Interquantile Shrinkage and Variable Selection in Quantile Regression

Liewen Jiang[a], Howard D. Bondell[a], Huixia Judy Wang[a,*]

[a]*Department of Statistics, North Carolina State University, Raleigh, NC 27606, U.S.A*

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.

Keywords: Fused adaptive lasso; Fused adaptive sup-norm; Oracle; Quantile regression; Smoothing; Variable selection

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.

*Corresponding author. Email: hwang3@ncsu.edu, Phone: (919) 513-1661

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, Jiang and Song (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 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, Wang and Bondell (2013) 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 model selection and 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), Aneiros-Perez et al. (2011), to name a few. Alternative penalization-based variable selection approaches include the SACD method developed by Fan and Li (2001), the Dantzig method proposed by Candes and Tao (2007), the OSCAR method by Bondell and Reich (2008), and so on. The readers are referred to Bülhmann 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 (2009a) 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, Zhou and Li (2012) 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_∞-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 work, 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 inter-quantile 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, $\mathbf{X} \in \boldsymbol{R}^p$ be the p-dimensional covariate vector, and $\{(y_i, \mathbf{x}_i), i = 1, \cdots, n\}$ be an observed sample. Suppose we are interested in K quantile levels $0 < \tau_1 < \ldots < \tau_K < 1$, where K is a finite integer. In this paper, we consider the linear quantile regression model

$$Q_{\tau_k}(\mathbf{x}) = \alpha_k + \mathbf{x}^T \boldsymbol{\beta}_k,$$

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

$$\sum_{i=1}^{n} \rho_{\tau_k}(y_i - \alpha_k - \mathbf{x}_i^T \boldsymbol{\beta}_k), \tag{1}$$

where $\rho_\tau(r) = \tau r I(r > 0) + (\tau - 1) r I(r \leq 0)$ 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

$$\sum_{k=1}^{K} \sum_{i=1}^{n} \rho_{\tau_k}(y_i - \alpha_k - \mathbf{x}_i^T \boldsymbol{\beta_k}). \tag{2}$$

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 coeficients at the two neighboring

quantile levels are constant. 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 τ_k, where $l = 1, \ldots, p$ and $k = 1, \ldots, K$. Let $\boldsymbol{\beta}_{(l)} = (\beta_{1,l}, \ldots, \beta_{K,l})^T$ be the collection of slopes for the K quantile levels for the l^{th} predictor, and the parameter vector $\boldsymbol{\theta} = (\alpha_1, \ldots, \alpha_K, \boldsymbol{\beta}_{(1)}^T, \ldots, \boldsymbol{\beta}_{(p)}^T)^T$. Define $\mathbf{z}_{ik}^T = (1, \mathbf{x}_i^T)\boldsymbol{T}_k$, where $\boldsymbol{T}_k = (\boldsymbol{D}_{k,1}, \boldsymbol{D}_{k,2}) \in \boldsymbol{R}^{(p+1)\times(p+1)K}$, $\boldsymbol{D}_{k,1}$ is a $(p+1) \times K$ matrix with 1 in the first row and the k^{th} column but zero elsewhere, $\boldsymbol{D}_{k,2} = (\mathbf{0}_p, \boldsymbol{I}_p)^T \otimes \boldsymbol{V}_k^T$, $\boldsymbol{V}_k$ is a $K \times 1$ vector with the k^{th} element being 1 but zero elsewhere. Here $\mathbf{0}_p$ is a $p \times 1$ zero vector, $\boldsymbol{I}_p$ 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

$$\sum_{k=1}^{K} \sum_{i=1}^{n} \rho_{\tau_k}(y_i - \mathbf{z}_{ik}^T \boldsymbol{\theta}). \tag{3}$$

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 $\hat{\boldsymbol{\theta}}_{FAL} = \arg\min_{\boldsymbol{\theta}} Q(\boldsymbol{\theta})$, where

$$Q(\boldsymbol{\theta}) = \sum_{k=1}^{K} \sum_{i=1}^{n} \rho_{\tau_k}(y_i - \mathbf{z}_{ik}^T \boldsymbol{\theta}) + n\lambda_n \Big(\sum_{l=1}^{p} \sum_{k=1}^{K} \tilde{w}_{k,l} |\theta_{Kl+k}| + \sum_{l=1}^{p} \sum_{k=2}^{K} \tilde{v}_{k,l} |\theta_{Kl+k} - \theta_{Kl+k-1}| \Big), \tag{4}$$

and $\lambda_n > 0$ is the tuning parameter controlling the degree of penalization. For each $l = 1, \ldots, p$, we set the adaptive weights $\tilde{w}_{k,l} = |\tilde{\theta}_{Kl+k}|^{-1} = |\tilde{\beta}_{k,l}|^{-1}$, $k = 1, \ldots, K$, and $\tilde{v}_{k,l} = |\tilde{\theta}_{Kl+k} - \tilde{\theta}_{Kl+k-1}|^{-1} = |\tilde{\beta}_{k,l} - \tilde{\beta}_{k-1,l}|^{-1}$, $k = 2, \ldots, K$, where $\tilde{\beta}$ (and hence $\tilde{\theta}$) is the initial estimator obtained from the conventional quantile regression without any shrinkage. The component-wise weight $\tilde{w}_{k,l}$ controls the speed at which the slope for predictor at quantile K is shrunk to zero: the closer

the initial estimator is to zero, the faster it will be shrunk to zero. Likewise, $\tilde{v}_{k,l}$ controls the degree of smoothness between quantile coefficient process: the closer the initial slopes are, the faster they will be shrunk towards each other. Since both $\tilde{w}_{k,l}$ and $\tilde{v}_{k,l}$ 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 $\boldsymbol{\theta} = (\alpha^T, \boldsymbol{\beta}^T)^T$ and hence $(\theta_{K+1}, \ldots, \theta_{2K}) = (\beta_1, \ldots, \beta_K)$. Then the adaptive weights $\tilde{w}_k = |\tilde{\theta}_{K+k}|^{-1}$ for $k = 1, \ldots, K$, while $\tilde{v}_k = |\tilde{\theta}_{K+k} - \tilde{\theta}_{K+k-1}|^{-1}$ for $k = 2, \ldots, K$. Define $\boldsymbol{\theta}_0 = (\theta_{k,0} : k = 1, \ldots, 2K)^T$ as the true value of $\boldsymbol{\theta}$. The set of indices $\mathcal{A}_1 = \{j : \theta_{j,0} \neq 0, j = K+1, \ldots, 2K\}$ contains the indices of the true nonzero quantile slope coefficients, while $\mathcal{A}_2 = \{j : \theta_{j,0} \neq \theta_{j-1,0}, j = K+2, \ldots, 2K\}$ includes the indices of quantile slope coefficients that differ from their preceding neighboring quantile level. Further we set $\mathcal{A}_3 = \mathcal{A}_2 \cup \{K+1\}$ so that $\mathcal{A}_3$ also includes the first slope index, which does not have a preceding neighbor. Let $\mathcal{A} = \{1, \ldots, K\} \cup (\mathcal{A}_1 \cap \mathcal{A}_3)$, then the parameter vector $\boldsymbol{\theta}_{\mathcal{A}} = (\theta_j : j \in \mathcal{A})^T$ 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 $\boldsymbol{\theta}_0 = (\alpha_1, \ldots, \alpha_4, 0, 2, 2, 3)^T$. By the above definitions, $\mathcal{A}_1 = \{K+2, K+3, K+4\} = \{6, 7, 8\}$ and $\mathcal{A}_3 = \{K+1, K+2, K+4\} = \{5, 6, 8\}$. Hence, $\mathcal{A} = \{1, \ldots, K\} \cup \{K+2, K+4\} = \{1, 2, 3, 4, 6, 8\}$ and $\boldsymbol{\theta}_{\mathcal{A}} = (\alpha_1, \ldots, \alpha_4, \beta_2, \beta_4)^T$, 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, $\mathcal{A}_1 = \{K+1, K+2, K+3\} = \{5, 6, 7\}$ and $\mathcal{A}_3 = \{K+1, K+4\} = \{5, 8\}$. Hence, the index set $\mathcal{A} = \{1, \ldots, K\} \cup \{K+1\} = \{1, 2, 3, 4, 5\}$, and $\boldsymbol{\theta}_{\mathcal{A}} = (\alpha_1, \ldots, \alpha_4, \beta_1)^T$, where the common nonzero quantile slope coefficients β_1, β_2 and β_3 are collapsed to β_1 as a unique

representative.

Under the assumption that the true model structure has been known beforehand, we can estimate $\boldsymbol{\theta}_{\mathcal{A}}$ by the oracle estimator

$$\hat{\boldsymbol{\theta}}_{\mathcal{A}} = \arg\min_{\boldsymbol{\theta}_{\mathcal{A}}} \sum_{k=1}^{K} \sum_{i=1}^{n} \rho_{\tau_k}(y_i - \mathbf{z}_{ik,\mathcal{A}}^T \boldsymbol{\theta}_{\mathcal{A}}), \tag{5}$$

where $\mathbf{z}_{ik,\mathcal{A}} = (z_{ik,j} : j \in \mathcal{A})^T$ is a subset of $\mathbf{z}_{ik}$ whose indices of elements belong to the set $\mathcal{A}$. 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 $\mathbf{X} = \mathbf{x}_i$, denoted as f_i, is continuous and has a bounded first derivative, and $f_i\{Q_{\tau_k}(\mathbf{x}_i)\}$ is uniformly bounded away from zero and infinity.

(A2) $\max_{1 \le i \le n} \|\mathbf{x}_i\| = o(n^{1/2})$.

(A3) For $1 \le k \le K$, there exist some positive definite matrices $\boldsymbol{\Gamma}_k$ and $\boldsymbol{\Omega}_k$ such that $\lim_{n\to\infty} n^{-1} \sum_{i=1}^{n} \mathbf{z}_{ik}\mathbf{z}_{ik}^T = \boldsymbol{\Gamma}_k$ and $\lim_{n\to\infty} n^{-1} \sum_{i=1}^{n} f_i\{Q_{\tau_k}(\mathbf{x}_i)\}\mathbf{z}_{ik}\mathbf{z}_{ik}^T = \boldsymbol{\Omega}_k$.

The following proposition shows the asymptotic property of the oracle estimator $\hat{\boldsymbol{\theta}}_{\mathcal{A}}$.

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

$$n^{1/2}(\hat{\boldsymbol{\theta}}_{\mathcal{A}} - \boldsymbol{\theta}_{\mathcal{A},0}) \xrightarrow{d} N(0, \boldsymbol{\Sigma}_{\mathcal{A}}), \text{ as } n \to \infty,$$

where $\boldsymbol{\theta}_{\mathcal{A},0}$ is the truth of $\boldsymbol{\theta}_{\mathcal{A}}$, and $\boldsymbol{\Sigma}_{\mathcal{A}} = \left(\sum_{k=1}^{K} \boldsymbol{\Omega}_{k,\mathcal{A}}\right)^{-1} \left\{\sum_{k=1}^{K} \tau_k(1-\tau_k)\boldsymbol{\Gamma}_{k,\mathcal{A}}\right\} \left(\sum_{k=1}^{K} \boldsymbol{\Omega}_{k,\mathcal{A}}\right)^{-1}$, $\boldsymbol{\Omega}_{k,\mathcal{A}}$ and $\boldsymbol{\Gamma}_{k,\mathcal{A}}$ are the top-left $(K+s)\times(K+s)$ submatrices of $\boldsymbol{\Omega}_k$ and $\boldsymbol{\Gamma}_k$, which are defined in condition (A3).

However, in practice, the model structure is typically unknown. Therefore, we estimate the full parameter vector $\boldsymbol{\theta} \in \boldsymbol{R}^{K+Kp}$ by $\hat{\boldsymbol{\theta}}_{FAL}$. Theorem 1 presents the asymptotic properties of $\hat{\boldsymbol{\theta}}_{FAL}$ 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: $Pr\Big[\{j : \hat{\theta}_{j,FAL} \neq 0, j = K+1, \ldots, 2K\} = \mathcal{A}_1$ and $\{j : \hat{\theta}_{j,FAL} \neq \hat{\theta}_{j-1,FAL}, j = K+2, \ldots, 2K\} = \mathcal{A}_2\Big] \to 1$;
(ii) asymptotic normality: $n^{1/2}(\hat{\boldsymbol{\theta}}_{\mathcal{A},FAL} - \boldsymbol{\theta}_{\mathcal{A},0}) \xrightarrow{d} N(0, \boldsymbol{\Sigma}_{\mathcal{A}})$, where $\boldsymbol{\Sigma}_{\mathcal{A}}$ 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_∞-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_∞-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 $\boldsymbol{d}_{(l)} = (\beta_{2,l} - \beta_{1,l}, \ldots, \beta_{K,l} - \beta_{K-1,l})^T$ as a vector of interquantile slope differences corresponding to the l^{th} predictor, which is essentially a linear transformation of $\boldsymbol{\beta}_{(l)}$. The fused adaptive sup-norm estimator $\hat{\boldsymbol{\theta}}_{FAS}$ can be obtained by minimizing

$$Q(\boldsymbol{\theta}) = \sum_{k=1}^{K} \sum_{i=1}^{n} \rho_{\tau_k}(y_i - \boldsymbol{z}_{ik}^T \boldsymbol{\theta}) + n\lambda_n \Big(\sum_{l=1}^{p} \tilde{w}_l \|\boldsymbol{\beta}_{(l)}\|_\infty + \sum_{l=1}^{p} \tilde{v}_l \|\boldsymbol{d}_{(l)}\|_\infty \Big), \tag{6}$$

where $\tilde{w}_l = (\|\tilde{\boldsymbol{\beta}}_{(l)}\|_\infty)^{-1}$ and $\tilde{v}_l = (\|\tilde{\boldsymbol{d}}_{(l)}\|_\infty)^{-1}$, $l = 1, \ldots, p$, are the group-wise adaptive weights. The initial estimators $\tilde{\boldsymbol{\beta}}_{(l)}$ and $\tilde{\boldsymbol{d}}_{(l)}$ 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 $\boldsymbol{\beta}_{(l),0}$ and $\boldsymbol{d}_{(l),0}$ as the truth of $\boldsymbol{\beta}_{(l)}$ and $\boldsymbol{d}_{(l)}$, respectively, and $\|\cdot\|$ as the L_1-norm. The index set $\mathcal{B}_1 = \{l : \|\boldsymbol{\beta}_{(l),0}\| \neq 0 \text{ and } \|\boldsymbol{d}_{(l),0}\| \neq 0\}$ 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 $\mathcal{B}_2 = \{l : \|\boldsymbol{\beta}_{(l),0}\| \neq 0 \text{ and } \|\boldsymbol{d}_{(l),0}\| = 0\}$ corresponds to the predictors with nonzero but common slope coefficients across all quantile levels, and $\mathcal{B}_3 = \{l : \|\boldsymbol{\beta}_{(l),0}\| = 0\}$ corresponds to the predictors with zero quantile slope coefficients across all quantile levels. Define $\mathcal{B} = \{(k,l) : k \in (1,\ldots,K), l \in \mathcal{B}_1\} \cup \{(k,l) : k = 1, l \in \mathcal{B}_2\}$ and $\boldsymbol{\beta}_{\mathcal{B}} = (\beta_{k,l} : (k,l) \in \mathcal{B})^T$, 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 $\boldsymbol{\theta}_{\mathcal{B}} = (\alpha_1,\ldots,\alpha_K,\boldsymbol{\beta}_{\mathcal{B}}^T)^T$ and $\boldsymbol{\theta}_{\mathcal{B},0}$ as the truth. For better understanding, we examine the following example.

Example 3. We still consider $K = 4$, but with $p = 3$. Let $\boldsymbol{\beta}_{(1)} = (0,2,2,3)^T$, $\boldsymbol{\beta}_{(2)} = (2,2,2,2)^T$, and $\boldsymbol{\beta}_{(3)} = (0,0,0,0)^T$. By the definitions, $1 \in \mathcal{B}_1, 2 \in \mathcal{B}_2, 3 \in \mathcal{B}_3$ and $\boldsymbol{\beta}_{\mathcal{B}} = (0,2,2,3,2)^T$, which includes the group $\boldsymbol{\beta}_{(1)}$ and collapses the common slope coefficients to the unique one in $\boldsymbol{\beta}_{(2)}$. 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 $\boldsymbol{\beta}_{(1)}$.

Without loss of generality, we reorder $\boldsymbol{\beta}_{\mathcal{B}}$ in the vector $\boldsymbol{\theta}_{\mathcal{B}}$ 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 $\hat{\boldsymbol{\theta}}_{\mathcal{B}}$ is defined as

$$\hat{\boldsymbol{\theta}}_{\mathcal{B}} = \arg\min_{\boldsymbol{\theta}_{\mathcal{B}}} \sum_{k=1}^{K} \sum_{i=1}^{n} \rho_{\tau_k}(y_i - \boldsymbol{z}_{ik,\mathcal{B}}^T \boldsymbol{\theta}_{\mathcal{B}}), \tag{7}$$

where $\boldsymbol{z}_{ik,\mathcal{B}}$ 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 $\boldsymbol{z}_{ik}$.

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

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

where $\boldsymbol{\Sigma}_{\mathcal{B}} = \left(\sum_{k=1}^{K} \boldsymbol{\Omega}_{k,\mathcal{B}}\right)^{-1} \left\{\sum_{k=1}^{K} \tau_k(1-\tau_k)\boldsymbol{\Gamma}_{k,\mathcal{B}}\right\} \left(\sum_{k=1}^{K} \boldsymbol{\Omega}_{k,\mathcal{B}}\right)^{-1}$, $\boldsymbol{\Omega}_{k,\mathcal{B}}$ *and* $\boldsymbol{\Gamma}_{k,\mathcal{B}}$ *are the top-left* $(K + Kq_1 + q_2) \times (K + Kq_1 + q_2)$ *submatrices of* $\boldsymbol{\Omega}_k$ *and* $\boldsymbol{\Gamma}_k$*, respectively.*

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

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: $Pr\Big([\{l : \|\hat{\boldsymbol{\beta}}_{(l)}\| \neq 0, \|\hat{\boldsymbol{d}}_{(l)}\| \neq 0\} = \mathcal{B}_1] \cap [\{l : \|\hat{\boldsymbol{\beta}}_{(l)}\| \neq 0, \|\hat{\boldsymbol{d}}_{(l)}\| = 0\} = \mathcal{B}_2] \cap [\{l : \|\hat{\boldsymbol{\beta}}_{(l)}\| = 0\} = \mathcal{B}_3]\Big) \to 1$*;*
(ii) asymptotic normality: $n^{1/2}(\hat{\boldsymbol{\theta}}_{\mathcal{B},FAS} - \boldsymbol{\theta}_{\mathcal{B},0}) \to N(0, \boldsymbol{\Sigma}_{\mathcal{B}})$*, where* $\boldsymbol{\Sigma}_{\mathcal{B}}$ *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

$$\hat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}} \sum_{k=1}^{K}\sum_{i=1}^{n} \rho_{\tau_k}(y_i - \boldsymbol{z}_{ik}^T\boldsymbol{\theta}), \text{ s.t. } \sum_{l=1}^{p}\sum_{k=1}^{K} \tilde{w}_{k,l}|\theta_{Kl+k}| + \sum_{l=1}^{p}\sum_{k=2}^{K} \tilde{v}_{k,l}|\theta_{Kl+k} - \theta_{Kl+k-1}| \leq t, \quad (8)$$

where $t > 0$ is a tuning parameter that plays a similar role as λ. 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

$$\hat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}} \sum_{k=1}^{K}\sum_{i=1}^{n} \rho_{\tau_k}(y_i - \boldsymbol{z}_{ik}^T\boldsymbol{\theta}), \text{ s.t. } \sum_{l=1}^{p} \tilde{w}_l\|\boldsymbol{\beta}_{(l)}\|_\infty + \sum_{l=1}^{p} \tilde{v}_l\|\boldsymbol{d}_{(l)}\|_\infty \leq t, \quad (9)$$

where the tuning parameter $t \in [0, t_2]$ with $t_2 = 2p$.

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

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

where $\text{loss(t)} = \sum_{k=1}^{K} \log\left[\sum_{i=1}^{n} \rho_{\tau_k}\{y_i - \boldsymbol{z}_{ik}^T \hat{\boldsymbol{\theta}}(t)\}\right]$ measures the goodness of fit; see Bondell et al. (2010) for a similar measure in multiple-quantile regression. The vector $\hat{\boldsymbol{\theta}}(t)$ 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 $edf(t)$ is also associated with the tuning parameter t. We set 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 (2009a), 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 (2009a), 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.

To conduct a simulation study, we consider a 6-dimensional case with $p = 6$ and quantile levels $\tau = \{0.3, 0.4, 0.5, 0.6, 0.7\}$. We run 500 simulations. The response y_i is generated from

$$y_i = \alpha + \beta_1 x_{i,1} + \ldots + \beta_6 x_{i,6} + \gamma x_{i,6} e_i, \quad i = 1, \ldots, 200, \tag{10}$$

where $\alpha = 0, \beta_1 = \beta_2 = 1, \beta_3 = \beta_4 = \beta_5 = 0, \beta_6 = 2$ and $\gamma = 3$. The error term $e_i \overset{i.i.d}{\sim} N(0, 1)$ and all predictors $x_{i,1}, \ldots, x_{i,6}$ are generated from $U(0, 1)$ independently. Thus, in the true model,

quantile slope coefficients $\beta_3(\tau) = \beta_4(\tau) = \beta_5(\tau) = 0$ and $\beta_1(\tau) = \beta_2(\tau) = 1$ for all τ. The only predictor that has a varying effect over τ is $x_{i,6}$, where $\beta_6(\tau) = \beta_6 + \gamma\Phi^{-1}(\tau)$ varies with respect to τ, and $\Phi^{-1}(\tau)$ is the τ^{th} quantile of $N(0, 1)$.

We assess the performance of the six approaches via various criteria. Table 1 gives the results for model structure identification ability. Note that we do not consider the RQ approach in Table 1 because it does not perform selection. The upper portion of Table 1 adopts AIC as the tuning method, while the lower portion is for BIC. Columns in Table 1 represent the various comparison criteria. The selection accuracy for slope coefficients at different quantile levels is reflected by TNb and TPb, where TNb (TPb) is defined as the proportion of times that every zero (nonzero) quantile slope coefficient is shrunk to zero (nonzero) correctly over 500 runs. The criterion TNd merely focuses on the interquantile slope differences and is defined as the proportion of times that the true zero slope differences at adjacent quantile levels are identified as zero. The criteria GTNb and GTPb indicate the groupwise selection accuracy. In this paper, we regard the quantile slope coefficients corresponding to one predictor as a group, and GTNb (GTPb) is defined as the averaged number of groups with zero (nonzero) slope coefficients across all quantile levels that are estimated as all zero (nonzero) over simulations. For the speicific design of data generation in this simulation study, we have 3 groups with slope coefficients being zero across all quantile levels, and the rest 3 with all nonzero ones as the truth.

In general, compared to AIC, BIC has a better capability of identifying the sparse and constant slope coefficients, see the higher TNb, TNd and GTNb in Table 1. However, BIC may suffer from some over-shrinkage, which results in the lower TPb and GTPb. If we compare different approaches, FAL usually performs better than both VAL and SCAD by imposing individual penalization on interquantile slope differences additionally.

The above statement is not convincing for me. The adaptive lasso and SCAD both have oracle properties for variable selection. I do not think we can conclude that adaptive lasso works better than SCAD due to its adaptive penalty.
Liewen: I agree. I just deleted that sentence. – Judy

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 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 Table 2. We adopt Mean of Integrated Squared Error (MISE), defined as the average of ISE over 500 simulations, where

$$ISE = \frac{1}{n}\sum_{i=1}^{n}\{(\alpha_k + \mathbf{x}_i^T\boldsymbol{\beta}_k) - (\hat{\alpha}_k + \mathbf{x}_i^T\hat{\boldsymbol{\beta}}_k)\}^2.$$

Here, $(\alpha_k + \mathbf{x}_i^T\boldsymbol{\beta}_k)$ and $(\hat{\alpha}_k + \mathbf{x}_i^T\hat{\boldsymbol{\beta}}_k)$ are the true and estimated τ_k^{th} conditional quantile of Y given $\mathbf{x}_i$.

Table 2 shows 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.

Remark 1. Results in Table 1 shows 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

$$y_i = \alpha + \beta_1 x_{i,1} + \beta_2 x_{i,2} + 5x_{i,2}e_i,\ i = 1, \ldots, 200,$$

where $x_{i,1} \overset{i.i.d}{\sim} U(0,1)$, $x_{i,2} \overset{i.i.d}{\sim} U(0,1)$, $e_i \overset{i.i.d}{\sim} N(0,1)$, $\alpha = 1, \beta_1 = 0$ and $\beta_2 = 2$. We consider three quantile levels $\tau = \{0.4, 0.5, 0.6\}$ for simplicity. Figure 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 0.6^{th} quantile slope is shrunk down to the 0.5^{th}, then down to

the 0.4^{th}, 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.

Table 1: The performance of VAL, FAL, VAS, FAS and SCAD methods in 6-dimensional case in model (10), where $\beta_6(\tau) = 2 + \Phi^{-1}(\tau)$ varies with τ, $\beta_3(\tau) = \beta_4(\tau) = \beta_5(\tau) = 0$, and $\beta_1(\tau) = \beta_2(\tau) = 1$.

AIC					
	TNb	TPb	TNd	GTNb	GTPb
VAL	0.198	0.594	0	1.556	2.562
FAL	0.264	0.744	0.226	1.72	2.734
VAS	0.634	0.984	0.05	2.498	2.982
FAS	0.198	1	0.958	1.618	3
SCAD	0.022	0.694	0	0.798	2.668
BIC					
	TNb	TPb	TNd	GTNb	GTPb
VAL	0.588	0.392	0.008	2.51	2.23
FAL	0.610	0.630	0.672	2.502	2.612
VAS	0.894	0.922	0.196	2.89	2.902
FAS	0.4	0.988	0.996	2.152	2.986
SCAD	0.118	0.422	0.002	1.508	2.136

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 all zero interquantile slope differences are identified as zero; GTNb/GTPb: the averaged number of groups with all zero/nonzero slope coefficients across quantiles being estimated as zero/nonzero over 500 simulations.

Table 2: MISE of VAL, FAL, VAS, FAS and SCAD methods in 6-dimensional case in model (10), where $\beta_6(\tau)$ varies with τ, $\beta_3(\tau) = \beta_4(\tau) = \beta_5(\tau) = 0$, and $\beta_1(\tau) = \beta_2(\tau) = 1$.

$100 \times MISE(100 \times S.E)$, AIC					
	$\tau = 0.3$	$\tau = 0.4$	$\tau = 0.5$	$\tau = 0.6$	$\tau = 0.7$
VAL	6.2 (0.3)	6.4 (0.3)	5.8(0.3)	5.7(0.3)	6.7(0.3)
FAL	6.2 (0.3)	5.8(0.3)	5.5(0.3)	5.5(0.3)	6.3(0.3)
VAS	5.7(0.3)	5.1(0.2)	5.1(0.2)	5.1(0.2)	6.8(0.3)
FAS	6.0 (0.3)	5.0(0.2)	4.8(0.2)	5.0(0.2)	6.2(0.3)
SCAD	5.8 (0.2)	6.0(0.3)	5.8(0.2)	5.8(0.2)	6.8(0.3)
$100 \times MISE(100 \times S.E)$, BIC					
	$\tau = 0.3$	$\tau = 0.4$	$\tau = 0.5$	$\tau = 0.6$	$\tau = 0.7$
VAL	6.5 (0.2)	7.5(0.4)	6.6(0.4)	6.5(0.3)	7.5(0.4)
FAL	6.3(0.3)	6.2(0.3)	5.7(0.3)	5.6(0.3)	6.9(0.3)
VAS	5.9(0.3)	5.3(0.2)	5.3(0.2)	5.4(0.2)	8.9(0.4)
FAS	6.2(0.3)	4.9(0.2)	4.7(0.2)	5.1(0.2)	6.6(0.3)
SCAD	7.4(0.3)	7.5(0.4)	7.1(0.3)	7.1(0.3)	8.5(0.4)
RQ	7.6(0.3)	7.0(0.3)	6.6(0.3)	6.6(0.3)	7.4(0.3)

MISE: Mean of Integrated Squared Errors among 500 simulations, followed by S.E in the parentheses. The top half of the table contains the results from using AIC in selecting the tuning parameter, while the bottom half is for BIC.

4. Analysis of 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 period 1965-1975, while the rest 90 observations are for period 1975-1985. Some countries may appear in both periods. Our preliminary analysis suggests that after removing the

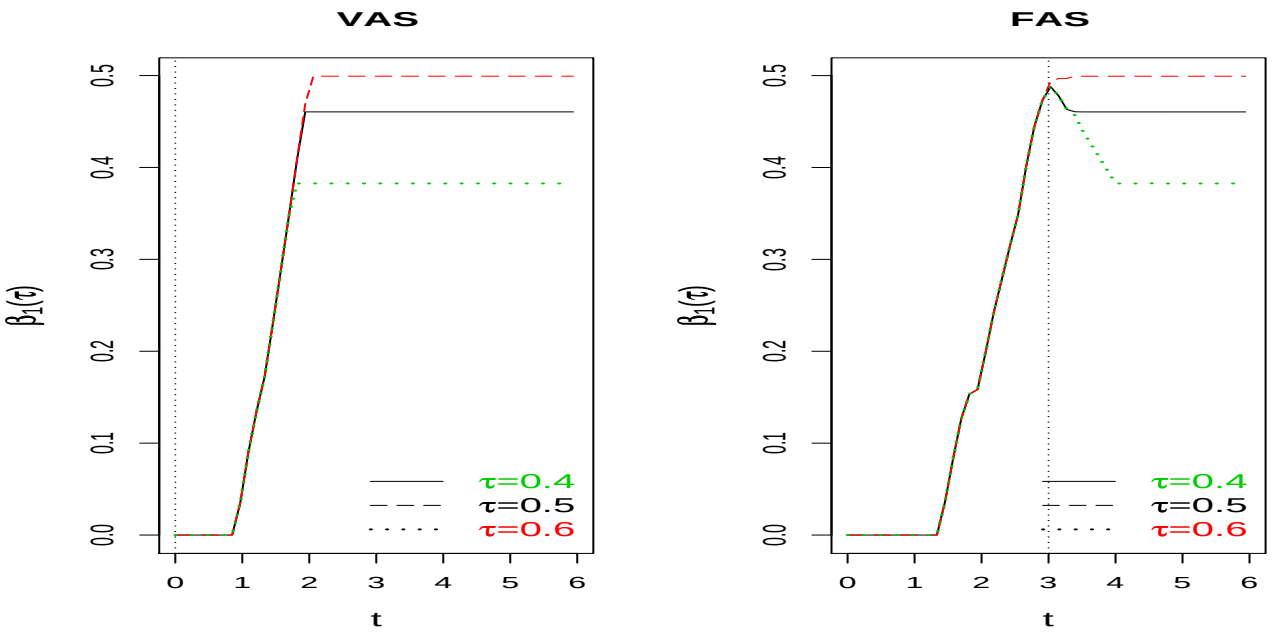

Figure 1: The solution path of $\beta_1(\tau)$ for VAS and FAS approaches at $\tau = 0.4$, $\tau = 0.5$ and $\tau = 0.6$. The dotted vertical line corresponds to the optimal t that gives the smallest BIC value.

predictors' effects at median, the data are weakly dependent between time points with correlation coefficient around -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), 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 (*ivst*), the ratio of public consumption and GDP growth (*pcon*), black market premium (*blakp*), political instability (*pol*) and growth rate terms trade (*ttrad*). All covariates are stan-

dardized 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 3. 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 coefficient to zero, but FAS couldn't 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 prediction accuracy, defined as

$$\text{PE} = \sum_{k=1}^{9} \sum_{j=1}^{50} \rho_{\tau_k}\{y_j - \mathbf{x}_j^T \hat{\boldsymbol{\beta}}(\tau_k)\},$$

where $\{(y_j, \mathbf{x}_j), j = 1, \ldots, 50\}$ are in the test set, $\hat{\boldsymbol{\beta}}(\tau_k)$ is the estimated coefficients at τ_k based on the training set. We repeat the cross validation 200 times and take the average of PE. Results in Table 4 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.

Table 3: The estimated quantile coefficients for predictors *fhe*, *edu*, *ivst* and *ttrad* in the economic growth data set.

	$\tau = 0.1$	$\tau = 0.2$	$\tau = 0.3$	$\tau = 0.4$	$\tau = 0.5$	$\tau = 0.6$	$\tau = 0.7$	$\tau = 0.8$	$\tau = 0.9$
					FAL, AIC				
fhe	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
edu	-0.39	-0.39	-0.03	-0.03	-0.03	-0.03	-0.03	0.00	0.00
ivst	0.67	0.67	0.67	0.67	0.67	0.67	0.67	0.60	0.60
ttrad	0.42	0.42	0.42	0.42	0.69	0.80	0.80	0.80	0.80
					FAS, AIC				
fhe	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
edu	-0.29	-0.29	-0.21	-0.14	-0.06	-0.12	-0.13	-0.06	-0.13
ivst	0.71	0.71	0.71	0.70	0.69	0.68	0.67	0.67	0.66
ttrad	0.51	0.46	0.52	0.58	0.64	0.70	0.75	0.80	0.80
					FAL, BIC				
fhe	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
edu	-0.35	-0.28	0.00	0.00	0.00	0.00	0.00	0.00	0.00
ivst	0.62	0.62	0.62	0.62	0.62	0.62	0.62	0.61	0.61
ttrad	0.39	0.39	0.39	0.39	0.70	0.70	0.70	0.70	0.70
					FAS, BIC				
fhe	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
edu	-0.12	-0.12	-0.12	-0.12	-0.12	-0.12	-0.12	-0.12	-0.12
ivst	0.63	0.63	0.63	0.63	0.63	0.63	0.63	0.63	0.63
ttrad	0.60	0.60	0.60	0.60	0.60	0.60	0.60	0.60	0.60

Table 4: Average prediction errors (PE) of the conventional quantile regression (RQ), VAL, FAL, VAS and FAS methods for the economic growth data. The tuning parameter is selected by AIC and BIC criteria. Entries in the parenthesis are the standard errors.

	AIC	BIC
VAL	257.68 (1.67)	258.17 (1.73)
FAL	253.86 (1.74)	254.23 (1.84)
VAS	253.95 (1.75)	255.17 (1.85)
FAS	251.97 (1.76)	252.76 (1.79)
RQ	258.94 (1.65)	258.94 (1.65)

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 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 on the scenario with low dimensionality only. How to handle high-dimensional data becomes a very active research area, especially with the advancement of modern technologies. As far as we know, there is a large field of applications of variable selection methods in high-dimensional data analysis. For example, Zou and Hastie (2005), Meier, Geer and Buhlmann (2009), Wang, Wu and Li (2012), to name a few. It is worth applying the proposed methodology to high-dimensional data in the near future, such as the functional data, which can be seen, in some sense, as very high dimensional data.

This last section needs to be worked on. I am not clear yet how to extend our method to functional data, and need read some related reference and the reference provided by Reviewer 1. Howard: can you try modify this Section first? If I got any ideas, I will add later. Thanks. – Judy

Acknowledgement

Bondell's research was supported in part by NSF grant DMS-1005612 and NIH grant PO1-CA-142538-01. Wang's research was supported by NSF grant DMS-1007420 and NSF CAREER Award DMS-1149355.

Akaike, H. (1974), A New Look at the Statistical Model Identification. *IEEE Transactions on Automatic Control* **19(6)**, 716-723.

Aneiros-Pérez, G., Ferraty, F. and Vieu, P. (2011), Variable Selection in Semi-functional regression model. In *Recent Advances in Functional Data Analysis and Related Topics*, Physica-Verlag, 17-23.

Barro, R. and Lee, J. (1994), Data Set for a Panel of 138 Countries. Cambridge, Mass.: National Bureau of Economic Research.

Belloni, A. and Chernozhukov, V. (2011). L_1-Penalized Quantile Regression in High-dimensional Sparse Models. *The Annals of Statistics* **39**, 82-130.

Bondell, H., Reich, B. and Wang, H. (2010), Non-crossing Quantile Regression Curve Estimation. *Biometrika* **97**, 825-838.

Bondell, H. and Reich, B. (2008), Simultaneous Regression Shrinkage, Variable Selection and Clustering of Predictors with OSCAR. *Biometrics* **64**, 115-123.

Bülhmann, P. and van de Geer, S. (2011). *Statistics for High-dimensional Data: Methods, Theory and Applications*. Springer.

Candes, E. and Tao, T. (2007), The Dantzig Selector: Statistical Estimation When p Is Much Larger Than n. *The Annals of Statistics* **35**, 2313-2351.

Efron, B., Hastie, T., Johnstone, I., Tibshirani, R. (2004), Least Angle Regression. *The Annals of Statistics* **32**, 407-499.

Fan, J. and Li, R. (2001), Variable Selection via Nonconcave Penalized Likelihood and Its Oracle Properties. *Journal of the American Statistical Association* **96**, 1348-1360.

Huang, J., Horowitz, L., Wei, F. (2010), Variable Selection in Nonparametric Additive Models. *The Annals of Statistics* **37**, 3779-3821.

Koenker, R. and Bassett, G. (1978), Regression Quantiles. *Econometrica* **4**, 33-50.

Koenker, R. (2004), Quantile Regression for Longitudinal Data. *Journal of Multivariate Analysis* **91**, 74-89.

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

Koenker, R. and Machado, J. (1999), Goodness of Fit and Related Inference Processes for Quantile Regression. *Journal of the American Statistical Association* **94**, 1296-1310.

Jiang, X., Jiang, J. and Song, X. (2012), Oracle Model Selection for Nonlinear Models Based on Weighted Composite Quantile Regression. *Statistica Sinica*, **22**, 1479-1506.

Jiang, L., Wang, H. and Bondell, H. (2013), Interquantile Shrinkage in Regression Models. *Journal of Computational and Graphical Statistics*, to appear.

Li, Y. and Zhu, J. (2007), Analysis of Array CGH Data for Cancer Studies Using Fused Quantile Regression. *Bioinformatics* **23**, 2470-2476.

Li, Y. and Zhu, J. (2008), L1-norm Quantile Regression. *Journal of Computational and Graphical Statistics* **17**, 163-185.

Meier, L., van de Geer, S. and Bühlmann, P. (2009), High-dimensional additive modeling. *The Annals of Statistics* **37**, 3779?821.

Meinshausen, N. (2007), Relaxed Lasso. *Computational Statistics and Data Analysis* **52**, 374-393.

Pollard, D. (1991), Asympototics for Least Absolute Deviation Regression Estimators. *Econometric Theory* **7**, 186-199.

Ravikumar, P., Lafferty, J., Liu, H. and Wasserman, L. (2009), Sparse Additive Models. *Journal of the Royal Statistical Society, series B* **71**, 1009-1030.

Schwarz, G. (1978), Estimating the Dimension of a Model. *The Annals of Statistics* **6**, 461-464.

Tibshirani, R. (1996), Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society, series B* **58**, 267-288.

Tibshirani, R., Saunders, M., Rosset, S., Zhu, J., and Knight, K. (2005), Sparsity and Smoothness via the Fused Lasso. *Journal of the Royal Statistical Society, series B* **67**, 91-108.

Wang, H. and Hu, J. (2011), Identification of Differential Aberrations in Multiple-sample Array CGH Studies. *Biometrics* **67**, 353-362.

Wang, H., Zhou, J. and Li, Y. (2012). Variable Selection for Censored Quantile Regression. *Statistica Sinica*, in press.

Wu, Y. and Liu, Y. (2009a), Variable Selection in Quantile Regression. *Statistica Sinica* **19**, 801-817.

Wu, Y. and Liu, Y. (2009b), Stepwise Multiple Quantile Regression Estimation Using Non-crossing Constraints. *Statistica and Its Interface* **2**, 299-310.

Yuan, M. and Lin, Y. (2006), Model Selection and Estimation in Regression with Grouped Variables. *Journal of the Royal Statistical Society, series B* **68**, 49-67.

Zou, H. (2006), The Adaptive Lasso and Its Oracle Properties. *Journal of the American Statistical Association*

101, 1418-1429.

Zou, H. and Yuan, M. (2008a), Composite Quantile Regression and The Oracle Model Selection Theory. *The Annals of Statistics* **36**, 1108-1126.

Zou, H. and Yuan, M. (2008b), Regularized Simultaneous Model Selection in Multiple Quantiles Regression. *Computational Statistics and Data Analysis* **52**, 5296-5304.

Zou, H. and Hastie, T. (2005), Regularization and Variable Selection via the Elastic Net. *Journal of the Royal Statistical Society, series B* **67**, 301-320.

Meier, L., Geer, S., and Buhlmann, P. (2009), High-dimensional additive modeling. *The Annals of Statistics* **37**, 3779-3821.

Wang, L., Wu, Y. and Li, R. (2012), Quantile Regression for Analyzing Heterogenity in Ultra-High Dimension. *Journal of the American Statistical Association* **107**, 214-222.

Wang, H. (2009), Inference on Quantile Regression for Heteroscedastic Mixed Models. *Statistica Sinica* **19**, 1247-1261.

Yin, G. S. and Cai, J. (2005), Quantile Regression Models with Multivariate Failure Time Data. *Biometrics* **61**, 151-161.

Appendix

We omit the proofs of Propositions 1 and 2, as they are similar to those in Jiang, Wang and Bondell (2013).

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

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

Lemma 2. *[Root-n consistency of* $\hat{\boldsymbol{\theta}}_{FAL}$*] Under conditions (A1)-(A3), if* $n^{1/2}\lambda_n \to 0$ *as* $n \to \infty$, *then* $\hat{\boldsymbol{\theta}}_{FAL} - \boldsymbol{\theta}_0 = O_p(n^{-1/2})$.

Proof. The proof of Lemma 2 is similar to that in Jiang, Wang and Bondell (2013) and thus is skipped.

Proof of Theorem 1. We first prove the consistency in selection. By reordering $\boldsymbol{\theta}$, we can decompose it as $\boldsymbol{\theta} = (\boldsymbol{\theta}^T_{\mathcal{A}}, \boldsymbol{\theta}^T_{\mathcal{A}_1^C}, \boldsymbol{\theta}^T_{\mathcal{A}^C \setminus \mathcal{A}_1^C})^T$, where $\mathcal{A}, \mathcal{A}_1$ are defined in Section 2.2.

Case 1. Suppose $\hat{\boldsymbol{\theta}}^T_{\mathcal{A}_1^C}$ is not selected correctly, that is, some elements in $\hat{\boldsymbol{\theta}}^T_{\mathcal{A}_1^C}$ are not estimated as zero. Without loss of generality, we assume there is only one element $\hat{\theta}_j \neq 0$, where $j \in \mathcal{A}_1^C$. Cases with more than one element in $\boldsymbol{\theta}_{\mathcal{A}_1^C}$ are not selected correctly basically follow the same elaborations, but with more complicated notations. Let $\boldsymbol{\theta}^*$ be a vector constructed by replacing $\hat{\theta}_j$ with 0, other elements are the same as $\hat{\boldsymbol{\theta}}$. We can show that $Q(\boldsymbol{\theta}^*) < Q(\hat{\boldsymbol{\theta}})$, which contradicts the fact that $\hat{\boldsymbol{\theta}}$ minimizes the objective function $Q(\boldsymbol{\theta})$ defined in (4). Define

$$L_n(\boldsymbol{u}) \;=\; \sum_{k=1}^{K}\sum_{i=1}^{n}\left\{\rho_{\tau_k}(y_i - \boldsymbol{z}_{ik}^T\boldsymbol{\theta}_0 - n^{-1/2}\boldsymbol{z}_{ik}^T\boldsymbol{u}) - \rho_{\tau_k}(y_i - \boldsymbol{z}_{ik}^T\boldsymbol{\theta}_0)\right\}.$$

As was shown in Jiang, Wang and Bondell (2013), $L_n(\boldsymbol{u})$ is a bounded quantity provided that $\|\boldsymbol{u}\|$ is bounded.

Is $L_n(\cdot)$ bounded all any u?

Liewen: I added some words here. I think as long as $\|u\|$ is bounded, $L_n(u)$ becomes bounded. – Judy

Note that

$$\begin{aligned} Q(\boldsymbol{\theta}^*) - Q(\hat{\boldsymbol{\theta}}) \;&=\; L_n\{n^{1/2}(\boldsymbol{\theta}^* - \boldsymbol{\theta}_0)\} - L_n\{n^{1/2}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)\} - n\lambda_n\tilde{w}_j|\hat{\theta}_j| + n\lambda_n\tilde{v}_j\left(|\hat{\theta}_{j-1}| - |\hat{\theta}_j - \hat{\theta}_{j-1}|\right) \\ &+\; n\lambda_n\tilde{v}_{j+1}\left(|\hat{\theta}_{j+1}| - |\hat{\theta}_{j+1} - \hat{\theta}_j|\right). \end{aligned}$$

Case 1a. If $\theta_{j,0} - \theta_{j-1,0} \neq 0$ and $\theta_{j+1,0} - \theta_{j,0} \neq 0$, then $\tilde{v}_j \to_p |\theta_{j,0} - \theta_{j-1,0}|^{-1}$ and $\tilde{v}_{j+1} \to_p |\theta_{j+1,0} - \theta_{j,0}|^{-1}$. Since $|\hat{\theta}_{j-1}| - |\hat{\theta}_j - \hat{\theta}_{j-1}| \leq |\hat{\theta}_j|$ and $|\hat{\theta}_{j+1}| - |\hat{\theta}_{j+1} - \hat{\theta}_j| \leq |\hat{\theta}_j|$, we have

$$Q(\boldsymbol{\theta}^*) - Q(\hat{\boldsymbol{\theta}}) \leq L_n\{n^{1/2}(\boldsymbol{\theta}^* - \boldsymbol{\theta}_0)\} - L_n\{n^{1/2}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)\} - n\lambda_n\tilde{w}_j|\hat{\theta}_j| + n\lambda_n\tilde{v}_j|\hat{\theta}_j| + n\lambda_n\tilde{v}_{j+1}|\hat{\theta}_j|, \quad (11)$$

where $n\lambda_n\tilde{v}_j|\hat{\theta}_j| = n^{1/2}\lambda_n\tilde{v}_j|n^{1/2}\hat{\theta}_j| \to_p 0$, and $n\lambda_n\tilde{v}_{j+1}|\hat{\theta}_j| = n^{1/2}\lambda_n\tilde{v}_{j+1}|n^{1/2}\hat{\theta}_j| \to_p 0$ as $n \to \infty$, due to the fact that $n^{1/2}\lambda_n \to 0$ and the root-n consistency of $\hat{\boldsymbol{\theta}}$. However, note that $-n\lambda_n\tilde{w}_j|\hat{\theta}_j| = -n\lambda_n(n^{1/2}\tilde{\theta}_j)^{-1}|n^{1/2}\hat{\theta}_j|$ dominates the right hand side of (11) as $n\lambda_n \to \infty$. Hence $Q(\boldsymbol{\theta}^*) < Q(\hat{\boldsymbol{\theta}})$ holds.

Case 1b. If one of $\theta_{j,0} - \theta_{j-1,0}$ and $\theta_{j+1,0} - \theta_{j,0}$ is 0, but the other one is nonzero. Without loss of generality, suppose $\theta_{j,0} - \theta_{j-1,0} \neq 0$, but $\theta_{j+1,0} - \theta_{j,0} = 0$. Under the assumption that only $\hat{\theta}_j \neq 0$ is not correctly selected, but other elements in $\boldsymbol{\theta}$ are selected correctly, we have $\hat{\theta}_{j+1} = 0$, given the truth $\theta_{j+1,0} = 0$. Hence

$$\begin{aligned} Q(\boldsymbol{\theta}^*) - Q(\hat{\boldsymbol{\theta}}) &= L_n\{n^{1/2}(\boldsymbol{\theta}^* - \boldsymbol{\theta}_0)\} - L_n\{n^{1/2}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)\} - n\lambda_n \tilde{w}_j |\hat{\theta}_j| \\ &+ n\lambda_n \tilde{v}_j (|\hat{\theta}_{j-1}| - |\hat{\theta}_j - \hat{\theta}_{j-1}|) - n\lambda_n \tilde{v}_{j+1} |\hat{\theta}_j|, \end{aligned} \tag{12}$$

where $n\lambda_n \tilde{v}_j (|\hat{\theta}_{j-1}| - |\hat{\theta}_j - \hat{\theta}_{j-1}|) \leq n^{1/2}\lambda_n \tilde{v}_j |n^{1/2}\hat{\theta}_j| \to_p 0$. So (12) is dominated by $-n\lambda_n \tilde{w}_j |\hat{\theta}_j| - n\lambda_n \tilde{v}_{j+1} |\hat{\theta}_j|$, thus $Q(\boldsymbol{\theta}^*) < Q(\hat{\boldsymbol{\theta}})$ holds.

Case 1c. If $\theta_{j-1,0} = \theta_{j,0} = \theta_{j+1,0} = 0$. Under the assumption that only $\hat{\theta}_j \neq 0$ is not selected correctly, we have $\hat{\theta}_{j-1} = \hat{\theta}_{j+1} = 0$, and

$$Q(\boldsymbol{\theta}^*) - Q(\hat{\boldsymbol{\theta}}) = L_n\{n^{1/2}(\boldsymbol{\theta}^* - \boldsymbol{\theta}_0)\} - L_n\{n^{1/2}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)\} - n\lambda_n \tilde{w}_j |\hat{\theta}_j| - n\lambda_n \tilde{v}_j |\hat{\theta}_j| - n\lambda_n \tilde{v}_{j+1} |\hat{\theta}_j| < 0.$$

Therefore, $Q(\boldsymbol{\theta}^*) < Q(\hat{\boldsymbol{\theta}})$ holds, which contradicts the fact that $\hat{\boldsymbol{\theta}}$ is the minimizer to $Q(\boldsymbol{\theta})$.

Case 2. Now suppose there exists one $j' \in \mathcal{A}^C \backslash \mathcal{A}_1^C$, where the true $d_{j',0} = 0$, but $\hat{d}_{j'} \neq 0$. Since $j' \notin \mathcal{A}_1^C$, we have $\theta_{j',0} \neq 0$, and $\theta_{j'-1,0} \neq 0$. In fact, the case of $\theta_{j'-1,0} = \theta_{j',0} = 0$ has been discussed in Case 1. Let $\boldsymbol{\theta}^*$ be a vector constructed by restricting $d^*_{j'} = 0$, that is, every element in $\boldsymbol{\theta}^*$ is the same as the one in $\hat{\boldsymbol{\theta}}$, except $\theta^*_{j'} \neq \hat{\theta}_{j'}$, which is to say, $d^*_{j'} \neq \hat{d}_{j'}$ and $d^*_{j'+1} \neq \hat{d}_{j'+1}$. Without loss of generality, we assume $d_{j'+1,0} \neq 0$. We can show that $Q(\boldsymbol{\theta}^*) < Q(\hat{\boldsymbol{\theta}})$. Note that

$$\begin{aligned} Q(\boldsymbol{\theta}^*) - Q(\hat{\boldsymbol{\theta}}) &= L_n\{n^{1/2}(\boldsymbol{\theta}^* - \boldsymbol{\theta}_0)\} - L_n\{n^{1/2}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)\} + n\lambda_n \tilde{w}_{j'} (|\theta^*_{j'}| - |\hat{\theta}_{j'}|) - n\lambda_n \tilde{v}_{j'} |\hat{d}_{j'}| \\ &+ n\lambda_n \tilde{v}_{j'+1} (|d^*_{j'+1}| - |\hat{d}_{j'+1}|). \end{aligned} \tag{13}$$

Since $\theta^*_{j'} = \theta^*_{j'-1} = \hat{\theta}_{j'-1}, n\lambda_n \tilde{w}_{j'} (|\theta^*_j| - |\hat{\theta}_{j'}|) \leq \sqrt{n}\lambda_n \tilde{w}_{j'} \sqrt{n}|\hat{\theta}_{j'} - \hat{\theta}_{j'-1}| \to_p 0$. Moreover, $\theta^*_{j'+1} = \hat{\theta}_{j'+1}$, and $|d^*_{j'+1}| - |\hat{d}_{j'+1}| \leq |\hat{\theta}_{j'} - \theta^*_{j'}| = |\hat{\theta}_{j'} - \hat{\theta}_{j'-1}|$. Thus, $n\lambda_n \tilde{v}_{j'+1} (|d^*_{j'+1}| - |\hat{d}_{j'+1}|) \leq \sqrt{n}\lambda_n \tilde{v}_{j'+1} \sqrt{n}|\hat{\theta}_{j'} - \hat{\theta}_{j'-1}| \to_p 0$. Therefore, $-n\lambda_n \tilde{v}_{j'} |\hat{d}_{j'}|$ dominates the right hand side of (13), and $Q(\boldsymbol{\theta}^*) < Q(\hat{\boldsymbol{\theta}})$ holds, which contradicts the fact that $\hat{\boldsymbol{\theta}}$ is the minimizer to $Q(\cdot)$.

The asymptotic normality can be shown in a similar way as in the proof of Theorem 1 in Jiang, Wang and Bondell (2013). We hence skip the proof here.

Lemma 3 (Root-n consistency of $\hat{\boldsymbol{\theta}}_{FAS}$). *Under conditions A1-A3, if $n^{1/2}\lambda_n \to 0$ as $n \to \infty$, we have $\hat{\boldsymbol{\theta}}_{FAS} - \boldsymbol{\theta}_0 = O_p(n^{-1/2})$.*

Proof. The proof is omitted since it follows similar arguments as in the proof of Lemma 3 in Jiang, Wang and Bondell (2013).

Proof of Theorem 2. We prove the consistency in selection first. Suppose there exists $l^{'} \in \mathcal{B}_3$, where the true $\|\boldsymbol{\beta}_{(l^{'}),0}\| = 0$, but in the estimator $\hat{\boldsymbol{\theta}}$, we have $\|\hat{\boldsymbol{\beta}}_{(l^{'})}\| \neq 0$. Let $\boldsymbol{\theta}^*$ be a vector constructed by restricting $\|\boldsymbol{\beta}^*_{(l^{'})}\| = 0$. We can show that $Q(\boldsymbol{\theta}^*) < Q(\hat{\boldsymbol{\theta}})$ always holds, which contradicts the fact that $\hat{\boldsymbol{\theta}}$ is the minimizer to $Q(\cdot)$ defined in (6).

$$Q(\boldsymbol{\theta}^*) - Q(\hat{\boldsymbol{\theta}}) = L_n\{n^{1/2}(\boldsymbol{\theta}^* - \boldsymbol{\theta}_0)\} - L_n\{n^{1/2}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)\} - n\lambda_n\tilde{w}_{l^{'}}\|\hat{\boldsymbol{\beta}}_{(l^{'})}\|_\infty - n\lambda_n\tilde{v}_{l^{'}}\|\hat{\boldsymbol{d}}_{(l^{'})}\|_\infty. \quad (14)$$

When the true $\|\boldsymbol{\beta}_{(l),0}\| = 0$, (14) is dominated by $-n\lambda_n\tilde{w}_{l^{'}}\|\hat{\boldsymbol{\beta}}_{(l^{'})}\|_\infty - n\lambda_n\tilde{v}_{l^{'}}\|\hat{\boldsymbol{d}}_{(l^{'})}\|_\infty$. Hence $Q(\boldsymbol{\theta}^*) < Q(\hat{\boldsymbol{\theta}})$ always holds.

Similarly, suppose there exists a $l_2 \in \mathcal{B}_2$, where the true $\|\boldsymbol{d}_{(l_2),0}\| = 0$, but in the estimator $\hat{\boldsymbol{\theta}}$, $\|\hat{\boldsymbol{d}}_{(l_2)}\| \neq 0$. Let $\boldsymbol{\theta}^*$ be a vector constructed by restricting $\|\boldsymbol{d}^*_{(l_2)}\| = 0$. Hence,

$$\begin{aligned} Q(\boldsymbol{\theta}^*) - Q(\hat{\boldsymbol{\theta}}) \quad = \quad & L_n\{n^{1/2}(\boldsymbol{\theta}^* - \boldsymbol{\theta}_0)\} - L_n\{n^{1/2}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)\} + n\lambda_n\tilde{w}_{l_2}(\|\boldsymbol{\beta}^*_{(l_2)}\|_\infty - \|\hat{\boldsymbol{\beta}}_{(l_2)}\|_\infty) \\ & -n\lambda_n\tilde{v}_{l_2}\|\hat{\boldsymbol{d}}_{(l_2)}\|_\infty, \end{aligned} \quad (15)$$

where the last term $-n\lambda_n\tilde{v}_{l_2}\|\hat{\boldsymbol{d}}_{(l_2)}\|_\infty$ dominates. Hence $Q(\boldsymbol{\theta}^*) < Q(\hat{\boldsymbol{\theta}})$. The asymptotic normality can be shown in a similar way as in the proof of Theorem 2 in Jiang, Wang and Bondell (2013).