Quantile-adaptive variable screening in ultra-high dimensional varying coefficient models

Junying Zhang\textsuperscript{a,b}, Riquan Zhang\textsuperscript{a,c,}\textsuperscript{*} and Zhiping Lu\textsuperscript{a}

\textsuperscript{a}School of Finance and Statistics, East China Normal University, Shanghai 200241, People’s Republic of China; \textsuperscript{b}Department of Mathematics, Taiyuan University of Technology, Taiyuan 030024, People’s Republic of China; \textsuperscript{c}Department of Mathematics, Shanxi Datong University, Datong 037009, People’s Republic of China

(Received 18 May 2015; accepted 9 July 2015)

The varying-coefficient model is an important nonparametric statistical model since it allows appreciable flexibility on the structure of fitted model. For ultra-high dimensional heterogeneous data it is very necessary to examine how the effects of covariates vary with exposure variables at different quantile level of interest. In this paper, we extended the marginal screening methods to examine and select variables by ranking a measure of nonparametric marginal contributions of each covariate given the exposure variable. Spline approximations are employed to model marginal effects and select the set of active variables in quantile-adaptive framework. This ensures the sure screening property in quantile-adaptive varying-coefficient model. Numerical studies demonstrate that the proposed procedure works well for heteroscedastic data.

\textbf{Keywords:} quantile regression; varying-coefficient independent screening; ultra-high dimension; heterogeneous data; variable selection; dimensionality reduction

\textit{AMS 2000 Subject Classifications:} Primary: 62G05; Secondary: 62E20

\section{Introduction}

Varying coefficient models have been widely applied as an adaptive generalization of the linear regression model to depict dynamic behaviors. Because of their flexibility and interpretability, much work has been done to perform parameter estimation and hypothesis testing, but mostly in the mean regression setting until some authors such as Honda \cite{11}, Cai and Xu \cite{1} and Kim \cite{12} paid attention to the quantile regression setting in the early 2000s.

Quantiles themselves were defined without moment conditions. Exploring several conditional quantiles can provide us with a more complete description of the data than exploring only the

\textsuperscript{*}Corresponding author. Email: zhangriquan@163.com

\copyright 2015 Taylor & Francis
conditional mean. These advantages have inspired many researchers to consider the quantile regression not only in parametric models but also in semi-parametric or nonparametric models. Regarding varying coefficient models, Honda [11] and Cai and Xu [1] explored a varying coefficient model for the conditional quantiles using local polynomials. Kim [12] proposed a polynomial spline-based methodology for the same model. For longitudinal data analysis, Wang et al. [17] developed the partially linear varying coefficient model in quantile regression. In spite of remarkable progresses in parameter estimation and hypothesis testing, there do not exist methods how to conduct variable selection efficiently for the varying coefficient model with ultra-high dimensional heterogeneous data in the quantile regression framework.

The ultra-high dimensional data are frequently encountered in a large variety of research areas such as genomic, functional magnetic resonance imaging, tomography, economics, and finance. Analysis of ultra-high dimensional data poses many challenges for statisticians and calls for new statistical methodologies and theories. With the sparsity assumption, variable selection improves the accuracy of estimation and gains scientific insights. Many significant variable selection methods have been developed, such as Bridge regression in [8], Lasso in [16], smoothly clipped absolute deviation (SCAD) and folded concave penalty in [4], the Elastic net in [23], Adaptive Lasso [22], and the Dantzig selector in [2]. Techniques on the implementation of folded concave penalized least-squares include the local linear approximation algorithm in [24] and the the plus algorithm in [19]. However, these methods do not perform well in ultra-high dimensional problems.

Fan and Lv [5] introduced a method of sure independent screening (SIS) to reduce dimensionality of the feature space to a moderate scale in ultra-high dimensional linear regression models via marginal correlation learning. By this method, some important variables are retained with probability approaching one. Fan and Song [7] extended the methodology to ultra-high dimensional generalized linear models and constructed a useful technical tool for establishing the sure screening results and bounding false selection rates. Other methods were put forward including data tilling method in [9], marginal partial likelihood method in [21], robust screening method by rank correlation in [14], distance correlation in [15], nonparametric independence screening (NIS) with additive models in [3], and variable screening in nonparametric varying-coefficient models with ultra-high dimensionality in [6]. Inspired by these previous work, our study will focus on variable screening in varying coefficient quantile regression models with ultra-high dimensionality, that is, dimensionality of nonpolynomial (NP) order of sample size.

The nonparametric quantile regression varying-coefficient model is studied under the conditional linear structure. Suppose that the observed data \( \{(X_i, Y_i), i = 1, \ldots, n\} \) be i.i.d. copy of \((X, Y)\), where \(X_i = (X_{i1}, \ldots, X_{ip})^T\). The covariates are denoted as \(X_1, \ldots, X_p\), so \(X\) also can be written as \((X_1, \ldots, X_p)^T\). The varying-coefficient model is that:

\[
Y = \beta^T(U)X + \epsilon, \tag{1}
\]

where \(U\) is some observable exposure variable, and \(\epsilon\) is the random noise with condition mean 0. The vector \(X\) is assumed to enter the model linearly, meanwhile it allows coefficient functions to vary smoothly with the exposure variable.

In the quantile-adaptive screening framework, we use \(B\)-spline approximation to estimate marginal effects of varying coefficient model. In this aspect, our method shares some similarity with that in [6]. The main technical challenge is to handle the nonsmooth loss function comparing with that in [6]. We derive the exponential bounds by applying the empirical process theory and establish the sure screening property under some general condition.

The rest of the article is organized as follows. In Section 2, we apply \(B\)-spline basis approximation and screening variables by ranking a measure of the estimators in marginal quantile nonparametric regression model. In Section 3 we derive the quantile nonparametric independent screening (QNIS) property under certain technical conditions. Iterative QNIS procedures
are established in Section 4. In Section 5, some simulation studies and a real data application are conducted to evaluate the performance of our proposed methods. Proofs are presented in Section 6.

2. Nonparametric quantile marginal screening

We assume that the functional coefficient vector $\beta(\cdot) = (\beta_1(\cdot), \ldots, \beta_p(\cdot))^T$ is sparse and that $\mathcal{M}_s = \{j, E\beta_j^2(U) > 0\}$ be the true sparse model with nonsparsity size $s_n = |\mathcal{M}_s|$. $p$ is allowed to grow with $n$ and sometimes written that $p = p_n$ whenever necessary.

2.1 Quantile-adaptive marginal regression

$Y$ and $X_j$ being independent means that $Q_\tau(Y|X_j, U) - Q_\tau(Y|U) = 0, \forall \tau \in (0, 1)$, where $Q_\tau(Y|X_j, U)$ is the $\tau$th conditional quantile of $Y$ given $X_j$ and exposure variable $U$, $Q_\tau(Y|U)$ is the $\tau$th quantile of $Y$ on $U$. So the set of active variable is equivalent to the set:

$$\mathcal{M}_s = \{1 \leq j \leq p : Q_\tau(Y|X_j, U) \text{ functional depend on } X_j\}.$$

We consider the marginal quantile regression of $Y$ on $X_j$ and $U$ and define

$$\{\hat{b}_{j_0}(U), \hat{b}_j(U)\} = \arg \min_{b_{j_0}(U), b_j(U) \in L_2(P)} \{E[\rho_\tau(Y - b_{j_0}(U) - b_j(U)X_j)]$$

$$- E(\rho_\tau(Y|U))\},$$

(2)

where $\rho_\tau (u) = u(\tau - I(u < 0))$ is quantile loss function. $P$ denotes the joint distribution of $(Y, U, X_j)$, $L_2(P)$ is the class of square integrable functions under the measure $P$ and $I(\cdot \cdot)$ denotes the indicator function. It is known that $Q_\tau(Y|X_j, U) = \hat{b}_{j_0}(U) + \hat{b}_j(U)X_j$. We rank the marginal effect of covariates by $||\tilde{f}_j||^2$, and define $||\tilde{f}_j||^2 = E(\tilde{f}_j^2)$,

$$\tilde{f}_j = \tilde{b}_{j_0}(U) + \tilde{b}_j(U)X_j - Q_\tau(Y|U),$$

(3)

where $Q_\tau(Y|U)$ is the $\tau$th quantile about $Y$.

Let $\{\hat{b}_{j_0}, \hat{b}_j\} = \arg \min_{b_{j_0}(U), b_j(U) \in L_2(P)} \sum_{i=1}^n \rho_\tau(Y_i - b_{j_0}(U_i) - b_j(U_i)X_{ji})$, and define

$$\hat{f}_j = \hat{b}_{j_0} + \hat{b}_jX_j - F_{\hat{Y},n}^{-1}(\tau).$$

(4)

where $F_{\hat{Y},n}^{-1}(\tau)$ is the $\tau$th sample quantile function based on $Y_1, \ldots, Y_n$ under $U$, $||a_j||^2_n = (1/n) \sum_{i=1}^n (a_{ji})^2$. Thus $\hat{f}_j$ is an estimator of $Q_\tau(Y|X_j, U) - Q_\tau(Y|U)$. If $X_j$ is independent of $Y$, $\hat{f}_j$ is close to 0. Now we select variable as follows,

$$\mathcal{M}_{\gamma_n} = \{1 \leq j \leq n : ||\hat{f}_j||^2_n \geq \gamma_n\},$$

(5)

where $\gamma_n$ is a predefined threshold value.

2.2 Marginal regression estimation with B-spline

The next step is to estimate the marginal effect $\hat{f}_j, j = 1, \ldots, p$, we approximate $b_{j_0}(U)$ and $b_j(U)$ by functions in $S_n$, the space of polynomial splines of degree $d_n \geq 1$ on $U$, a compact set. Let
\{B_k, k = 1, \ldots, d_n\} denote its normalized B-spline bases with \|B_k\|_{\infty} \leq 1, where \| \cdot \|_{\infty} is the supremum norm. Then

\[ b_j(U) \approx \sum_{k=1}^{d_n} \mu_{jk} B_k(U), \quad j = 1, \ldots, p, \]

\[ b_j(U) \approx \sum_{k=1}^{d_n} v_{jk} B_k(U), \quad j = 1, \ldots, p, \]

where \{\mu_{jk}\}_{k=1}^{d_n} and \{v_{jk}\}_{k=1}^{d_n} are scalar coefficients.

We now consider the following sample version of the marginal quantile regression problem:

\[
\min_{\mu_j, \nu_j \in \mathbb{R}^{d_n}} \frac{1}{n} \sum_{i=1}^{n} \rho_{\tau}(Y_i - B^T(U_i) \mu_j - B^T(U_i) \nu_j X_{ji}),
\]

where \(\mu_j = (\mu_{j1}, \ldots, \mu_{jd_n})^T, \nu_j = (v_{j1}, \ldots, v_{jd_n})^T\) and \(B(\cdot) = (B_1(\cdot), \ldots, B_{d_n}(\cdot))^T\).

To obtain the solution of Equation (6) we assumed that the regression observations \((Y, X)\) are in general position (defined by Roger Koenker [13]) and for some \(h \subset \{1, \ldots, n\}\), that \(|h| = 2d_n\) is the size of the set \(h\),

\[ -\tau I_{2d_n} \leq \xi(h) \leq (1 - \tau) I_{2d_n}, \]

where \(I_{2d_n}\) denotes \(2d_n\) dimensional column vector of all entries 1, \(\xi^T(h) = \sum_{i \in h} \psi_{\tau}(Y_i - X_i^T b(h) H_{nj}^{-1}(h)), \psi_{\tau}(u) = \tau - I(u < 0).\) The solution of Equation (6) is given by

\[
\hat{b}_j(h) = (\hat{\mu}_j^T, \hat{\nu}_j^T)^T = (H_{nj}(h))^{-1} Y(h),
\]

\[
H_{nj} = (B_n, \Phi_{nj}) = \begin{pmatrix}
B^T(U_1), & X_{j1} B^T(U_1) \\
\vdots & \vdots \\
B^T(U_n), & X_{jn} B^T(U_n)
\end{pmatrix}
\]

is an \(n \times 2d_n\) matrix, and \(H_{nj}(h)\) is an \(2d_n \times 2d_n\) reversible matrix from \(H_{nj}\) (Theorem 2.1 at Roger Koenker [13]).

So we derive the the estimate of the marginal effect \(f_j\) as

\[
\hat{f}_j = H_j \hat{b}_j(h) - F_{Y,n,U}^{-1}(\tau),
\]

where \(H_j = (B_1(U), \ldots, B_{d_n}(U), B_1(U)X_j, \ldots, B_{d_n}(U)X_j)\).

Next we apply the similar method to define the expression of \(\tilde{f}_j\). That is

\[
\tilde{f}_j = H_j [E \tilde{H}_j]^{-1} EY - Q(Y|U),
\]

where \(E \tilde{H}_j = \text{diag}(EB_1(U), \ldots, EB_{d_n}(U), EB_1(U)X_j, \ldots, EB_{d_n}(U)X_j)\).

3. Theoretical properties

3.1 Sure screening property

Here and below we give the following regularity conditions to establish the sure screening properties.
(i) The conditional quantile function $\beta_j(U)$ belongs to $F$, where $F$ is a class of functions defined on $[0, 1]$ whose $l$th derivative satisfies a Lipschitz condition of order $c: |f^d(s) - f^d(t)| \leq c_0 |s - t|$, for some positive constant $c_0$, $s, t \in [0, 1]$, where $l$ is a nonnegative integer and $c \in (0, 1)$ satisfies $l = d + c > 0.5$.

(ii) $\min_{j \in \mathcal{M}_t} E(Q_t(Y|X_j, U) - Q_t(Y|U))^2 \geq c_1 n^{-1}$, for some $0 \leq t \leq 2l/(2l + 1)$ and some positive constant $c_1$.

(iii) The conditional density $f_{Y|X_j, U}$ is bounded away from 0 and $\infty$ on $[Q_t(Y|X_j, U) - \eta, Q_t(Y|X_j, U) + \eta]$, for some $\eta > 0$, uniformly in $(X_j, U)$.

(iv) $d_{n}^{-2l} n^l = O(1), d_{n^2}^{-2l} = o(1)$.

(v) Suppose for $j = 1, \ldots, p$, $\text{Var}[X_j|U]$ is uniformly bounded away from 0 and infinity. That is, there exist some positive constants $h_1$ and $h_2$, such that $0 < h_1 \leq \text{Var}[X_j|U] \leq h_2 < \infty$.

(vi) The density function $g$ of $U$ is bounded away from zero and infinity on $U$. That is, $0 < T_1 \leq g(u) \leq T_2 < \infty$ for some constants $T_1$ and $T_2$.

(vii) Suppose for all $j = 1, \ldots, p$, there exists a positive constant $K_1$ and $r_1 \geq 2$, such that

$$P(|X_j| > t|U) \leq \exp(1 - (t/K_1)^{r_1}),$$

uniformly on $U$, for any $t \geq 0$. Furthermore, let $m(X^*) = E[Y|X, U]$, where $X^* = (X^T, U)^T$. Suppose there exists some positive constants $K_2$ and $r_2$ satisfying $r_1 r_2 / (r_1 + r_2) \geq 1$, such that

$$P(|m(X^*)| > t|U) \leq \exp(1 - (t/K_2)^{r_2})$$

uniformly on $U$, for any $t \geq 0$.

Conditions (i)–(iv) were used by He et al. [10] to prove variable screening property of additive model. Condition (v)–(vii) are assumed to restrict the moment of covariate and response on exposure variable and bear some similarity with that [6].

We assume that the covariates are standardized, that is:

$$EX_j = 0 \quad \text{and} \quad EX_j^2 = 1 \quad j = 1, \ldots, p.$$ 

**Proposition 1** If conditions (i)–(iv) hold, then $\min_{j \in \mathcal{M}_t} \|f_j\|^2 \geq c_6 n^{-1}$, for some positive constant $c_6$ and all $n$ sufficiently large.

The proof of Proposition 2 is given in the supplemental material [20].

**Theorem 1** Assume conditions (i)–(vii) hold.

(1) For any $h_1 > 0$, there exist positive constants $h_1, c_7, c_8, t_2, t_3$, such that

$$P \left( \max_{1 \leq j \leq p} \|f_j\|^2_n - \|f_j\|^2 \geq h_1 n^{-1} \right) < s_n \left[ 6s_n \exp(-c_7 n^{1-4d}) + 4s_n \exp(-c_8 d_n^{-2} n^{1-2d}) + 12d_n^2 \exp(-t_2 d_n^{-3} n) + 6d_n^2 \exp(-t_3 d_n^{-2} n^{-1}) \right]. \quad (10)$$

(2) If $\gamma_n = c_6 n^{-1}/2$, then

$$P(\mathcal{M}_* \subset \mathcal{M}_{\gamma_n}) \geq 1 - s_n \left[ 6s_n \exp(-c_7 n^{1-4d}) + 4s_n \exp(-c_8 d_n^{-2} n^{1-2d}) + 12d_n^2 \exp(-t_2 d_n^{-3} n) + 6d_n^2 \exp(-t_3 d_n^{-2} n^{-1}) \right].$$
The proof of Theorem 1 is given in the supplemental material [20].

Remark 1 According to Theorem 1, we can handle the number of covariates

\[ p = o(\exp(n^{1-\delta})). \]

3.2 False selection rates

From Proposition 1, the ideal case for vanishing false-rate is when

\[ \max_{j \notin M} \| \hat{f}_j \|_2^2 = o(n^{-\delta}), \]

so that important and unimportant variables are naturally separated. By Theorem 1(1), when Equation (10) tends to 0, we have with probability tending to 1 that

\[ \max_{j \notin M} \| \hat{f}_j \|_2^2 \leq cn^{-\delta}, \quad \text{for any } c > 0. \]

For this reason, by choosing \( \gamma_n \) as in Theorem 1(2), the model selection consistency can be established:

\[ P(M_{\tau n} = M_*) = 1 - o(1). \]

We now consider the more general case.

Theorem 2 Suppose the conditions in Theorem 1 hold, then for any \( \gamma_n = c_9 d_n n^{-\delta} \),

\[
P \left( |M_{\gamma_n}| \leq O \left( n' \lambda_{\max} \left( \sum \right) \right) \right) \geq 1 - s_n \left( 6s_n \exp(-c_5 n^{1-4\delta}) + 4s_n \exp(-c_6 d_n^{-2} n^{1-2\delta}) + 12d_n^2 \exp(-t_2 d_n^{-3} n) + 6d_n^2 \exp(-t_3 d_n^{-2} n^{-\delta}) \right),
\]

where \( \Sigma = EH_j^T H_j \).

The proof of Theorem 2 is given in the supplemental material [20].

4. The method of QNIS

Applying the method of QNIS to select the variables, inevitably, we would suffer from false negative (i.e. miss some important predictors that are marginally weakly correlated but jointly correlated with the response), and false positive (FP) (i.e. select some unimportant predictors which are highly correlated with the important ones). Hence, we use group-SCAD penalty as our selection strategy by an iterative framework by which to reduce the large-scale variable to a moderate-scale variable, enhancing the performance of our proposed method.

4.1 Description of the Algorithm

We build a data-driven thresholding \( \gamma_n \) for independence screening, by extending the random permutation idea of Zhao and Li [21], which allows only \( 1 - q \) proportion (for a given \( q \in [0, 1] \)) of inactive variables to enter the model when \( X \) and \( Y \) are not related (the null model). We
decouple $X_i$ and $Y_i$ by random permutation, so that the resulting data \{$(X_{\pi(i)}, Y_i)$\} follow a null model. Where $\{\pi_1, \ldots, \pi_n\}$ is a permutation of \{1, \ldots, n\}. The algorithm proceeds as follows:

1. For $j \in 1, \ldots, p$, compute
   \[
   \| \hat{j}_n \|^2 = \| H_j \hat{\beta}_j(h) - F_{Y, n}^{-1}(\tau) \|^2,
   \]
   where it is defined in Equations (3) and (4) using \{$(Y, U, X_j), j = 1, \ldots, p$\}. Select the top $K$ marginal utilities $\| \hat{j}_n \|^2$, resulting in the index subset $\mathcal{M}_0$.
2. For given quantile $\tau \in (0, 1)$, by quantile regression $Y$ on \{$U, X_j, j \in \mathcal{M}_0$\}, we get the estimators (relating to $\tau$), including intercept $\hat{\beta}_{00}$ and functional coefficients’ estimators $\{\hat{\beta}_{0j}, j \in \mathcal{M}_0\}$. Conditioning on $\mathcal{M}_0$, the $n$-dimensional partial residual is
   \[
   Y^* = Y(U) - \hat{\beta}_{00}(U) - \sum_{j \in \mathcal{M}_0} X_j \hat{\beta}_{0j}(U),
   \]
   where $Y(U)$ is the sample $\tau$-quantile of response on $U$. Next, compute $\| \hat{j}_n \|^2$ by data \{$(Y^*, U, X_j), j \in \mathcal{M}_0$\}, denoted $\| \hat{j}_n^* \|^2$, which reflects the additional utility of each covariate conditioning on the selected set $\mathcal{M}_0$. Random permutation on the partial residual $Y^*$, which yields $Y^{**}$. After a quantile regression for the decoupled data \{$(Y^{**}, U, X_j), j \in \mathcal{M}_0$\}, we get $\| \hat{j}_n^{**} \|^2$. Let $\gamma_{q}^*$ be the $q$-ranked magnitude of $\| \hat{j}_n^{**} \|^2, j \in \mathcal{M}_0$. The active variable set of covariates is chosen as
   \[
   \mathcal{A}_1 = \{ j : \| \hat{j}_n^* \|^2 > \gamma_{q}^*, j \in \mathcal{M}_0 \} \cup \mathcal{M}_0.
   \]
   In our numerical examples, we use $q = 1$.
3. We apply the Group-SCAD penalty on $\mathcal{A}_k$ to select a subset of variables $\mathcal{M}_k$, and $\mathcal{M}_k = \{ j \in \mathcal{A}_k : \gamma_{j,k} \neq 0 \}$ as follows:
   \[
   \min_{\mathbf{y} \in \mathbb{R}^p} 1/n \sum_{i=1}^n \rho_\tau(Y_i - \mathbf{B}(U_i)\mathbf{y}_0) - \sum_{j \in \mathcal{A}_i} \mathbf{B}(U_i)X_jy_j + \sum_{j \in \mathcal{A}_i} p_k(\| \mathbf{y}_j \|_B),
   \]
   where $\| \mathbf{y}_j \|_B = \sqrt{1/n \sum_{i=1}^n (\sum_{k=1}^d \mathbf{B}_k(U_i)\gamma_{jk})^2}$, and $p_k(\cdot)$ is the SCAD penalty such that
   \[
   p_k'(|x|) = \lambda I(|x| \leq \lambda) + \frac{(a\lambda - |x|)_+}{a - 1} I(|x| > \lambda),
   \]
   with $p_k(0) = 0$. From Fan and Li [4], we set $a = 3.7$. We select $\lambda$ by BIC (Bayesian information criterion) criteria $n\log(\hat{\sigma}_l^2 + ed_n \log n)$, where $e$ is the number of covariates chosen.
4. Repeat the step 2–3 where we replace $\mathcal{M}_{k-1}$ by $\mathcal{M}_k, k = 1, 2, \ldots$, and get $\mathcal{A}_{k+1}$ and $\mathcal{M}_{k+1}$. Iterate until $\mathcal{M}_{k+1} = \mathcal{M}_k$ or $|\mathcal{M}_{k+1}| \geq s_0$, for some prescribed integer $s_0$.

4.2 Greedy method
Following Wang [18], Fan and Song [7] and Fan et al. [6] we also implemented the greedy method of QNIS. We select the top $K$ variables that have the largest marginal norms $\| \hat{j}_n \|^2$, that is, $|\mathcal{M}_0| = K$. We apply the second step above to get $\mathcal{M}_k$, then back to the second step to get $\mathcal{A}_k$. We then iterate those steps as the third step above.
5. Numerical studies

In this section we present several simulation examples to assess the performance of our proposed methods (QNIS) with the conditional linear structure as in Equation (1). We also compare our method (QNIS) with the method (NIS) of Fan et al. [6]. We apply the common setup for the following simulation which are: cubic B-spline, \( L_n = 7 \), sample size \( n = 500 \), the number of variables \( p = 1000 \), and the number of simulation \( N = 100 \) for each example.

Example 1  The example is adopted from Fan et al. [6]. Let \( \{W_1, \ldots, W_p\} \) be i.i.d. standard normal and \( \{U_1, U_2\} \) be i.i.d. standard uniformly distributed random variables. We get \( \{Y, U, X\} \) as follows:

\[
X_j = \frac{W_j + t_1 U_1}{1 + t_1}, j = 1, \ldots, p, \quad U = \frac{U_2 + t_2 U_1}{1 + t_2},
\]

\[
Y = 2X_1 + 3U \cdot X_2 + (U + 1)^2 \cdot X_3 + \frac{4 \sin(2\pi U)}{2 - \sin(2\pi U)} \cdot X_4 + \exp(X_5^2 + X_6) \cdot \epsilon.
\]

where \( \epsilon \) has Cauchy distribution. Hence \( \text{corr}(X_j, X_k) = 0 \) for \( j \neq k \) when \( t_1 = t_2 = 0 \). As \( t_1 = 3 \) and \( t_2 = 1 \), \( \text{corr}(X_j, X_k) = 0.43 \) for \( j \neq k \) and \( \text{corr}(X_j, U) = 0.46 \).

We simulate the model with quantile is that \( \tau = 0.5 \) and \( \tau = 0.75 \). We use the conditional permutation method for different \( K \) (When \( K = 0 \), it is unconditional permutation). The screening threshold is fitted as \( \tau_q \) with \( q = 1 \). Report the average of the true positive (TP) number, model size, the lower bound of the marginal signal of true variables and upper bound of the marginal signal of false variables for different correlation setting.

From Table 1, we find that for heteroscedastic data, QNIS method perform better than NIS in selecting important variables for interested quantiles, which is only similar to the case of quantile 0.5. When \( K = 0 \), the correlation among covariates is large. It is hardly any differentiability between the marginal utilities of the true variables and the false ones. This makes the unconditional random permutation not a feasible method to determine the screening threshold in practice. When we make a few choices of tuning parameter \( K \), the screening threshold is taken as \( \tau_q \) with \( q = 1 \), the effect has been dramatic. When \( K = 4 \) or \( K = 8 \) all the true variables have already included in the first \( K \) variables (i.e. \( M_0 \setminus M_q = \emptyset \)), hence the minimum of true signal is not available, that is, ‘NA’. Based on Table 1, we find the conditional permutation method for different quantile of interest selecting thresholding maintains the sure screening properties and dramatically reduces the model sizes for heteroscedastic data.

Example 2  Let \( \{U, X\} \), and \( \epsilon \) be the same as in Example 1. We construct the model as follows:

\[
Y = 3U \cdot X_1 + (U + 1)^2 \cdot X_2 + (U - 2)^3 \cdot X_3 + 3(\sin(2\pi U)) \cdot X_4
\]

\[
+ \exp(U) \cdot X_5 + 2 \cdot X_6 + 2X_7 + 3\sqrt{U} \cdot X_8 + \exp(X_9^2 + X_{10} + X_{11} + X_{12}) \cdot \epsilon.
\]

In Example 2 we explore the performance of Conditional-NIS and Greedy-NIS method. We take \( K = 5 \) when using conditional-NIS method. We report the average number of TP, FP and prediction error (PE).

Table 2 reports the results of the simulated model specified in Example 2. For different quantiles, different number of covariates is selected, that is, the QNIS method is better than NIS for heterogeneous data. Correlation between covariates or between covariates and exposure variables can influence the performance of sure screening. When the covariates are independent or weakly correlated TP is the same as the number of covariates in the true model and FP is rare. As the correlation gets stronger TP decreases and FP increases. It seems that Greedy-NIS selects
Table 1. Model size and marginal signals for different correlation setting.

| Model \(K\) | \(\tau\) | TP \(\tau\) | Size | \(\min_{j \in M_+ \setminus M_0} \| \hat{f}_j \|_n^2\) | \(\min_{j \in M_+ c \setminus M_0} \| \hat{f}_j \|_n^2\) | \(\min_{j \in \{1, ..., p\} \setminus M_0} \| \hat{f}_j \|_n^2\) |
|---|---|---|---|---|---|---|
| \(0\) | 0.5 | 4 | 8.95 | 3.53 | 2.33 | 1.96 |
| QCIS | 0.75 | 7 | 9.32 | 5.21 | 1.33 | 1.15 |
| NIS | 0.5 | 4 | 369.41 | 0.43 | 0.32 | 0.23 |
| \(1\) | 0.75 | 7 | 661.58 | 0.85 | 0.61 | 0.44 |
| QCIS | 1.0 | 4 | 6.68 | 0.96 | 0.22 | 0.12 |
| NIS | 1.0 | 4 | 886.49 | 0.61 | 0.58 | 0.22 |
| \(4\) | 0.75 | 7 | 203.55 | 0.35 | 0.22 | 0.30 |
| QCIS | 1.0 | 4 | 286.47 | 0.41 | 0.36 | 0.35 |
| NIS | 1.0 | 4 | 202.50 | 0.16 | 0.11 | 0.11 |
| \(8\) | 0.75 | 7 | 5.70 | NA | 0.06 | 0.06 |
| QCIS | 1.0 | 4 | 8.92 | NA | 0.05 | 0.05 |
| NIS | 1.0 | 4 | 8.43 | 0.61 | 0.44 | 0.05 |

Table 2. Average values of TP, FP and PE for the model in Example 2.

| Model \(q\) | Correlation \(\chi^2 \chi^2 - \Upsilon\) | Conditional-NIS | Greedy-NIS |
|---|---|---|---|
| QCIS | \(q = 0.5\) | \(t_1 = 0, t_2 = 0\) | 0 | 0 | 8 | 0.20 | 1.21 | 8 | 8.55 | 1.32 |
| | | \(t_1 = 2, t_2 = 0\) | 0.25 | 0 | 8 | 0.13 | 1.12 | 8 | 0.54 | 1.15 |
| | | \(t_1 = 2, t_2 = 1\) | 0.25 | 0.36 | 7.85 | 0.08 | 2.10 | 7.65 | 0.21 | 2.11 |
| | | \(t_1 = 3, t_2 = 0\) | 0.43 | 0 | 7.90 | 0.11 | 1.14 | 7.90 | 0.64 | 1.12 |
| | | \(t_1 = 3, t_2 = 1\) | 0.43 | 0.46 | 7.80 | 0.12 | 1.51 | 7.80 | 0.68 | 1.52 |
| | \(q = 0.75\) | \(t_1 = 0, t_2 = 0\) | 0 | 0 | 12 | 0.18 | 1.22 | 12 | 9.23 | 1.25 |
| | | \(t_1 = 2, t_2 = 0\) | 0.25 | 0 | 12 | 0.15 | 1.10 | 12 | 0.56 | 1.13 |
| | | \(t_1 = 2, t_2 = 1\) | 0.25 | 0.36 | 11.75 | 0.15 | 1.99 | 11.55 | 0.23 | 2.03 |
| | | \(t_1 = 3, t_2 = 0\) | 0.43 | 0 | 11.85 | 0.11 | 1.18 | 11.90 | 0.68 | 1.16 |
| | | \(t_1 = 3, t_2 = 1\) | 0.43 | 0.46 | 11.70 | 0.12 | 1.60 | 11.45 | 0.24 | 1.65 |
| NIS | \(t_1 = 0, t_2 = 0\) | 0 | 0 | 8 | 0.21 | 1.24 | 8 | 10.71 | 1.57 |
| | | \(t_1 = 2, t_2 = 0\) | 0.25 | 0 | 8 | 0.13 | 1.17 | 8 | 0.60 | 1.16 |
| | | \(t_1 = 2, t_2 = 1\) | 0.25 | 0.36 | 7.80 | 0.20 | 2.16 | 7.55 | 0.26 | 2.26 |
| | | \(t_1 = 3, t_2 = 0\) | 0.43 | 0 | 7.90 | 0.10 | 1.21 | 7.98 | 0.71 | 1.29 |
| | | \(t_1 = 3, t_2 = 1\) | 0.43 | 0.46 | 7.75 | 0.18 | 1.65 | 7.35 | 0.28 | 1.84 |

slightly more FPs than Conditional-NIS, this result being from in each step Greedy-NIS selects the top variables by fitting the residuals conditional on previously chosen variable set and tends to overfit.
6. Real data analysis

6.1 Tecator Infratec Food Data

We illustrate the performance of our method through a real data analysis on the Meatspec data set which contains data from the Tecator Infratec Food Analyzer. It is available in the R add-on package *faraway* and on Statlib. The 100 predictors are channel spectrum measurements, and the response variable is ‘fat’. A total of $n = 215$ observations are available. We artificially increase covariant by i.i.d. standard normal random variables $\{Z_1, \ldots, Z_p\}$ and the standard uniform distribution $U$ so that

$$V_j = \frac{(Z_j + 2U)}{2}, \quad j = 101, \ldots, p.$$ 

Now we do regression on $\{U, V_1, \ldots, V_p\}$. We take $p = 1000$, respectively.

In Meatspec data set all variables have strongly relevance with response ‘fat’. To validate the effectiveness of our methods, we add heterogeneous as $(V_{101}^2 + V_{102})\varepsilon$, $\varepsilon \sim N(0, 1)$ to model, that is, the regression model is

$$\text{fat} = \beta_1(U)V_1 + \cdots + \beta_{100}(U)V_{100} + \beta_{101}(U)V_{101}$$

$$+ \cdots + \beta_{1000}(U)V_{1000} + (V_{101}^2 + V_{102})\varepsilon.$$

We repeat 100 times and report the average PE and model size. The simulation results are presented in Table 3.

From Table 3, our QNIS methods are effective with heterogeneous data. For different quantile the number of covariates is different, which cannot be achieved in NIS.

6.2 Boston Housing Data

To further demonstrate the performance of our proposed method in real practice, we present here a Boston Housing Data with $n = 506$ and 14 variables. Most empirical specification for the Boston Housing Data is that the dependent variable is $\text{medv}$ which is the median value of owner-occupied homes whose description can be found in the manual of R package mlbench. We assume the model is as following:

$$\log(\text{medv}) = \beta_0(U) + \beta_1(U)\ast\text{rm}^2 + \beta_2(U)\ast\text{age} + \beta_4(U)\ast\log(\text{rad})$$

$$+ \beta_5(U)\ast\text{tax} + \beta_6(U)\ast\text{ptratio} + \beta_7(U)(b - 0.63)^2 + \beta_8(U)\ast\log(\text{Istat})$$

$$+ \beta_9(U)\ast\text{crim} + \beta_{10}(U)\ast zn + \beta_{11}(U)\ast\text{indus}$$

$$+ \beta_{12}(U)\ast\text{chas} + \beta_{13}(U)\ast\text{nox}^2$$

$$+ \beta_{14}(U)\ast X_{14} + \cdots + \beta_{1000}(U)\ast X_{1000} + \varepsilon,$$

(11)

| Table 3. Model size and PE for Meatspec data. |
|-----------------------------------------------|
| **Method**          | **Quant** | **Size** | **PE** |
|---------------------|-----------|----------|--------|
| Conditional-QNIS    | 0.5       | 99.85    | 0.053  |
|                     | 0.75      | 101.80   | 0.068  |
| Conditional-NIS     |           | 99.76    | 0.060  |
| Greedy-QNIS         | 0.5       | 98.86    | 0.098  |
|                     | 0.75      | 101.27   | 0.052  |
| Greedy-NIS          |           | 98.96    | 0.071  |
where $U = \log(\text{dis})$, $X_j = (Z_j + 3W)/4$ ($j = s + 1, \ldots, p$), $Z_j$ is the standard normal random variables and $W$ follows the standard uniform distribution. Here $(U, X_1, \ldots, X_s) (s = 12)$ are the independent variables in Boston Housing Data and the variables $(X_j)_{j=s+1}^p$ are artificial predictors, are irrelevant to the housing price. We take quantiles as 0.5, 0.75, $p = 1000$ in which to randomly select $n = 406$ samples as training set, and compute prediction mean squared error (PE) on the rest 100 samples. We repeat 100 times and report the average PE and model size on the rest 100 samples. The simulation results are presented in Table 4.

From Table 4, we can see that our methods are very effective in selecting active variables and filtering noise variables in high dimensional setting, and acquire deferent model size for deferent quantile, respectively. In Conclusion, the proposed QNIS methodology is very useful which can select relatively true model and be applied as practical forecast under sparsity assumption.

### Disclosure statement

No potential conflict of interest was reported by the authors.

### Funding

The research was supported in part by National Natural Science Foundation of China [11171112, 11201190, 11101158], Doctoral Fund of Ministry of Education of China [20130076110004], Program of Shanghai Subject Chief Scientist [14XD1401600] and the 111 Project of China [B14019].

### Supplemental data and research materials

Supplemental material ‘Quantile-adaptive variable screening in ultra-high dimensional varying coefficient models’ can be accessed at doi:10.1080/02664763.2015.1072141. We provide additional technical details in the supplemental material.

### References

[1] Z. Cai and X. Xu, Nonparametric quantile estimations for dynamic smooth coefficient models, J. Amer. Statist. Assoc. 103 (2008), pp. 1595–1608.

[2] E. Candes and T. Tao, The Dantzig selector: Statistical estimation when $p$ is much larger than $n$ (with discussion), J. Ann. Statist. 35 (2007), pp. 2313–2404.

[3] J. Fan, Y. Feng, and R. Song, Nonparametric independence screening in sparse ultra-highdimensional additive models, J. Amer. Statist. Assoc. 106 (2011), pp. 544–557.

[4] J. Fan and R. Li, Variable selection via nonconcave penalized likelihood and its oracle properties, J. Amer. Statist. Assoc. 96 (2001), pp. 1348–1360.

[5] J. Fan and J. Lv, Sure independence screening for ultrahigh dimensional feature space (with discussion), J. R. Stat. Soc. B. 70 (2008), pp. 849–911.
[6] M. Fan, Y. Ma, and W. Dai, Nonparametric independence screening in sparse ultra-high dimensional varying coefficient models, 2013. Available at arxiv:1303.0458v1.

[7] J. Fan and R. Song, Sure independence screening in generalized linear models with NP-dimensionality, Ann. Statist. 38 (2010), pp. 3567–3604.

[8] I.E. Frank and J.H. Friedman, A statistical view of some chemometrics regression tools (with discussion), Technometrics 35 (1993), pp. 109–148.

[9] P. Hall, D.M. Titterington, and J.-H. Xue, Tilting methods for assessing the influence of components in a classifier, J. R. Stat. Soc. B. 71 (2009), pp. 783–803.

[10] X. He, L. Wang, and H.G. Hong, Quantile-adaptive model-free variable screening for high-dimensional heterogeneous data, Ann. Statist. 41 (2013), pp. 342–369.

[11] T. Honda, Quantile regression in varying coefficient models, J. Statist. Plann. Inference 121 (2004), pp. 113–125.

[12] M.-O. Kim, Quantile regression with varying coefficients, Ann. Statist. 35 (2007), pp. 92–108.

[13] R. Koenker, Quantile Regression, Econometric Society Monographs, Vol. 38, Cambridge University Press, Cambridge, 2005.

[14] G Li, H. Peng, J. Zhang, and L. Zhu, Robust rank correlation based screening, Ann. Statist. 40 (2012), pp. 1846–1877.

[15] R. Li, W. Zhong, and L. Zhu, Feature screening via distance correlation learning, J. Amer. Statist. Assoc. 107 (2012), pp. 1129–1139.

[16] R. Tibshirani, Regression shrinkage and selection via the lasso, J. R. Statist. Soc. B. 58 (1996), pp. 267–288.

[17] H. J. Wang, Z. Zhu, and J. Zhou, Quantile regression in partially linear varying coefficient models, Ann. Statist. 37 (2009), pp. 3841–3866.

[18] H. Wang, Forward regression for ultra-high dimensional variable screening, J. Amer. Statist. Assoc. 104 (2009), pp. 1512–1524.

[19] C.-H. Zhang, Nearly unbiased variable selection under minimax concave penalty, Ann. Statist. 38 (2010), pp. 894–942.

[20] L. Zhang, Quantile-adaptive variable screening in ultra-high dimensional varying coefficient models, 2015. doi:10.1080/02664763.2015.1014891.

[21] D.S. Zhao and Y. Li, Principled sure independence screening for Cox models with ultra-high-dimensional covariates, manuscript, Harvard University, 2010.

[22] H. Zou, The adaptive Lasso and its oracle properties, J. Amer. Statist. Assoc. 101 (2006), pp. 1418–1429.

[23] H. Zou and T. Hastie, Addendum: Regularization and variable selection via the elastic net, J. R. Statist. Soc. B. 67 (2005), pp. 301–320.

[24] H. Zou and R. Li, One-step sparse estimates in nonconcave penalized likelihood models, Ann. Statist. 36 (2008), pp. 1509–1533.