leads to a slow convergence rate, while a small value of λ results in failing to identify active covariates. In what follows, a backtracking method is employed to find λ such that the objective function monotonically decreases after each iteration. Specifically, we choose the step-size $\lambda^{[l]}$ at the lth iteration as the minimum value such that

$$\tilde{\mathcal{U}}_n(D^{[l+1]}) \leq \tilde{\mathcal{U}}_n(D^{[l]}) - \frac{\varrho\lambda^{[l]}}{2K}\|D^{[l+1]} - D^{[l]}\|_F^2, \tag{8}$$

where $\varrho \in (0, 1)$ is a fixed small constant. The proposed algorithm is presented in the following Algorithm 1.

Algorithm 1. Let L be a pre-specified positive integer.

Step 1. Choose an initial value for $D^{[0]}$, such as $D^{[0]} = 0$;

Step 2. For each $l \in \{0, 1, \ldots, L\}$,

Step 2.1. Compute $D^{[l+1]}$ by equation (7);

Step 2.2. Stop Step 2 if the linear search criterion (8) is satisfied; otherwise, take the step-size to be $2\lambda^{[l]}$ and return to *Step 2.1*;

Step 3. Stop the algorithm if $l > L$ or $\|D^{[l+1]} - D^{[l]}\|_F < \delta\|D^{[l]}\|_F$, where $\delta > 0$ is a prespecified tolerance parameter. Otherwise, increase l, and return to *Step 2.1*.

In our simulation studies and real data analysis, we take $L = 1000$, and set $\delta = \varrho = 10^{-5}$.

4 Theoretical Properties

4.1 Convergence Analysis of Algorithm

To show the convergence property of the proposed algorithm, we need the following Lipschitz condition:

$$\|\dot{\ell}_\tau(\boldsymbol{\beta}_1) - \dot{\ell}_\tau(\boldsymbol{\beta}_2)\|_2 \leq \phi\|\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2\|_2,$$

where ϕ is a positive constant independent of τ. The Lipschitz condition is satisfied if the largest eigenvalue of $\ddot{\ell}_\tau(\boldsymbol{\beta})$ is uniformly bounded in $\boldsymbol{\beta}$ and τ. A more serious concern is whether for each $l \geq 0$ the step size $\lambda^{[l]}$ is bounded or not. Following similar arguments in Gong et al. (2013), the Lipschitz condition, together with criterion (8), guarantees the boundedness of the step size $\lambda^{[l]}$ in Step 2.2. The following theorem summarizes the convergence property of Algorithm 1.

Theorem 1. *Let $\{D^{[l]}\}$ be the sequence generated by Algorithm 1. If $\lambda^{[l]} > \phi/(1-\varrho)$, then as $l \to \infty$, there exists at least one subsequence such that $\{D^{[l]}\}$ is convergent. In addition, if the stopping criterion is $K^{-1/2}\|D^{[l+1]} - D^{[l]}\|_F \leq \varepsilon$, we have that Algorithm 1 stops in a finite number of steps, where $\varepsilon > 0$ is a prespecified small constant.*

The proof can be found in the online Supplement Material, which indicates that the proposed algorithm yields an approximate solution. The next theorem presents an upper

bound for the estimation error of $D^{[l]}$. Let $\tilde{\phi} = \min_{0<\|x\|_0\leq 3t}\{x^\top \ddot{\ell}_\tau(\boldsymbol{\beta})x\}/(x^\top x) > 0$ be the restricted eigenvalue, where $\|a\|_0 = \sum_{j=1}^p I(a_j \neq 0)$ for a vector $a = (a_1, \ldots, a_p)^\top \in \mathbb{R}^p$. The restricted eigenvalue condition is frequently used in the literature of high dimensional data analysis (Candes and Tao (2007); Belloni and Chernozhukov (2011)). Let $D^\star$ denote the true value of D.

Theorem 2 (Upper Bound of Algorithm 1)**.** *If $s \leq t$ and $\phi < \lambda^{[l]} < \tilde{\phi}/\{1 - 1/(4\sqrt{2})\}$, then*

$$\|D^{[l]} - D^\star\|_F \leq 2^{-l}\|D^{[0]} - D^\star\|_F + \sqrt{\frac{8}{\phi}}\|\dot{\Psi}(D^\star)\|_F.$$

Theorem 2, combining with the convergence property of Algorithm 1, implies that there exists at least one subsequence such that the difference between the limiting point and the true value $D^\star$ can be bounded by $\|\dot{\Psi}(D^\star)\|_F$. Moreover, if we take the initial value $D^{[0]} = 0$, after at most $l = [\log_2(\|D^\star\|_F/\|\dot{\Psi}(D^\star)\|_F)] + 1$ iterations, the sequence $\{D^{[l]}\}$ satisfies that $\|D^{[l]} - D^\star\|_F \leq (1 + \sqrt{8/\phi}\|\dot{\Psi}(D^\star)\|_F$. Thus, in a finite number of steps, the estimation error can be controlled by $\|\dot{\Psi}(D^\star)\|_F$.

4.2 Sure Screening Property

Let M be an arbitrary subset of $\{1, \ldots, p\}$, and $M_t = \{M : |M| \leq t\}$. Define the collections of over-fitted models with model size t as $M_+^t = \{M : M^\star \subset M_t\}$. To study the asymptotic properties of the proposed SCQR, we need the following regularity conditions:

(C1) $\log(p) = o(n^{\xi_0})$ for $0 < \xi_0 < 1$.

(C2) There exist some positive constants $\omega_1,\ \omega_2,\ \xi_1$ and ξ_2 such that for a given hard-thresholding parameter t in (2), the true mode size $s \le t < \omega_1 n^{\xi_1}$, and

$$\min_{j\in M^\star}\Big[\int_\Theta (\beta^\star_{\tau,j})^2 d\tau\Big]^{1/2} \ge \omega_2 n^{-\xi_2}.$$

Condition (C2) suggests that the minimum signal of the active set is bounded away from zero, but it is allowed to converge to zero in order $O(n^{-\xi_2})$. This encompasses what is considered by Xu and Chen (2014) for the generalized linear model.

(C3) Let $\epsilon = Y - \mathbf{X}^\top \boldsymbol{\beta}^\star_\tau$, and $F(\cdot|\mathbf{x})$ and $f(\cdot|\mathbf{x})$ be the cumulative distribution function and density function of ϵ given $\mathbf{X} = \mathbf{x}$, respectively. There exist positive constants ν and δ^* free of τ such that for sufficiently large n and each vector $\boldsymbol{u} \in \{\boldsymbol{v} : \|\boldsymbol{v}_M\|_2 < \delta^*,\ M \in M^{2t}_+\}$,

$$\frac{1}{n^{1-\xi_2}}\sum_{i=1}^n \int_0^{\mathbf{X}_i^\top \boldsymbol{u}} \Big[F\big(\frac{s}{n^{\xi_2}}|\mathbf{X}_i\big) - F\big(0|\mathbf{X}_i\big)\Big] ds \ge \nu \|\boldsymbol{u}\|_2^2.$$

Condition (C3) is similar to condition (2) of Zou and Yuan (2008), which is used to establish the asymptotic properties of composite quantile regression. Indeed, condition (C3) can be replaced by some sufficient conditions that are commonly used in quantile regression. Some examples are given in the online Supplement Material.

(C4) For $\boldsymbol{X}_i = (1, x_{i1}, \ldots, x_{ip})^\top$, there exists a positive constant m such that $\sup_{i,j} |x_{ij}| \le m$.

Condition (C4) is commonly used in the context of high-dimensional data analysis (Wang, Wu and Li (2012); Lee, Noh and Park (2014)). This assumption can be relaxed

to a tail probability inequality that there exist some positive constants m_0, m_1 and α such that for sufficiently large η, $P\{|x_{ij}| > \eta\} \leq m_0 \exp\{-m_1\eta^\alpha\}$. In this case, the theoretical results still hold with slight modifications in the proofs.

Theorem 3 (Sure Screening Property)**.** *Suppose that conditions (C1)-(C4) hold with* $\max\{\xi_1 + \xi_2, \xi_1/2 + 2\xi_2\} < (1 - \xi_0)/2$. *Then for sufficiently large* K,

$$P\big\{M^\star \subset \hat{M}_t\big\} \to 1 \quad as \quad n \to \infty.$$

Theorem 3 states that with probability tending to one, all relevant variables can be identified by carrying out the SCQR within at most $O(n^{\xi_1})$ times, which is a number much smaller than n under condition $\max\{\xi_1 + \xi_2, \xi_1/2 + 2\xi_2\} < (1 - \xi_0)/2$. Based on Theorem 3, the strong screening consistency (Huang, Li and Wang (2014)) is further provided in the following corollary.

Corollary 1. *Under the conditions of Theorem 3, we have*

$$P\big\{M^\star = \hat{M}_s\big\} \to 1 \quad as \quad n \to \infty.$$

Corollary 1 suggests that if one has prior knowledge on the model size s, the selected model $\hat{M}_s$ is exactly the true model $M^\star$ with probability approaching one. This corollary is important because it guarantees that the true model is one of our candidate models $\{\hat{M}_1, \ldots, \hat{M}_{\tilde{t}}\}$ as long as $\tilde{t} \geq s$. The consistency property for the EBIC procedure is established in the following theorem.

Theorem 4. *Suppose that conditions (C1)-(C4) hold with* $\max\{\xi_1 + \xi_2, \xi_1/2 + 2\xi_2\} < (1 -$

$\xi_0)/2$. *If* $E(|\epsilon|) < \infty$, $C_n^{-1} = o(1)$, *and* $C_n \log(n)/(n^{1-\xi_1}) = o(1)$. *Then* $P\{M^\star = \hat{M}\} \to 1$ *as* $n \to \infty$.

Theorem 4 suggests that with probability approaching one, the true model index can be correctly identified by the SCQR when the EBIC is employed as the stopping criterion.

Theorem 5. *Under conditions (C1)-(C4) with* $\max\{\xi_1 + \xi_2, \xi_1/2 + 2\xi_2\} < (1 - \xi_0)/2$, *we have that there exists a constant* $c_0 > 0$ *such that*

$$P\Big\{\Big[\int_\Theta \|\hat{\boldsymbol{\beta}}_\tau - \boldsymbol{\beta}_\tau^\star\|_2^2 d\tau\Big]^{1/2} \geq c_0 n^{-\xi_2}\Big\} \to 0 \quad as \quad n \to \infty.$$

Theorem 5 indicates that the integral square error of the proposed estimate can be bounded by $O_p(n^{-\xi_2})$. This, combining with Theorem 3, implies that the SCQR procedure can perform variable screening and parameter estimation simultaneously. The consistency property of Algorithm 1 is guaranteed by the following theorem.

Theorem 6. *Suppose that conditions (C1)-(C4) hold with* $\max\{\xi_1 + \xi_2, \xi_1/2 + 2\xi_2\} < (1 - \xi_0)/2$. *If* $E(|\epsilon|) < \infty$ *and* $\phi < \lambda^{[l]} < \tilde{\phi}/\{1 - 1/(4\sqrt{2})\}$. *Then there exists a constant* $c_1 > 0$ *such that after* $l = [\log_2(\|D^\star\|_F/\|\dot{\Psi}(D^\star)\|_F)] + 1$ *iterations,*

$$P\Big\{\Big[\int_\Theta \|\hat{\boldsymbol{\beta}}_\tau^{[l]} - \boldsymbol{\beta}_\tau^\star\|_2^2 d\tau\Big]^{1/2} \geq c_1 n^{-\xi_2}\Big\} \to 0 \quad as \quad n \to \infty.$$

Theorems 5 and 6 indicate that the estimates generated by Algorithm 1 and problems (2) and (3) have the same consistency rate $O_p(n^{-\xi_2})$. Theorem 6 also implies the following result, which indicates the sure screening property of Algorithm 1.

Corollary 2 (Sure Screening of Algorithm 1)**.** *Under the conditions of Theorem 6, we have that after* $l = [\log_2(\|D^\star\|_F/\|\dot{\Psi}(D^\star)\|_F)] + 1$ *iterations,*

$$P\big\{M^\star \subset \hat{M}_t^{[l]}\big\} \to 1 \quad as \quad n \to \infty.$$

Remark 2. To guarantee the sure screening property, Xu and Chen (2014) proposed to use an appropriate Lasso-type initial value in their algorithm. However, the Lasso-type estimate may be unstable and time-consuming under ultra-high dimensional settings. Corollary 2 generalizes their results and states that zero is a reasonable initial value for Algorithm 1. This finding further enriches the SCQR method from a practical perspective.

5 Real Data Analysis

In this section, the proposed method is applied to a gene expression dataset to investigate gene regulation in the mammalian eye and to identify genetic variations relevant to human eye disease (Scheetz et al. (2006)). This dataset has 31,042 gene expression probe sets on 120 rats, and the gene expression levels are analyzed on a log scale with base 2. The response variable of interest is the expression of gene TRIM32 (probe 1389163_at), which is known to cause human hereditary diseases of the retina. As in Huang, Ma and Zhang (2008), Wang, Wu and Li (2012), and Zheng, Peng and He (2015), the main aim of this analysis is to study how the response variable depends on the gene expression of other probes. The dataset is available in **R** package *"flare",* which has been processed to exclude probes that are not expressed or lack variation. There are 200 probes left as covariates.

As in Zheng, Peng and He (2015), two reasonable choices for Θ are considered: $(0.2, 0.8)$ and $(0.25, 0.75)$. The bandwidth is chosen as $h = 1.9n^{-1/3}$ and $C_n = \log(p)/2$ in the EBIC. For comparison, two other methods are also considered: our proposed method with Θ degenerating to one point τ, denoted by SQR(τ) with $\tau \in \{0.25 + 0.05k, k = 0, 1, \ldots, 10\}$; the method of simply taking the union of active covariate sets identified by SQR(τ) at each τ, denoted by USQR. To evaluate each method, we consider 400 random partitions. For each partition, the data are randomly divided into two equal datasets: a training dataset and a testing dataset. Based on the training dataset, we implement the screening methods and obtain the estimate of $\boldsymbol{\beta}_\tau$. Subsequently, we compute the prediction error:

$$\text{PE}(\Theta) = \frac{1}{|\mathcal{T}|} \sum_{i \in \mathcal{T}} \int_\Theta \rho_\tau(Y_i - \mathbf{X}_i^\top \hat{\boldsymbol{\beta}}_\tau) d\tau,$$

where $\mathcal{T} = \{i : \text{the } i\text{th subject in the testing dataset}\}$ is the testing set index. For the SQR(τ), we treat the coefficient functions as constants over $\tau \in \Theta$ and calculate PE(Θ). A smaller value of the prediction error indicates a better performance.

The results averaged over 400 random partitions are reported in Table 1. The table indicates that the SCQR procedure selects four same genes that are significantly related to the response variable for two different choices of Θ. This suggests that the SCQR method is robust to the selection of Θ, which is a desirable feature from the perspective of model selection. For the SQR(τ) method, the chosen set of probes varies with τ. For instance, the genes 1370551_a_at and 1398389_at are selected by the SQR(τ) with τ from 0.4 to 0.6, but they are overlooked at lower and higher τ's. These may suggest a heterogeneous

relationship across different quantile levels. Further, the results indicate that 3 probes are selected by the SQR(0.4), but no probe is selected by the SQR(0.35). This implies that the SQR(τ) method may be sensitive to the choice of τ. For the USQR procedure, a total number of 5 probes are selected both for $\Theta = (0.25, 0.75)$ and $\Theta = (0.2, 0.8)$. Compared to the selection results of the USQR, the SCQR yields slightly smaller predictor errors.

6 Conclusions

This article considered a sparse composite quantile regression method for analyzing ultra-high dimensional heterogeneous data across a continuous range of quantile levels. An efficient iterative algorithm was developed to implement our proposed method. The properties of the proposed procedure were provided. Specifically, the theoretical results suggest that the SCQR method with ultra-high dimensional covariates can successfully identify active covariates with probability approaching one. Meanwhile, the SCQR method yields consistent estimates of coefficients. Furthermore, the proposed algorithm enjoys consistent properties in terms of variable screening and parameter estimation.

Supplementary Material

The online Supplementary Material includes simulation studies, some sufficient conditions for (C3) and the proofs of Proposition 1 and Theorems 1-6.

Acknowledgement

The authors thank the Co-Editor, Professor Hans-Georg Müller, an associate editor, and two referees for their insightful comments and suggestions that greatly improved the article. This research was partly supported by the National Natural Science Foundation of China (Nos. 11771431, 11690015 and 11901087), the Key Laboratory of RCSDS, CAS (No. 2008DP173182), the Hubei Natural Science Foundation of China (No. 2018CFB256), the Fundamental Research Funds for the Central Universities (No. CCNU19QN084), and the Program for Young Excellent Talents, UIBE (No. 19YQ15).

References

Belloni, A. and Chernozhukov, V. (2011). ℓ_1-penalized quantile regression in high dimensional sparse models. *The Annals of Statistics* **39**, 82-130.

Candes, E. and Tao, T. (2007). The Dantzig selector: statistical estimation when p is much larger than n (with discussion). *The Annals of Statistics* **35**, 2313-2404.

Fan, J., Fan, Y. and Barut, E. (2014). Adaptive robust variable selection. *The Annals of Statistics* **42**, 324-351.

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

Fan, Y., Kong, Y., Li, D. and Lv, J. (2017). Interaction pursuit with feature screening and selection. Available at arXiv:1605.08933.

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

Fan, J. and Lv, J. (2010). A selective overview of variable selection in high dimensional feature space. *Statistica Sinica* **20**, 101-148.

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

Fan, J. and Song, R. (2010). Sure independence screening in generalized linear models with NP-dimensionality. *The Annals of Statistics* **38**, 3567-3604.

Gong, P., Zhang, C., Lu, Z., Huang, J. and Ye, J. (2013). A general iterative shrinkage and thresholding algorithm for non-convex regularized optimization problems. *Proceedings of the 30th International Conference on Machine Learning* **28**, 37-45.

Gu, Y., Fan, J., Kong, L., Ma, S. and Zou, H. (2018). ADMM for high-dimensional sparse penalized quantile regression. *Technometrics* **60**, 319-331.

He, X., Wang, L. and Hong, H. (2013). Quantile-adaptive model-free variable screening for high-dimensional heterogeneous Data. *The Annals of Statistics* **41**, 342-69.

Heller, G. (2007). Smoothed rank regression with censored data. *Journal of the American Statistical Association* **102**, 552-559.

Huang, D., Li, R. and Wang, H. (2014). Feature screening for ultrahigh dimensional categorical data with applications. *Journal of Business and Economic Statistics* **32**, 237-244.

Huang, J., Ma, S. and Zhang, C.-H. (2008). Adaptive Lasso for sparse high-dimensional regression models. *Statistica Sinica* **18**, 1603-1618.

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

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

Koenker, R. and D'Orey, V. (1987). Algorithm as 229: Computing regression quantiles. *Journal of the Royal Statistical Society, Series C* **36**, 383-393.

Kong, Y., Li, D., Fan, Y. and Lv, J. (2017). Interaction pursuit in high-dimensional multi-response regression via distance correlation. *The Annals of Statistics* **45**, 897-922.

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

Li, X., Ma, X. and Zhang, J. (2018). Conditional quantile correlation screening procedure for ultrahigh-dimensional varying coefficient models. *Journal of Statistical Planning and Inference* **197**, 69-92.

Liu, J., Li, R. and Wu, R. (2014). Feature selection for varying coefficient models with

ultrahigh-dimensional covariates. *Journal of the American Statistical Association* **109**, 266-274.

Ma, S., Li, R. and Tsai, C-L. (2017). Variable screening via quantile partial correlation. *Journal of the American Statistical Association* **112**, 650-663.

Ma, X. and Zhang, X. (2016). Robust model-free feature screening via quantile correlation. *Journal of Multivariate Analysis* **143**, 472-480.

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

Scheetz, T., Kim, K.-Y., Swiderski, R., Philp, A., Braun, T., Knudtson, K., Dorrance, A., Dibona, G., Huang, J., Casavant, T., Sheffield, V. and Stone, E. (2006). Regulation of gene expression in the mammalian eye and its relevance to eye disease. *Proceedings of the National Academy of Sciences of the United States of America* **103**, 14429-14434.

Song, R., Lu, W., Ma, S. and Jeng, J. (2014). Censored rank independence screening for high-dimensional survival data. *Biometrika* **101**, 799-814.

Wang, H. (2009). Forward regression for ultra-high dimensional variable screening. *Journal of the American Statistical Association* **104**, 1512-1524.

Wang, L., Wu, Y. and Li, R. (2012). Quantile regression for analyzing heterogeneity in ultra-high dimension. *Journal of the American Statistical Association* **107**, 214-222.

Wu, Y. and Yin, G. (2015). Conditional quantile screening in ultrahigh-dimensional heterogeneous data. *Biometrika* **102**, 65-76.

Xu, C. and Chen, J. (2014). The sparse MLE for ultrahigh-dimensional feature screening. *Journal of the American Statistical Association* **109**, 1257-1269.

Xu, K. (2017). Model-free feature screening via a modified composite quantile correlation. *Journal of Statistical Planning and Inference* **188**, 22-35.

Yang, G., Hou, S., Wang, L. and Sun, Y. (2018). Feature screening in ultrahigh-dimensional additive Cox model. *Journal of statistical computation and simulation* **88**, 1117-1133,

Zhang, S. and Zhou, Y. (2018). Variable screening for ultrahigh dimensional heterogeneous data via conditional quantile correlations. *Journal of Multivariate Analysis* **165**, 1-13.

Zheng, Q., Peng, L. and He, X. (2015). Globally adaptive quantile regression with ultra-high dimensional data. *The Annals of Statistics* **43**, 2225-2258.

Zheng, Q., Peng, L. and He, X. (2018). High dimensional censored quantile regression. *The Annals of Statistics* **46**, 308-343.

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

Zou, H. and Yuan, M. (2008). Composite quantile regression and the oracle model selection theory. *The Annals of Statistics* **36**, 1108-1126.

Table 1. *Probe sets identified by various methods.*

			PE(Θ)	
Θ	Method	Probes	$[0.25, 0.75]$	$[0.2, 0.8]$
$[0.25, 0.75]$	SCQR	"1370551_a_at, 1374106_at, 1384862_at, 1389457_at"	0.020(0.002)	-
$[0.2, 0.8]$	SCQR	"1370551_a_at, 1374106_at, 1384862_at, 1389457_at"	-	0.028(0.003)
$[0.25, 0.75]$	USQR	5 probes	0.034(0.002)	-
$[0.2, 0.8]$	USQR	5 probes	-	0.042(0.002)
0.25	SQR	0 probes	0.040(0.003)	0.047(0.003)
0.30	SQR	0 probes	0.038(0.003)	0.046(0.004)
0.35	SQR	0 probes	0.035(0.003)	0.042(0.004)
0.40	SQR	"1370551_a_at, 1384886_at, 1398389_at"	0.029(0.004)	0.035(0.005)
0.45	SQR	"1370429_at, 1370551_a_at, 1398389_at"	0.022(0.002)	0.027(0.003)
0.50	SQR	"1370551_a_at, 1398389_at"	0.021(0.003)	0.025(0.003)
0.55	SQR	"1370551_a_at, 1374106_at, 1398389_at"	0.022(0.002)	0.027(0.003)
0.60	SQR	"1370429_at, 1370551_a_at, 1398389_at"	0.028(0.004)	0.034(0.004)
0.65	SQR	0 probes	0.032(0.004)	0.039(0.005)
0.70	SQR	0 probes	0.034(0.004)	0.043(0.004)
0.75	SQR	0 probes	0.037(0.003)	0.045(0.004)