Interquantile shrinkage and variable selection in quantile regression

Liewen Jiang, Howard D. Bondell, Huixia Judy Wang

Department of Statistics, North Carolina State University, Raleigh, NC 27606, USA

ABSTRACT

Examination of multiple conditional quantile functions provides a comprehensive view of the relationship between the response and covariates. In situations where quantile slope coefficients share some common features, estimation efficiency and model interpretability can be improved by utilizing such commonality across quantiles. Furthermore, elimination of irrelevant predictors will also aid in estimation and interpretation. These motivations lead to the development of two penalization methods, which can identify the interquantile commonality and nonzero quantile coefficients simultaneously. The developed methods are based on a fused penalty that encourages sparsity of both quantile coefficients and interquantile slope differences. The oracle properties of the proposed penalization methods are established. Through numerical investigations, it is demonstrated that the proposed methods lead to simpler model structure and higher estimation efficiency than the traditional quantile regression estimation.

1. Introduction

Quantile regression (Koenker and Bassett, 1978) can provide more comprehensive statistical views than regression at the mean when studied at multiple quantiles. Conventional multiple-quantile regression methods often carry out analysis at each quantile level separately. However, if the quantile coefficients share some common features across quantile levels, modeling at multiple quantiles jointly to utilize such commonality can improve the estimation efficiency. In addition, prediction accuracy and model interpretability can be improved by eliminating irrelevant variables in multiple-quantile regression models.

For example, in regression models with independent and identically distributed (i.i.d.) errors, the quantile slope coefficients are constant across all quantile levels. Assuming this special model, Zou and Yuan (2008a) proposed a composite quantile regression method to estimate the common slopes, and they further adopted an adaptive Lasso method (Zou, 2006) to select nonzero slopes. In addition, Jiang et al. (2012) proposed a weighted composite quantile regression method and the sparse counterparts for nonlinear models by assuming that the parameters in the nonlinear functions are constant across quantiles. However, in practice, the i.i.d. error assumption may be violated, and the slope coefficients may appear constant only at a certain quantile region. One may use the hypothesis testing method such as the Wald test in Koenker (2005) to identify the commonality of quantile slopes, but this method becomes infeasible when the number of quantiles or the number of predictors are large. Jiang et al. (in press) used a penalized regression approach to smooth neighboring quantiles. In this paper, we propose new penalization methods to perform automatic estimation, and detection of zero coefficients and quantile regions with constant slope coefficients.
Penalization has become a very popular tool in variable selection. Tibshirani (1996) proposed the popular Least Absolute Shrinkage and Selection Operator (Lasso) method for variable selection. Efron et al. (2004) discussed least angle regression and its connection with Lasso and forward stagewise linear regression. Other adaptations of Lasso appeared in Tibshirani et al. (2005), Zou and Hastie (2005), Zou (2006), Yuan and Lin (2006), Meinshausen (2007), Huang et al. (2010), Meier et al. (2009), Ravikumar et al. (2009), and Aneiros-Pérez et al. (2011), to name a few. Alternative penalization-based variable selection approaches include the SCAD method developed by Fan and Li (2001), the Dantzig method proposed by Candès and Tao (2007), and the OSCAR method by Bondell and Reich (2008), among others. The readers are referred to Bühlmann and van de Geer (2011) for a more comprehensive coverage of different penalization methods.

The penalization idea was also adopted in quantile regression in different contexts. Li and Zhu (2008) studied $L_1$-norm quantile regression and computed the entire solution path. Wu and Liu (2009) discussed SCAD and adaptive Lasso in quantile regression and demonstrated their oracle properties. Li and Zhu (2007) and Wang and Hu (2011) analyzed comparative genomic hybridization data using fused quantile regression. Belloni and Chernozhukov (2011) studied quantile regression with Lasso penalty in high-dimensional models. Wang et al. (2013) studied variable selection in censored quantile regression. In all of these works, analysis is carried out at each given quantile level separately. Zou and Yuan (2008b) proposed a simultaneous multiple quantiles regression approach, where a group $F_\infty$-norm penalty was adopted to eliminate covariates that have no impacts on all quantile levels. Therefore, this method will retain a covariate in all the quantile regression models as long as it has some nonzero effect on at least one quantile level. However, the method is not able to identify common quantile slopes.

We develop two new penalization methods, Fused Adaptive Lasso (FAL) and Fused Adaptive Sup-norm (FAS). In the aforementioned references, penalizations are employed to select variables that have nonzero impacts on the mean or a given quantile of the response distribution. Our proposed methods differ from these existing works, as we target on variable selection and quantile smoothing for regression at multiple quantiles. By adopting fused penalization, we penalize the quantile slopes and their successive differences at neighboring quantile levels, so that the sparsity and the interquantile constancy of coefficients can be identified simultaneously. Therefore, the proposed methods can not only simplify the model structure but also improve the estimation efficiency by borrowing strength across quantiles to estimate the common slopes.

The rest of the paper is organized as follows. In Section 2 we present details of the proposed penalization methods including their asymptotic properties, and discuss some computational issues. The performance of the proposed methods are assessed through a simulation study in Section 3 and the analysis of an international economic growth data in Section 4. All technical details are provided in the Appendix.

2. Proposed method

2.1. Model setup

Let $Y$ be the response variable, $X \in \mathbb{R}^p$ be the $p$-dimensional covariate vector, and $(y_i, x_i), \ i = 1, \ldots, n$ be an observed sample. Suppose we are interested in $K$ quantile levels $0 < \tau_1 < \cdots < \tau_K < 1$, where $K$ is a finite integer. In this paper, we consider the linear quantile regression model

$$Q_{\tau_k}(x) = \alpha_k + x^T \beta_k,$$

where $\alpha_k \in \mathbb{R}$ is the intercept and $\beta_k \in \mathbb{R}^p$ is the slope vector at the quantile level $\tau_k$. Here $Q_{\tau_k}(x)$ denotes the $r$th conditional quantile of $Y$ given $X = x$, that is, $P(Y \leq Q_{\tau_k}(x) | X = x) = \tau$. The conventional quantile regression method estimates the parameter vector $(\alpha_k, \beta_k^T)$ at the given quantile level $\tau_k$ by minimizing the quantile objective function

$$\sum_{i=1}^{n} \rho_{\tau_k}(y_i - \alpha_k - x_i^T \beta_k),$$

(1)

where $\rho_{\tau}(r) = \tau r (r > 0) + (\tau - 1) r (r \leq 0)$ is the quantile check function and $I(\cdot)$ is the indicator function (Koenker and Bassett, 1978). If we consider the regression at multiple quantile levels, minimizing (1) is equivalent to minimizing the combined quantile loss function

$$\sum_{k=1}^{K} \sum_{i=1}^{n} \rho_{\tau_k}(y_i - \alpha_k - x_i^T \beta_k),$$

(2)

In some applications, however, it is likely that some covariates have no impacts on $Y$ at certain quantile levels. Including the irrelevant variables in the multiple-quantile regression model complicates the model and may decrease the predictability. It is also possible that some slope coefficient is constant in certain quantile regions but varies in others. The conventional quantile regression approach ignores such shared information across quantiles and thus may lose estimation efficiency.

In this paper, we adopt the fused penalization idea and propose two approaches by shrinking quantile slope coefficients and the interquantile slope differences towards zero simultaneously. A zero interquantile slope difference means that the
slope coefficients at the two neighboring quantile levels are tested. Therefore, the fused penalization leads to an automatic identification of sparse and constant slope coefficients.

We fix some notations before presenting the proposed procedures. Let $\beta_{k,l}$ be the slope coefficient corresponding to the $l$th predictor at the quantile level $\tau_k$, where $l = 1, \ldots, p$ and $k = 1, \ldots, K$. Let $\beta_{(l)} = (\beta_{1,l}, \ldots, \beta_{K,l})^T$ be the collection of slopes for the $K$ quantile levels for the $l$th predictor, and the parameter vector $\theta = (\alpha_1, \ldots, \alpha_K, \beta_{(1)}, \ldots, \beta_{(p)})^T$. Define $Z_{ik} = (1, X_i') \tau_k$, where $X_i = (D_{k,1}, D_{k,2}) \in \mathbb{R}^{(p+1) \times (p+1)K}$. $D_{k,1}$ is a $(p + 1) \times K$ matrix with 1 in the first row and the $k$th column but zero elsewhere, $D_{k,2} = (\theta_0, l)^T \otimes V_k$. $V_k$ is a $K \times 1$ vector with the $k$th element being 1 but zero elsewhere. Here $\theta_0$ is a $p \times 1$ zero vector, $l$ is a $p \times p$ identity matrix, and $\otimes$ denotes the Kronecker product. With these reparameterizations, the combined quantile objective function (2) can be written as

$$
\sum_{k=1}^{K} \sum_{l=1}^{n} \rho_{\tau_k}(y_i - Z_{ik}^T \theta).
$$

(3)

In order to detect the insignificant and the constant quantile slope coefficients, we propose to shrink the slope coefficients $\{\beta_{k,l} : k = 1, \ldots, K, l = 1, \ldots, p\}$ and the interquantile slope differences $\{\beta_{k,l} - \beta_{k-1,l} : k = 2, \ldots, K, l = 1, \ldots, p\}$ towards zero simultaneously, resulting in a simpler model structure and inducing the smoothness across quantiles.

2.2 Fused adaptive Lasso estimator

Our first method employs a weighted $L_1$ penalization on the quantile slope coefficients and interquantile slope differences. The fused adaptive Lasso estimator is defined as $\hat{\theta}_{\text{FL}} = \arg \min_\theta Q(\theta)$, where

$$
Q(\theta) = \sum_{k=1}^{K} \sum_{l=1}^{n} \rho_{\tau_k}(y_i - Z_{ik}^T \theta) + n\lambda_n \left( \sum_{l=1}^{p} \sum_{k=1}^{K} \tilde{w}_{ik}^l |\theta_{k+1} - \theta_k| + \sum_{l=1}^{p} \sum_{k=2}^{K} \tilde{v}_{ik} |\theta_{k+1} - \theta_{k-1} - \theta_k| \right).
$$

(4)

and $\lambda_n > 0$ is the tuning parameter controlling the degree of penalization. For each $l = 1, \ldots, p$, we set the adaptive weights $\tilde{w}_{ik,l} = |\theta_{k+1} - \theta_k|^{-1}$, $k = 1, \ldots, K$, and $\tilde{v}_{ik,l} = |\theta_{k+1} - \theta_{k-1} - \theta_k|^{-1}$, $k = 2, \ldots, K$, where $\beta$ (and hence $\theta$) is the initial estimator obtained from the conventional quantile regression without any shrinkage. The component-wise weight $\tilde{w}_{ik,l}$ controls the speed at which the slope for the $l$th predictor at quantile $\tau_k$ is shrunk to zero: the closer the initial estimator is to zero, the faster it will be shrunk to zero. Likewise, $\tilde{v}_{ik,l}$ controls the degree of smoothness between quantile coefficient processes: the closer the initial slopes are, the faster they will be shrunk towards each other. Since both $\tilde{w}_{ik,l}$ and $\tilde{v}_{ik,l}$ incorporate some prior information about quantile slope coefficients and their interquantile differences, they often lead to more appropriate shrinkage.

Before discussing the asymptotic properties of the fused adaptive Lasso estimator, we first study the properties of the oracle estimator, which is obtained as if the true model structure were known. For simple illustration, we consider $p = 1$ throughout the discussion of this section. Notations are thus simplified as $\theta = (\alpha^T, \beta^T)^T$ and hence $(\theta_{k+1}, \ldots, \theta_{2K}) = (\beta_1, \ldots, \beta_k)$. Then the adaptive weights $\tilde{w}_k = |\theta_{k+1} - \theta_k|^{-1}$ for $k = 1, \ldots, K$, while $\tilde{v}_k = |\theta_{k+1} - \theta_{k-1} - \theta_k|^{-1}$ for $k = 2, \ldots, K$. Define $\theta_0 = (\theta_{0,j} : j = 1, \ldots, 2K)^T$ as the true value of $\theta$. The set of indices $A_1 = \{j : \theta_{0,j} \neq 0, j = K + 1, \ldots, 2K\}$ contains the indices of the true nonzero quantile slope coefficients, while $A_2 = \{j : \theta_{0,j} \neq \theta_{0,j-1}, j = K + 2, \ldots, 2K\}$ includes the indices of quantile slope coefficients that differ from their preceding neighboring quantile level. Further we set $A_3 = A_2 \cup \{K + 1\}$ so that $A_3$ also includes the first slope index, which does not have a preceding neighbor. Let $A = \{1, \ldots, K\} \cup (A_1 \cap A_3)$, then the parameter vector $\theta_A = (\theta : j \in A)^T$ contains all quantile intercepts along with all nonzero unique quantile slope coefficients that need to be estimated under the oracle model. To better understand the notations, we examine the following two examples.

Example 1. Suppose we consider $K = 4$ quantile levels with $p = 1$, and the true quantile slope coefficients are $\beta_1 = 0, \beta_2 = \beta_3 = 2, \beta_4 = 3$, that is, the true parameter vector $\theta_0 = (\alpha_1, \ldots, \alpha_4, 0, 2, 2, 3)^T$. By the above definitions, $A_1 = \{K + 2, K + 3, K + 4\} = \{6, 7, 8\}$ and $A_2 = \{K + 1, K + 2, K + 4\} = \{5, 6, 8\}$. Hence, $A = \{1, \ldots, K\} \cup (A_2 \cup A_3) = \{1, 2, 3, 4, 6, 8\}$ and $\theta_A = (\alpha_1, \ldots, \alpha_4, \beta_2, \beta_4)^T$, which includes all quantile coefficients that appear in the oracle model, since $\beta_1 = 0$ and $\beta_3 = \beta_2$.

Example 2. We consider the same setting as in Example 1 but let $\beta_1 = \beta_2 = \beta_3 = 2, \beta_4 = 0$. In this case, $A_1 = \{K + 1, K + 2, K + 3\} = \{5, 6, 7\}$ and $A_2 = \{K + 1, K + 4\} = \{5, 8\}$. Hence, the index set $A = \{1, \ldots, K\} \cup (A_1 \cup A_2) = \{1, 2, 3, 4, 5\}$, and $\theta_A = (\alpha_1, \ldots, \alpha_4, \beta_1)^T$, where the common nonzero quantile slope coefficients $\beta_1, \beta_2$ and $\beta_3$ are collapsed to $\beta_1$ as a unique representative.

Under the assumption that the true model structure has been known beforehand, we can estimate $\theta_A$ by the oracle estimator

$$
\hat{\theta}_A = \arg \min_{\theta_A} \sum_{k=1}^{K} \sum_{l=1}^{n} \rho_{\tau_k}(y_i - Z_{ik,A}^T \theta_A),
$$

(5)
where \( z_{ik,A} = (z_{ik,j} : j \in A)^T \) is a subset of \( z_k \) whose indices of elements belong to the set \( A \). For notational convenience, we assume the first \( s < K \) quantile slope coefficients are nonzero and unique. Other more general cases follow the similar exposition, but with more complicated notations. To establish the asymptotic properties of the proposed estimators, the following regularity conditions are assumed throughout this paper.

(A1) For \( k = 1, \ldots, K, \ i = 1, \ldots, n \), the conditional density function of \( Y \) given \( X = x_i \), denoted as \( f_i \), is continuous and has a bounded first derivative, and \( f_i(Q_{\tau_k}(x_i)) \) is uniformly bounded away from zero and infinity.

(A2) \( \max_{1 \leq i \leq n} \| x_i \| = o(n^{1/2}) \).

(A3) For \( 1 \leq k \leq K \), there exist some positive definite matrices \( \Gamma_k \) and \( \Omega_k \) such that

\[
\lim_{n \to \infty} n^{-1} \sum_{i=1}^{n} z_{ik} z_{ik}^T = \Gamma_k \quad \text{and} \quad \lim_{n \to \infty} n^{-1} \sum_{i=1}^{n} f_i(Q_{\tau_k}(x_i)) z_{ik} z_{ik}^T = \Omega_k.
\]

The following proposition shows the asymptotic property of the oracle estimator \( \hat{\theta}_A \).

**Proposition 1.** Under the conditions (A1)–(A3), we have

\[
n^{1/2}(\hat{\theta}_A - \theta_{A,0}) \xrightarrow{d} N(0, \Sigma_A), \quad \text{as } n \to \infty,
\]

where \( \theta_{A,0} \) is the truth of \( \theta_A \), and \( \Sigma_A = \left( \sum_{k=1}^{K} \Omega_{k,A}^{-1} \right)^{-1} \left( \sum_{k=1}^{K} \tau_k (1 - \tau_k) \Gamma_{k,A} \right) \left( \sum_{k=1}^{K} \Omega_{k,A}^{-1} \right)^{-1} \). \( \Omega_{k,A} \) and \( \Gamma_{k,A} \) are the top-left \((K + s) \times (K + s)\) submatrices of \( \Omega_k \) and \( \Gamma_k \), which are defined in condition (A3).

However, in practice, the model structure is typically unknown. Therefore, we estimate the full parameter vector \( \theta \in \mathbb{R}^{K+p} \) by \( \hat{\theta}_{FAL} \). Theorem 1 presents the asymptotic properties of \( \hat{\theta}_{FAL} \) with \( p = 1 \).

**Theorem 1.** Suppose that conditions (A1)–(A3) hold. If \( n^{1/2} \lambda_n \to 0 \) and \( n \lambda_n \to \infty \), we have

(i) consistency in selection: \( \Pr[\{ j : \hat{\theta}_{j,FAL} \neq 0, j = K + 1, \ldots, 2K \} = A_1 \text{ and } \{ j : \hat{\theta}_{j,FAL} \neq \hat{\theta}_{j-1,FAL}, j = K + 2, \ldots, 2K \} = A_2 \} \to 1 \);

(ii) asymptotic normality: \( n^{1/2}(\hat{\theta}_{A,FAL} - \theta_{A,0}) \xrightarrow{d} N(0, \Sigma_A) \), where \( \Sigma_A \) is the covariance matrix of the oracle estimator given in Proposition 1.

### 2.3. Fused adaptive sup-norm estimator

Note that in fused adaptive Lasso estimation, the slope coefficients and the interquantile slope differences are penalized individually. In reality, however, if a predictor has no effect over all quantile levels, it may be desired to entirely remove the predictor from the multiple-quantile regression models. In the mean regression problem, Yuan and Lin (2006) considered selecting grouped variables. A typical example is the multi-factor analysis of variance (ANOVA), where each factor may have multiple levels and can be expressed as a group of dummy variables. Instead of selecting individual variables, Yuan and Lin (2006) proposed group Lasso to select important factors. In the multiple-quantile regression model, Zou and Yuan (2008b) estimated the quantile coefficients simultaneously, and imposed an \( L_\infty \)-norm on quantile slope coefficients to achieve group-wise sparsity. Hence the predictor is either in or excluded from all quantile regression models. We adopt the group-wise shrinkage idea in this section, and propose a fused adaptive sup-norm approach by imposing the \( L_\infty \)-penalty on both quantile slope coefficients and their interquantile slope differences. Such penalty, by treating quantile slope coefficients corresponding to one predictor as a group, will lead to the shrinkage of the entire group towards zero.

To illustrate the fused adaptive sup-norm, we consider a more general case with \( p \) predictors. For \( l = 1, \ldots, p \), denote \( d_{(l)} = (\beta_{1,l} - \beta_{1,1}, \ldots, \beta_{K,l} - \beta_{K,1})^T \) as a vector of interquantile slope differences corresponding to the \( l \)th predictor, which is essentially a linear transformation of \( \hat{\beta}_l \). The fused adaptive sup-norm estimator \( \hat{\theta}_{FAS} \) can be obtained by minimizing

\[
Q(\theta) = \sum_{k=1}^{K} \sum_{i=1}^{n} \rho_{\tau_k}(y_{i} - z_{ik}^T \theta) + n \lambda_n \left( \sum_{l=1}^{p} \sum_{i=1}^{n} \tilde{w}_i \| \beta_{(l),i} \|_\infty + \sum_{l=1}^{p} \tilde{v}_i \| d_{(l)} \|_\infty \right),
\]

where \( \tilde{w}_i = (\| \hat{\beta}_{(l),i} \|_\infty)^{-1} \) and \( \tilde{v}_i = (\| d_{(l)} \|_\infty)^{-1} \), \( l = 1, \ldots, p \), are the group-wise adaptive weights. The initial estimators \( \hat{\beta}_{(l)} \) and \( \hat{d}_{(l)} \) are calculated from the conventional quantile regression method. The tuning parameter \( \lambda_n > 0 \) controls the degree of group-wise penalization on the quantile coefficients and their interquantile differences.

We first discuss properties of the estimator from the oracle model. Define \( \hat{\beta}_{(l),0} \) and \( \hat{d}_{(l),0} \) as the truth of \( \beta_{(l),0} \) and \( d_{(l),0} \), respectively, and \( \| \cdot \| \) as the \( L_1 \)-norm. The index set \( B_1 = \{ l : \| \beta_{(l),0} \| \neq 0 \text{ and } \| d_{(l),0} \| \neq 0 \} \) corresponds to the predictors with at least one nonzero quantile coefficient, and the slope coefficients are not entirely constant across all quantiles. The index set \( B_2 = \{ l : \| \beta_{(l),0} \| \neq 0 \text{ and } \| d_{(l),0} \| = 0 \} \) corresponds to the predictors with nonzero but common slope coefficients across all quantile levels, and \( B_3 = \{ l : \| \beta_{(l),0} \| = 0 \} \) corresponds to the predictors with zero quantile slope coefficients.
across all quantile levels. Define $\mathcal{B} = \{(k, l) \colon k \in \{1, \ldots, K\}, l \in \mathcal{B}_1 \} \cup \{(k, l) \colon k = 1, l \in \mathcal{B}_2 \}$ and $\beta_\mathcal{B} = (\beta_{k,l} : (k, l) \in \mathcal{B})^T$, where $k$ corresponds to the $k$th quantile, and $l$ corresponds to the $l$th predictor. In other words, if the slope coefficients corresponding to one predictor are neither zero nor constant across all quantile levels, all quantile coefficients for this predictor will be kept in the model. If the slope coefficients corresponding to one predictor are a nonzero constant across all quantile levels, this predictor is also considered in the oracle model, but the quantile slope coefficients are collapsed to the unique one, for which case we select the slope coefficient at the first quantile level as a representative. On the other hand, if the predictor has zero coefficients across all quantile levels, this predictor is excluded from the oracle model. Denote $\theta_\mathcal{B} = (\alpha_1, \ldots, \alpha_K, \beta_\mathcal{B}^T)^T$ and $\theta_{\mathcal{B},0}$ as the truth. For better understanding, we examine the following example.

Example 3. We still consider $K = 4$, but with $p = 3$. Let $\beta_{(1)} = (0, 2, 2, 3)^T$, $\beta_{(2)} = (2, 2, 2, 2)^T$, and $\beta_{(3)} = (0, 0, 0, 0)^T$. By the definitions, $1 \in \mathcal{B}_1, 2 \in \mathcal{B}_2, 3 \in \mathcal{B}_3$ and $\beta_\mathcal{B} = (0, 2, 2, 3)^T$, which includes the group $\beta_{(1)}$, and collapses the common slope coefficients to the unique one in $\beta_{(2)}$. The oracle model of fused adaptive sup-norm either keeps, or excludes the predictor completely from all multiple-quantile regression models. Hence, unlike the oracle model of the fused adaptive Lasso, where no individual zero slope coefficient is included, some zero slope coefficients may still be in the oracle model of the fused adaptive sup-norm, such as the first quantile coefficient in $\beta_{(1)}$.

Without loss of generality, we reorder $\beta_{\mathcal{B}}$ in the vector $\theta_\mathcal{B}$ and assume the indices of predictors $\{1, \ldots, q_1\} \in \mathcal{B}_1$ and $\{q_1 + 1, \ldots, q_1 + q_2\} \in \mathcal{B}_2$, where $q_1 = |\mathcal{B}_1|$ and $q_2 = |\mathcal{B}_2|$. Here, $|\cdot|$ denotes the size of the set. If we assume the model structure has been known beforehand, the oracle estimator $\hat{\theta}_{\mathcal{B}}$ is defined as

$$\hat{\theta}_{\mathcal{B}} = \arg\min_{\theta_\mathcal{B}} \sum_{k=1}^{K} \sum_{i=1}^{n} \rho_{\tau_k}(y_i - z_{ik,\mathcal{B}}^T \theta_{\mathcal{B}}),$$

(7) where $z_{ik,\mathcal{B}}$ contains the first $K(q_1 + 1)$ elements, and the $[K(q_1 + 1) + 1]$th, $[K(q_1 + 2) + 1]$th, $\ldots$, $[K(q_1 + q_2) + 1]$th elements of $z_{ik}$.

Proposition 2. Under conditions (A1)–(A3), we have

$$n^{1/2}(\hat{\theta}_{\mathcal{B}} - \theta_{\mathcal{B},0}) \xrightarrow{d} N(0, \Sigma_{\mathcal{B}}), \quad \text{as } n \to \infty,$$

where $\Sigma_{\mathcal{B}} = \left(\sum_{k=1}^{K} \Omega_{k,\mathcal{B}} \right)^{-1} \left\{ \sum_{k=1}^{K} \tau_k (1 - \tau_k) \Gamma_{k,\mathcal{B}} \right\}^{-1}$, $\Omega_{k,\mathcal{B}}$ and $\Gamma_{k,\mathcal{B}}$ are the top-left $(K + Kq_1 + q_2) \times (K + Kq_1 + q_2)$ submatrices of $\Omega_k$ and $\Gamma_k$, respectively.

Since the model structure is unknown in practice, we consider the full parameter vector $\theta$ and estimate it by $\hat{\theta}_{\text{FAS}}$. For notational convenience, $\hat{\beta}_{(t),\text{FAS}}$ and $\hat{d}_{(t),\text{FAS}}$, the estimation of quantile slope coefficients and their interquantile slope differences for the $t$th predictor, are simplified as $\hat{\beta}_{(t)}$ and $\hat{d}_{(t)}$.

Theorem 2. Suppose that conditions (A1)–(A3) hold. If $n^{1/2} \lambda_n \to 0, n \lambda_n \to \infty$ as $n \to \infty$, we have

(i) consistency in selection: $\Pr\{[l : \|\hat{\beta}_{(t)}\| \neq 0, \|\hat{d}_{(t)}\| \neq 0] = \mathcal{B}_1 \} \cap [l : \|\hat{\beta}_{(t)}\| \neq 0, \|\hat{d}_{(t)}\| = 0] = \mathcal{B}_2 \} \cap [l : \|\hat{\beta}_{(t)}\| = 0] = \{l : \|\hat{\beta}_{(t)}\| = 0\} = \mathcal{B}_3 \} \to 1$;

(ii) asymptotic normality: $n^{1/2}(\hat{\theta}_{\text{B},\text{FAS}} - \theta_{\mathcal{B},0}) \to N(0, \Sigma_{\mathcal{B}})$, where $\Sigma_{\mathcal{B}}$ is the covariance matrix of the oracle estimator given in Proposition 2.

2.4. Computation

The aforementioned minimization problems in (4) and (6) are equivalent to linearly constrained minimization problems, which can be solved via linear programming. For example, minimizing (4) is equivalent to solving

$$\hat{\theta} = \arg\min_{\theta} \sum_{k=1}^{K} \sum_{l=1}^{K} \rho_{\tau_k}(y_l - z_l^T \theta), \quad \text{s.t.} \sum_{l=1}^{K} \sum_{k=1}^{K} \tilde{w}_{k,l} |\theta_{k+l} | + \sum_{l=1}^{K} \sum_{k=2}^{K} \tilde{w}_{k,l} |\theta_{k+l} - \theta_{k+l-1} | \leq t,$$

(8)

where $t > 0$ is a tuning parameter that plays a similar role as $\lambda$. Adopting this constrained minimization in (8) gives us a natural range of the tuning parameter, that is, $t \in [0, t_1]$, where $t_1 = Kp + (K - 1)p$. Similarly, (6) can be formulated as

$$\hat{\theta} = \arg\min_{\theta} \sum_{k=1}^{K} \sum_{l=1}^{K} \rho_{\tau_k}(y_l - z_l^T \theta), \quad \text{s.t.} \sum_{l=1}^{K} \tilde{w}_l \|\theta_{(l)}\|_\infty + \sum_{l=1}^{K} \tilde{u}_l \|\hat{d}_{(l)}\|_\infty \leq t,$$

(9)

where the tuning parameter $t \in [0, t_2]$ with $t_2 = 2p$. 
The performance of VAL, FAL, VAS, FAS and SCAD methods in Case 1, where $\beta_0(\tau) = 2 + 3\Phi^{-1}(\tau)$ varies with $\tau$, $\beta_2(\tau) = \beta_4(\tau) = 0$, and $\beta_1(\tau) = \beta_5(\tau) = 1$.

<table>
<thead>
<tr>
<th></th>
<th>AIC</th>
<th>BIC</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>TNb</td>
<td>TPb</td>
</tr>
<tr>
<td>VAL</td>
<td>0.198</td>
<td>0.594</td>
</tr>
<tr>
<td>FAL</td>
<td>0.264</td>
<td>0.744</td>
</tr>
<tr>
<td>VAS</td>
<td>0.634</td>
<td>0.984</td>
</tr>
<tr>
<td>FAS</td>
<td>0.198</td>
<td>1</td>
</tr>
<tr>
<td>SCAD</td>
<td>0.022</td>
<td>0.694</td>
</tr>
</tbody>
</table>

AIC/BIC: tuning parameter selection criterion; TNb/TPb: the proportion of times that all zero/nonzero quantile slope coefficients are identified as zero/nonzero; TNd: the proportion of times that interquantile slope differences are identified as zero for nonzero constant slope coefficients; GTNb/GTPb: the average number of groups with all zero/nonzero slope coefficients across quantiles being estimated as zero/nonzero over 500 simulations.

The solutions to (8) and (9) are related to the tuning parameter $t$. We consider both Akaike Information Criterion (AIC) (Akaike, 1974) and Bayesian Information Criterion (BIC) (Schwarz, 1978) to select $t$, defined as

$$AIC(t) = \text{loss}(t) + \frac{1}{n} \text{edf}(t),$$
$$BIC(t) = \text{loss}(t) + \frac{\log(n)}{2n} \text{edf}(t),$$

where $\text{loss}(t) = \sum_{k=1}^{K} \log \left[ \sum_{i=1}^{n} \rho_{\hat{\tau}}(y_i - \hat{z}_k \hat{\theta}(t)) \right]$ measures the goodness of fit; see Bondell et al. (2010) for a similar measure in multiple-quantile regression. The vector $\hat{\theta}(t)$ is the solution to (8) or (9) with the tuning parameter $t$. For the second term, BIC uses $\log(n)/2n$ as a multiplier, while AIC adopts $1/n$. If $n > e^2, \log(n)/(2n) > 1/n$ is always true, and BIC emphasizes more on simplifying the model structure. As a tradeoff, BIC may not be able to obtain the same estimation accuracy compared to AIC. The effective degree of freedom $\text{edf}(t)$ is also associated with the tuning parameter $t$. We set $\text{edf}$ as the number of nonzero unique quantile slope coefficients over predictors in both fused adaptive Lasso and fused adaptive sup-norm approaches.

### 3. Simulation study

In this section, we compare the following six approaches: FAL, FAS, Adaptive Lasso Variable Selection (VAL) proposed in Wu and Liu (2009), Adaptive Sup-norm Variable Selection (VAS) proposed in Zou and Yuan (2008b), Smoothly Clipped Absolute Deviation (SCAD) method that was originally proposed by Fan and Li (2001) and extended to the quantile regression setting by Wu and Liu (2009), and the conventional regression of quantiles (RQ) method. Simply speaking, VAL, VAS and SCAD are the approaches penalizing the quantile slope coefficients only, where the penalty terms in VAL and VAS are in the same format as the ones being imposed on the quantile slope coefficients in FAL and FAS, respectively.

We generate data from the following model

$$y_i = \alpha + \beta_1 x_{i1} + \cdots + \beta_6 x_{i6} + \sigma(x_i) e_i, \quad i = 1, \ldots, 200,$$

where $\alpha = 0$, $\beta_1 = \beta_2 = 1$, $\beta_3 = \beta_4 = \beta_5 = 0$, $\beta_6 = 2$, $e_i \sim N(0, 1)$ and all predictors $x_{i1}, \ldots, x_{i6}$ are generated from $U(0, 1)$ independently. We consider two different cases. In Case 1, we let $\sigma(x_i) = 3x_{i6}$, thus the true quantile slope coefficients are $\beta_3(\tau) = \beta_4(\tau) = \beta_5(\tau) = 0$ and $\beta_1(\tau) = \beta_2(\tau) = 1$ for all $\tau$. The only predictor that has a varying effect over $\tau$ is $x_{i6}$, where $\beta_6(\tau) = \beta_6 + \gamma \Phi^{-1}(\tau)$ varies with respect to $\tau$, and $\Phi^{-1}(\tau)$ is the $\tau$th quantile of $N(0, 1)$. In Case 2, we set $\sigma(x_i) = x_{i1} + x_{i2} + x_{i6}$, so this corresponds to an example with nonzero slope coefficients varying across all quantile levels. For both cases, we consider quantile levels 0.3, 0.4, 0.5, 0.6 and 0.7, and we repeat the simulation for 500 times.

We assess the performance of the six approaches via various criteria. Tables 1 and 2 summarize the results for model structure identification ability in Cases 1 and 2, respectively, when either AIC or BIC is used for tuning. Note that we do

### Table 1

<table>
<thead>
<tr>
<th></th>
<th>AIC</th>
<th>BIC</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>TNb</td>
<td>TPb</td>
</tr>
<tr>
<td>VAL</td>
<td>0.198</td>
<td>0.594</td>
</tr>
<tr>
<td>FAL</td>
<td>0.264</td>
<td>0.744</td>
</tr>
<tr>
<td>VAS</td>
<td>0.634</td>
<td>0.984</td>
</tr>
<tr>
<td>FAS</td>
<td>0.198</td>
<td>1</td>
</tr>
<tr>
<td>SCAD</td>
<td>0.022</td>
<td>0.694</td>
</tr>
</tbody>
</table>

AIC/BIC: tuning parameter selection criterion; TNb/TPb: the proportion of times that all zero/nonzero quantile slope coefficients are identified as zero/nonzero; TNd: the proportion of times that interquantile slope differences are identified as zero for nonzero constant slope coefficients; GTNb/GTPb: the average number of groups with all zero/nonzero slope coefficients across quantiles being estimated as zero/nonzero over 500 simulations.

### Table 2

The performance of VAL, FAL, VAS, FAS and SCAD methods in Case 2, where $\beta_1(\tau) = \beta_2(\tau) = 1 + \Phi^{-1}(\tau)$ and $\beta_6(\tau) = 2 + \Phi^{-1}(\tau)$ vary with $\tau$, and $\beta_3(\tau) = \beta_4(\tau) = \beta_5(\tau) = 0$.

<table>
<thead>
<tr>
<th></th>
<th>AIC</th>
<th>BIC</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>TNb</td>
<td>TPb</td>
</tr>
<tr>
<td>VAL</td>
<td>0.072</td>
<td>0.358</td>
</tr>
<tr>
<td>FAL</td>
<td>0.074</td>
<td>0.558</td>
</tr>
<tr>
<td>VAS</td>
<td>0.198</td>
<td>0.98</td>
</tr>
<tr>
<td>FAS</td>
<td>0.114</td>
<td>0.976</td>
</tr>
<tr>
<td>SCAD</td>
<td>0.01</td>
<td>0.486</td>
</tr>
</tbody>
</table>

AIC/BIC: tuning parameter selection criterion; TNb/TPb: the proportion of times that all zero/nonzero quantile slope coefficients are identified as zero/nonzero; TNd: the proportion of times that interquantile slope differences are identified as zero for nonzero constant slope coefficients; GTNb/GTPb: the average number of groups with all zero/nonzero slope coefficients across quantiles being estimated as zero/nonzero over 500 simulations.
Error (MISE), defined as the average of ISE over 500 simulations with constant effect overall quantile levels, FAS tends to detect the structure with more accuracy. (see Table 1) by imposing the group-wise shrinkage on interquantile slope differences additionally. Hence, if a predictor has we focus on identifying the constancy among quantiles slope coefficients, FAS returns at tremendously higher TNd than VAS carry out additional analysis to understand this phenomenon, and the discussion can be found in Remark 1. However, if nonzero values for TNd in Table 1 only occur accidentally in the case that neighboring quantiles are both shrunk to zero, hence equal, although they should be nonzero but equal. One surprising observation is that VAS returns higher TNb and GTNb than FAS regardless of the selection criteria. We carry out additional analysis to understand this phenomenon, and the discussion can be found in Remark 1. However, if we focus on identifying the constancy among quantile slope coefficients, FAS returns a tremendously higher TNd than VAS (see Table 1) by imposing the group-wise shrinkage on interquantile slope differences additionally. Hence, if a predictor has constant effect over all quantile levels, FAS tends to detect the structure with more accuracy.

The estimation efficiency of six different approaches is compared in Tables 3 and 4. We adopt Mean of Integrated Squared Error (MISE), defined as the average of ISE over 500 simulations with

\[
\text{ISE} = \frac{1}{n} \sum_{i=1}^{n} \left( (\alpha_k + x_i^T \beta_k) - (\hat{\alpha}_k + x_i^T \hat{\beta}_k) \right)^2.
\]

where \(\alpha_k + x_i^T \beta_k\) and \(\hat{\alpha}_k + x_i^T \hat{\beta}_k\) are the true and estimated \(r\)th conditional quantile of \(Y\) given \(x_i\).

Tables 3 and 4 show that although BIC leads to sparser models, it sacrifices some estimation efficiencies as a tradeoff with larger MISEs. Moreover, since FAL and FAS induce smoothness among quantiles, they yield smaller MISEs compared to VAL, VAS, SCAD and RQ methods in Case 1 where some quantile slope coefficients are constant across quantile levels. However, in Case 2 where all nonzero slopes vary across quantiles, FAL and FAS show no clear improvement of MISE over VAL, VAS and SCAD, even though they still perform generally better than the RQ method.

Remark 1. Results in Tables 1 and 2 show that FAS has lower TNb than VAS. More investigation suggests that for data with zero coefficients, FAS tends to shrink the coefficients to a nonzero constant across quantiles without further shrinkage, but VAS can shrink them down to exactly zero. To help understand this phenomenon, we look at a simpler example with \(p = 2\). The data is generated from

\[y_i = \alpha + \beta_1 x_{i,1} + \beta_2 x_{i,2} + 5x_{i,2} e_i, \quad i = 1, \ldots, 200,\]
where \( x_{i1} \sim U(0, 1), x_{i2} \sim U(0, 1), e_i \sim N(0, 1), \alpha = 1, \beta_1 = 0 \) and \( \beta_2 = 2 \). We consider three quantile levels \( \tau = \{0.4, 0.5, 0.6\} \) for simplicity. Fig. 1 shows the full solution path of the FAS and VAS methods for estimating \( \beta_1(\tau) \) at \( \tau = 0.4, 0.5 \) and 0.6. Due to the fused penalty employed for the FAS method, as \( t \) goes to zero, the estimates at different quantiles merge with each other first, and then the common slope is shrunk to zero together. However, for the VAS method, due to the sup norm, the 0.6th quantile slope is shrunk down to the 0.5th, then down to the 0.4th, and all the three solution lines go to zero afterwards. Comparing the solution paths for FAS and VAS, it is easy to see that it costs more for FAS to shrink the quantile slopes from a larger common constant to exactly zero than VAS, while this act leads to one less degree of freedom for both methods. Therefore, this additional cost prohibits FAS from further shrinking the common slope to be exactly zero.

### 4. Analysis of an economic growth data

In this section, we analyze an economic growth data by adopting the fusion approaches, FAL and FAS, and compare them with the non-fusion approaches, VAL and VAS. The data was originally taken from Barro and Lee (1994) and later studied by Koenker and Machado (1999).

In the data set, there are 161 observations. The first 71 observations correspond to 71 countries from the period 1965–1975, while the remaining 90 observations are for the period 1975–1985. Some countries may appear in both periods, and due to this fact, we may expect some correlation among the observations across time. To investigate this dependence, we examined the residuals after removing the predictor effects at the median. Upon examining the residuals corresponding to the same countries over the time periods, we found a sign-correlation of essentially zero (in fact, it was \(-0.026\)). It was shown that for quantile regression with locally dependent data, estimators obtained by assuming working independence are still consistent with minimal efficiency loss when compared to the most efficient estimator unless the dependencies are very strong; see Yin and Cai (2005), and Wang (2009). Therefore, we ignore the dependence in the following analysis. The response is the averaged annual growth percentages of per Capita Gross Domestic Product (GDP growth), and 13 covariates are involved in total: the initial per capital GDP (igdp), male middle school education (mse), female middle school education (fse), female higher education (fhe), male higher education (mhe), life expectancy (lexp), human capital (hcap), the ratio of eduction and GDP growth (edu), the ratio of investment and GDP growth (ivist), the ratio of public consumption and GDP growth (pcon), black market premium (blapk), political instability (pol) and growth rate terms trade (tttrad). All covariates are standardized to lie in the interval \([0, 1]\) before analysis, and we focus on \( \tau = \{0.1, 0.2, \ldots , 0.9\} \). Our purpose is to investigate the effects of covariates on multiple conditional quantiles of the GDP growth.

Koenker and Machado (1999) studied the effects of covariates on the conditional quantiles of the GDP growth by adopting the conventional quantile regression method (RQ). In this study, we consider penalization approaches to identify the model structure and estimate the multiple quantiles simultaneously. We select some representative predictors and list their estimated coefficients using both AIC and BIC as the selection criteria in Table 5. In general, BIC leads to more shrinkage than AIC for every predictor. For the predictor edu, since FAL is a component-wise shrinkage method, it can shrink individual quantile slope coefficients to zero, but FAS could not shrink any individual one to zero unless all of them are zero over quantiles (see the results for the predictor fhe). On the other hand, FAS is more likely to set slope coefficients to be constant over all quantile levels (see the results for the predictors edu and ttrad based on the BIC section).

To evaluate the prediction accuracy of different methods, we carry out a cross validation by randomly splitting data into a testing set with 50 observations and a training set with 111 observations. We adopt Prediction Error (PE) to assess the

### Table 4

MISE of VAL, FAL, VAS, FAS and SCAD methods in Case 2, where \( \beta_1(\tau) = \beta_2(\tau) = 1 + \phi^{-1}(\tau) \) and \( \beta_3(\tau) = 2 + \phi^{-1}(\tau) \) vary with \( \tau \), and \( \beta_4(\tau) = \beta_5(\tau) = \beta_6(\tau) = 0 \).

<table>
<thead>
<tr>
<th>Method</th>
<th>( \tau = 0.3 )</th>
<th>( \tau = 0.4 )</th>
<th>( \tau = 0.5 )</th>
<th>( \tau = 0.6 )</th>
<th>( \tau = 0.7 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>VAL</td>
<td>9.4 (0.3)</td>
<td>9.3 (0.3)</td>
<td>9.7 (0.3)</td>
<td>9.5 (0.3)</td>
<td>10.1 (0.3)</td>
</tr>
<tr>
<td>FAL</td>
<td>9.4 (0.3)</td>
<td>9.0 (0.3)</td>
<td>9.1 (0.3)</td>
<td>9.0 (0.3)</td>
<td>10.2 (0.3)</td>
</tr>
<tr>
<td>VAS</td>
<td>8.1 (0.3)</td>
<td>7.9 (0.3)</td>
<td>8.1 (0.3)</td>
<td>8.5 (0.2)</td>
<td>11.5 (0.3)</td>
</tr>
<tr>
<td>FAS</td>
<td>10.5 (0.3)</td>
<td>8.1 (0.3)</td>
<td>7.7 (0.2)</td>
<td>9.1 (0.2)</td>
<td>13.0 (0.4)</td>
</tr>
<tr>
<td>SCAD</td>
<td>8.5 (0.3)</td>
<td>8.5 (0.3)</td>
<td>8.7 (0.3)</td>
<td>8.5 (0.3)</td>
<td>8.4 (0.3)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Method</th>
<th>( \tau = 0.3 )</th>
<th>( \tau = 0.4 )</th>
<th>( \tau = 0.5 )</th>
<th>( \tau = 0.6 )</th>
<th>( \tau = 0.7 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>VAL</td>
<td>8.7 (0.3)</td>
<td>9.7 (0.3)</td>
<td>10.2 (0.3)</td>
<td>10.2 (0.3)</td>
<td>10.7 (0.4)</td>
</tr>
<tr>
<td>FAL</td>
<td>9.1 (0.3)</td>
<td>8.6 (0.3)</td>
<td>8.8 (0.3)</td>
<td>9.5 (0.3)</td>
<td>12.7 (0.4)</td>
</tr>
<tr>
<td>VAS</td>
<td>6.8 (0.2)</td>
<td>6.9 (0.2)</td>
<td>8.1 (0.3)</td>
<td>11.2 (0.4)</td>
<td>17.5 (0.5)</td>
</tr>
<tr>
<td>FAS</td>
<td>11.4 (0.4)</td>
<td>7.5 (0.3)</td>
<td>7.6 (0.2)</td>
<td>11.1 (0.3)</td>
<td>18.7 (0.4)</td>
</tr>
<tr>
<td>SCAD</td>
<td>9.8 (0.3)</td>
<td>10.3 (0.4)</td>
<td>9.8 (0.3)</td>
<td>9.6 (0.3)</td>
<td>10.1 (0.3)</td>
</tr>
</tbody>
</table>

MISE: Mean of Integrated Squared Errors among 500 simulations. SE: standard error of the MISE. The top half of the table contains the results from using AIC in selecting the tuning parameter, while the bottom half is for BIC.
Thesolutionpathof $\beta_1(\tau)$ for VAS and FAS approaches at $\tau = 0.4$, $\tau = 0.5$ and $\tau = 0.6$. The dotted vertical line corresponds to the optimal $t$ that gives the smallest BIC value.

**Table 5**
The estimated quantile coefficients for predictors fhe, edu, ivst and ttrad in the economic growth data.

<table>
<thead>
<tr>
<th></th>
<th>$\tau = 0.1$</th>
<th>$\tau = 0.2$</th>
<th>$\tau = 0.3$</th>
<th>$\tau = 0.4$</th>
<th>$\tau = 0.5$</th>
<th>$\tau = 0.6$</th>
<th>$\tau = 0.7$</th>
<th>$\tau = 0.8$</th>
<th>$\tau = 0.9$</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>FAL, AIC</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>fhe</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>edu</td>
<td>-0.39</td>
<td>-0.39</td>
<td>-0.03</td>
<td>-0.03</td>
<td>-0.03</td>
<td>-0.03</td>
<td>-0.03</td>
<td>-0.03</td>
<td>-0.03</td>
</tr>
<tr>
<td>ivst</td>
<td>0.67</td>
<td>0.67</td>
<td>0.67</td>
<td>0.67</td>
<td>0.67</td>
<td>0.67</td>
<td>0.67</td>
<td>0.67</td>
<td>0.60</td>
</tr>
<tr>
<td>ttrad</td>
<td>0.42</td>
<td>0.42</td>
<td>0.42</td>
<td>0.42</td>
<td>0.42</td>
<td>0.69</td>
<td>0.80</td>
<td>0.80</td>
<td>0.80</td>
</tr>
<tr>
<td><strong>FAS, AIC</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>fhe</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>edu</td>
<td>-0.29</td>
<td>-0.29</td>
<td>-0.21</td>
<td>-0.14</td>
<td>-0.06</td>
<td>-0.12</td>
<td>-0.13</td>
<td>-0.06</td>
<td>-0.13</td>
</tr>
<tr>
<td>ivst</td>
<td>0.71</td>
<td>0.71</td>
<td>0.71</td>
<td>0.70</td>
<td>0.69</td>
<td>0.68</td>
<td>0.67</td>
<td>0.67</td>
<td>0.67</td>
</tr>
<tr>
<td>ttrad</td>
<td>0.51</td>
<td>0.46</td>
<td>0.52</td>
<td>0.58</td>
<td>0.64</td>
<td>0.70</td>
<td>0.75</td>
<td>0.80</td>
<td>0.80</td>
</tr>
<tr>
<td><strong>FAL, BIC</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>fhe</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>edu</td>
<td>-0.35</td>
<td>-0.28</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>ivst</td>
<td>0.62</td>
<td>0.62</td>
<td>0.62</td>
<td>0.62</td>
<td>0.62</td>
<td>0.62</td>
<td>0.62</td>
<td>0.62</td>
<td>0.61</td>
</tr>
<tr>
<td>ttrad</td>
<td>0.39</td>
<td>0.39</td>
<td>0.39</td>
<td>0.39</td>
<td>0.70</td>
<td>0.70</td>
<td>0.70</td>
<td>0.70</td>
<td>0.70</td>
</tr>
<tr>
<td><strong>FAS, BIC</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>fhe</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>edu</td>
<td>-0.12</td>
<td>-0.12</td>
<td>-0.12</td>
<td>-0.12</td>
<td>-0.12</td>
<td>-0.12</td>
<td>-0.12</td>
<td>-0.12</td>
<td>-0.12</td>
</tr>
<tr>
<td>ivst</td>
<td>0.63</td>
<td>0.63</td>
<td>0.63</td>
<td>0.63</td>
<td>0.63</td>
<td>0.63</td>
<td>0.63</td>
<td>0.63</td>
<td>0.63</td>
</tr>
<tr>
<td>ttrad</td>
<td>0.60</td>
<td>0.60</td>
<td>0.60</td>
<td>0.60</td>
<td>0.60</td>
<td>0.60</td>
<td>0.60</td>
<td>0.60</td>
<td>0.60</td>
</tr>
</tbody>
</table>

Prediction accuracy, defined as

$$PE = \sum_{k=1}^{9} \sum_{j=1}^{50} \rho_{\tau_k} (y_j - x_j^T \hat{\beta}(\tau_k)),$$

where $\{(y_j, x_j), j = 1, \ldots, 50\}$ are in the test set, and $\hat{\beta}(\tau_k)$ is the estimated coefficient at $\tau_k$ based on the training set. We repeat the cross validation 200 times and take the average of PE. Results in Table 6 show that the fusion methods, FAL and FAS, have better prediction accuracy than the non-fusion methods, VAL and VAS, respectively, and all of these approaches...
outperform the conventional quantile regression method regardless of the selection criteria. As for the comparison between the two tuning parameter selection methods, BIC leads to more shrinkage but slightly lower prediction accuracy as a tradeoff. Overall, we see that the additional penalty to smooth across the quantiles gives better performance than only performing variable selection without the penalization across quantiles.

5. Conclusion and discussion

Examination of multiple conditional quantile functions is very useful in exploring a comprehensive relationship between the response and covariates. In this article, we propose two fused penalization methods in a multiple-quantile regression setting, which can identify the interquantile commonality and nonzero quantile coefficients simultaneously. As a consequence, the estimation efficiency and model interpretability will be enhanced, especially if there indeed exist common slope coefficients and irrelevant predictors.

In this paper, we work in the linear quantile regression setting. With the advancement of more complex data, extensions to nonparametric and functional quantile regression deserve further investigation. In this setting, interest would focus on the possibility of parallel quantile functions and selection of both complete functions, as well as parts of functions.

Acknowledgments

HDB’s research was supported in part by NSF grant DMS-1005612 and NIH grant P01-CA-142538. HJW’s research was supported by NSF grant DMS-1007420 and NSF CAREER Award DMS-1149355.

Appendix

We omit the proofs of Propositions 1 and 2, as they are similar to those in Jiang et al. (in press).

Lemma 1 (Convexity Lemma). Let \( \{h_n(u) : u \in U\} \) be a sequence of random convex functions defined on a convex, open subset \( U \) of \( \mathbb{R}^d \). Suppose \( h(u) \) is a real-valued function on \( U \) for which \( h_n(u) \to h(u) \) in probability for each \( u \in U \). Then for each compact subset \( K \) of \( U \), \( \sup_{u \in K} |h_n(u) - h(u)| \to 0 \) in probability.

Proof. The proof can be found in Pollard (1991).

Lemma 2 (Root-n Consistency of \( \hat{\theta}_{FAL} \)). Under conditions (A1)–(A3), if \( n^{1/2} \lambda_n \to 0 \) as \( n \to \infty \), then \( \hat{\theta}_{FAL} - \theta_0 = O_p(n^{-1/2}) \).

Proof. The proof of Lemma 2 is similar to that in Jiang et al. (in press) and thus is skipped.

Proof of Theorem 1. We first prove the consistency in selection. By reordering \( \theta \), we can decompose it as \( \theta = (\theta^T_A, \theta^T_{A_1}, \theta^T_{A_3 \setminus A_1})^T \), where \( A, A_1 \) are defined in Section 2.2.

Case 1. Suppose \( \theta^T_{A_1} \) is not selected correctly, that is, some elements in \( \hat{\theta}^T_{A_1} \) are not estimated as zero. Without loss of generality, we assume there is only one element \( \hat{\theta}_j \neq 0 \), where \( j \in A_1 \). Cases with more than one element in \( \theta^T_{A_1} \) not selected correctly basically follow the same elaborations, but with more complicated notations. Let \( \theta^* \) be a vector constructed by replacing \( \hat{\theta}_j \) with 0, other elements are the same as \( \hat{\theta} \). We can show that \( Q(\theta^*) < Q(\hat{\theta}) \), which contradicts the fact that \( \hat{\theta} \) minimizes the objective function \( Q(\theta) \) defined in (4). Define

\[
L_n(u) = \sum_{k=1}^{K} \sum_{i=1}^{n} \left\{ \rho_{\tau_k}(y_i - z_{ik}^T \theta_0 - n^{-1/2} z_{ik}^T u) - \rho_{\tau_k}(y_i - z_{ik}^T \theta_0) \right\}.
\]

As was shown in Jiang et al. (in press), \( L_n(u) \) is a bounded quantity provided that \( \|u\| \) is bounded.
Note that
\[
Q(\theta^*) - Q(\hat{\theta}) = L_n \{ n^{1/2}(\theta^* - \theta_0) \} - L_n \{ n^{1/2}(\hat{\theta} - \theta_0) \} - n\lambda_\alpha \bar{w}_j |\hat{\theta_j} - \theta_j| - n\lambda_\alpha \bar{v}_j |\hat{\theta_j} - \theta_j|
\]
\[+ n\lambda_\alpha \bar{v}_{j+1} \left( |\hat{\theta}_{j+1} - \theta_{j+1}| - |\hat{\theta}_j - \theta_j| \right). \tag{11} \]

Case 1a. If \( \theta_j - \theta_{j-1} \neq 0 \) and \( \theta_{j+1} - \theta_j \neq 0 \), then \( \bar{v}_j \rightarrow p |\theta_j - \theta_{j-1}|^{-1} \) and \( \bar{v}_{j+1} \rightarrow p |\theta_{j+1} - \theta_j|^{-1} \). Since \( |\hat{\theta}_{j-1} - \theta_{j-1}| \leq |\hat{\theta}_j - \theta_j| \) and \( |\hat{\theta}_{j+1} - \theta_{j+1}| \leq |\hat{\theta}_j - \theta_j| \), we have
\[
Q(\theta^*) - Q(\hat{\theta}) \leq L_n \{ n^{1/2}(\theta^* - \theta_0) \} - L_n \{ n^{1/2}(\hat{\theta} - \theta_0) \} - n\lambda_\alpha \bar{w}_j |\hat{\theta_j} - \theta_j| - n\lambda_\alpha \bar{v}_j |\hat{\theta_j} - \theta_j| - n\lambda_\alpha \bar{v}_{j+1} |\hat{\theta}_{j+1} - \theta_j|.
\]

Case 1b. If one of \( \theta_j - \theta_{j-1} \neq 0 \) and \( \theta_{j+1} - \theta_j = 0 \), but the other one is nonzero. Without loss of generality, suppose \( \theta_j - \theta_{j-1} \neq 0 \), \( \theta_{j+1} - \theta_j = 0 \). Under the assumption that only \( \hat{\theta}_j \neq 0 \) is not correctly selected, but other elements in \( \theta^* \) are correctly selected, we have \( \hat{\theta}_{j-1} = \hat{\theta}_{j+1} = 0 \), given the truth \( \theta_{j+1} - \theta_j = 0 \). Hence
\[
Q(\theta^*) - Q(\hat{\theta}) = L_n \{ n^{1/2}(\theta^* - \theta_0) \} - L_n \{ n^{1/2}(\hat{\theta} - \theta_0) \} - n\lambda_\alpha \bar{w}_j |\hat{\theta_j} - \theta_j| - n\lambda_\alpha \bar{v}_j |\hat{\theta_j} - \theta_j| - n\lambda_\alpha \bar{v}_{j+1} |\hat{\theta}_{j+1} - \theta_j|.
\]

Case 2. Now suppose there exists one \( j' \in A^C \setminus A_1^C \), where the true \( d_{j',0} = 0 \), but \( \hat{d}_{j'} \neq 0 \). Since \( j' \notin A_1^C \), we have \( \theta_{j',0} \neq 0 \), and \( \theta_{j'-1} \neq 0 \). In fact, the case of \( \theta_{j'-1} = \theta_{j',0} = 0 \) has been discussed in Case 1. Let \( \theta^* \) be a vector constructed by restricting \( \dot{d}_{j'} = 0 \), that is, every element in \( \theta^* \) is the same as the one in \( \hat{\theta} \), except \( \theta^*_{j'} \neq \hat{\theta}_{j'} \), which is to say, \( \dot{d}_{j'} \neq \hat{d}_{j'} \) and \( \dot{d}_{j'+1} \neq \hat{d}_{j'+1} \). Without loss of generality, we assume \( d_{j'+1} = 0 \). We can show that \( Q(\theta^*) < Q(\hat{\theta}) \). Note that
\[
Q(\theta^*) - Q(\hat{\theta}) = L_n \{ n^{1/2}(\theta^* - \theta_0) \} - L_n \{ n^{1/2}(\hat{\theta} - \theta_0) \} + n\lambda_\alpha \bar{w}_j (|\theta^*_{j'} - \theta_{j'}| - |\hat{\theta}_{j'} - \theta_{j'}|) + n\lambda_\alpha \bar{v}_j (|\hat{\theta}_{j'} - \theta_{j'}| - |\hat{\theta}_{j'} - \theta_{j'}|) + n\lambda_\alpha \bar{v}_{j+1} (|\hat{\theta}_{j+1} - \theta_{j+1}| - |\hat{\theta}_j - \theta_j|). \tag{13} \]

Since \( \theta^*_{j'} = \theta^*_{j'+1} = \theta_{j'-1} - \hat{\theta}_{j'} - n\lambda_\alpha \bar{w}_j (|\theta^*_{j'} - \theta_{j'}| - |\hat{\theta}_{j'} - \theta_{j'}|) \leq n\lambda_\alpha \bar{w}_j (|\hat{\theta}_{j'} - \theta_{j'}| - |\hat{\theta}_{j'} - \theta_{j'}| - |\theta_{j'-1} - \hat{\theta}_{j'}|) \rightarrow p 0 \). Moreover, \( \theta^*_{j'} = \hat{\theta}_{j'} + \hat{\theta}_{j'-1} \), and \( |\hat{\theta}_{j'} - \theta_{j'}| - |\theta_{j'-1} - \hat{\theta}_{j'}| \leq |\hat{\theta}_{j'} - \theta_{j'}| \). Thus, \( n\lambda_\alpha \bar{v}_{j+1} (|\hat{\theta}_{j'} - \theta_{j'}| - |\hat{\theta}_{j'} - \theta_{j'}|) \leq n\lambda_\alpha \bar{v}_{j+1} (|\hat{\theta}_{j'} - \theta_{j'}| - |\hat{\theta}_{j'} - \theta_{j'}| - |\theta_{j'-1} - \hat{\theta}_{j'}|) \rightarrow p 0 \). Therefore, \( n\lambda_\alpha \bar{v}_j (|\hat{\theta}_{j'} - \theta_{j'}| - |\hat{\theta}_{j'} - \theta_{j'}|) \) dominates the right hand side of (13), and \( Q(\theta^*) < Q(\hat{\theta}) \) holds, which contradicts the fact that \( \hat{\theta} \) is the minimizer to \( Q(\theta) \).
where the last term $-n\lambda_\nu \|\hat{d}_{\nu})\|_\infty$ dominates. Hence $Q(\theta^*) < Q(\hat{\theta})$. The asymptotic normality can be shown in a similar way as in the proof of Theorem 2 in Jiang et al. (in press).

References


Bondell, H., Reich, B., 2008. Simultaneous regression shrinkage, variable selection and clustering of predictors with OSCAR. Biometrics 64, 115–123.


