Statistica Sinica **27** (2017), 1619-1638
doi:https://doi.org/10.5705/ss.202016.0081

DANTZIG-TYPE PENALIZATION FOR MULTIPLE QUANTILE REGRESSION WITH HIGH DIMENSIONAL COVARIATES

Seyoung Park[1], Xuming He[2] and Shuheng Zhou[2]

[1] *Yale University and* [2] *University of Michigan*

Abstract: We study joint quantile regression at multiple quantile levels with high-dimensional covariates. Variable selection performed at individual quantile levels may lack stability across neighboring quantiles, making it difficult to understand and to interpret the impact of a given covariate on conditional quantile functions. We propose a Dantzig–type penalization method for sparse model selection at each quantile level which, at the same time, aims to shrink differences of the selected models across neighboring quantiles. We show model selection consistency, and investigate the stability of the selected models across quantiles. We also provide asymptotic normality of post–model–selection parameter estimation in the multiple quantile framework. We use numerical examples and data analysis to demonstrate that the proposed Dantzig–type quantile regression model selection method provides stable models for both homogeneous and heterogeneous cases.

Key words and phrases: Fused lasso, high dimensional data, model selection, quantile regression, stability.

1. Introduction

Quantile regression has become a widely used method to evaluate the effect of regressors on the conditional distribution of a response variable (Koenker (2005)). Compared with linear regression analysis, quantile regression is less sensitive to the misspecification of error distributions and provides more comprehensive information on the relationship between the response variable and covariates. Due to the ubiquity of high-dimensional problems in a variety of modern applications ranging from signal processing to genomics, it is critical to understand quantile regression in high-dimensional settings. We focus on cases where p, the number of covariates, is greater than n, the sample size.

There has been a line of recent work on variable selection for quantile regression models (Li and Zhu (2008); Zou and Yuan (2008a,b); Wu and Liu (2009)). In the high dimensional setting, penalization methods with the ℓ_1 penalty (Belloni and Chernozhukov (2011); Wang (2013)), the weighted ℓ_1 penalty (Zheng, Gallagher and Kulasekera (2013); Fan, Fan and Barut (2014)), and the smoothly

clipped absolute deviation (SCAD) penalty (Wang, Wu and Li (2012); Fan, Xue and Zou (2014)) have been used to obtain consistent model selection. Belloni and Chernozhukov (2011) establish consistency in parameter estimation with the ℓ_1 penalty. Wang, Wu and Li (2012) consider the SCAD penalty, and show that the oracle estimate is one of the local minima of a non-convex optimization problem. Fan, Fan and Barut (2014) use the weighted ℓ_1 penalty based on the SCAD penalty function, and establish model selection consistency and asymptotic normality.

Although these works establish nice theoretical properties, empirical evidence suggests that the sets of variables selected at nearby quantiles often differ excessively. Stability of selected variables across quantiles is desirable both for interpretation of results and understanding the impact of a particular covariate on the conditional quantile functions. For example, a covariate that is selected at quantiles 0.5 and 0.6, but not at 0.55, would not be much appreciated unless there is a strong reason. The motivation and the main contribution of our work is to show that joint modeling across quantiles can lead to stable models. Zou and Yuan (2008a,b), Bang and Jhun (2012), Jiang, Wang and Bondell (2013), Peng, Xu and Kutner (2014), and Volgushev, Wagener and and Dette (2014) consider joint quantile regression and provide consistent estimators. He (1997), Dette and Volgushev (2008), Bondell, Reich and Wang (2010), and Jang and Wang (2015) study non-crossing quantile regression at multiple quantiles. Zheng, Peng and He (2015) focus on the selection of all variables that impact one of the quantile functions. The present paper aims to identify what impacts each quantile function by allowing subsets of covariates for each quantile to vary slowly across quantiles.

In this paper, we consider joint quantile regression in the high dimensional setting, where the number of potential covariates, as well as the number of quantiles, is allowed to increase with n. The penalty that we use consists of two components: the first shrinks the magnitudes of the coefficients toward zero; the second controls the rate of changes in coefficients at adjacent quantiles. Both contribute to sparse and stable model selection across quantiles. We propose to minimize the combined penalty in a way similar to the Dantzig selector proposed by Candes and Tao (2007). Throughout this paper, the size of set differences of the selected models at adjacent quantiles and the size of the union of the selected covariates across all quantiles of interest is utilized to quantify the stability of selected models. Moreover, we study a post–selection quantile regression estimator and establish its asymptotic distribution.

The rest of the paper is organized as follows. In Section 2, we describe the quantile regression model and our method. Its theoretical properties are presented in Section 3. An implementation of the proposed method is described in Section 4. In Section 5, we show consistency in model selection. Section 6 discusses post–selection joint quantile regression and its theoretical properties. We show simulation results in Section 7. A data example and some concluding remarks are given in Section 8 and Section 9, respectively. Proofs and the additional simulation study are presented in the Supplementary material.

2. Model and Method

Let $X = (x_1, \cdots, x_n)^T$ be an $n \times p$ fixed design matrix and $Y = (y_1, \cdots, y_n)^T \in \mathbb{R}^n$ be an n-dimensional response vector. Consider the following quantile regression model at multiple quantiles $0 < \tau_1 < \cdots < \tau_{K_n} < 1$, where K_n is allowed to increase with n:

$$Y = X\beta(\tau_k) + \epsilon^{(k)} \quad (k = 1, \ldots, K_n). \tag{2.1}$$

Here $\beta(\tau_k) \in \mathbb{R}^p$ is a τ_k-th quantile coefficient vector in the sense that ${x_i}^T\beta(\tau_k)$ is the τ_k-th quantile of y_i evaluated at x_i, which is called the conditional quantile of y_i given x_i. The $\epsilon^{(k)} = (\epsilon_1^{(k)}, \cdots, \epsilon_n^{(k)})^T$ is an n-dimensional vector with mutually independent elements and

$$\mathbb{P}\left[\epsilon_i^{(k)} \leq 0 \mid x_i\right] = \tau_k \quad (i = 1, \ldots, n;\ k = 1, \ldots, K_n).$$

In the special case in which we have a linear model with i.i.d. errors, $\epsilon^{(k)}$ would depend on k only through a location shift. Our model assumes that the conditional quantile of y_i given x_i is linear at each τ_k, but no distributional assumptions are made on $\epsilon^{(k)}$. Let $T^{(k)}$ be the support set of $\beta(\tau_k)$ and $B^{(k)}$ be the set of indices where the quantile coefficients at the τ_k-th quantile are different from those at the τ_{k-1}-th quantile;

$$\begin{aligned} T^{(k)} &= \{j \in \{1, \ldots, p\} : \beta_j(\tau_k) \neq 0\} \quad (k = 1, \ldots, K_n), \\ B^{(k)} &= \{j \in \{1, \ldots, p\} : \beta_j(\tau_k) \neq \beta_j(\tau_{k-1})\} \quad (k = 2, \ldots, K_n). \end{aligned} \tag{2.2}$$

Let $s_k = |T^{(k)}|$ denote the sparsity level of the model for the τ_k-th quantile. We consider a high-dimensional sparse model with $\max(n, K_n) = o(p)$, where $p = o\left(\exp(n^b)\right)$ for some constant $b > 0$. Let $s_0 := \max_k s_k$. Our goal is to recover support sets $T^{(k)}$ $(k = 1, \ldots, K_n)$, $B^{(k)}$ $(k = 2, \ldots, K_n)$, and coefficient vectors $\beta(\tau_k)$ $(k = 1, \ldots, K_n)$.

Let $w^{(k)}$ $(k = 1, \ldots, K_n)$ and $v^{(k)}$ $(k = 2, \ldots, K_n)$ be p-dimensional vectors

of non-negative weights. Let λ be a regularization parameter, and $r_k > 0$ for $k = 1, \ldots, K_n$ be constraint parameters to be chosen. We consider the convex optimization problem:

$$\min_{\mathcal{B}=[\beta^{(1)},\cdots,\beta^{(K_n)}]\in\mathbb{R}^{p\times K_n}} \sum_{k=1}^{K_n}\sum_{j=1}^{p} w_j^{(k)}|\beta_j^{(k)}| + \lambda \sum_{k=2}^{K_n}\sum_{j=1}^{p} v_j^{(k)} \frac{|\beta_j^{(k)} - \beta_j^{(k-1)}|}{|\tau_k - \tau_{k-1}|}, \quad (2.3)$$

$$\text{s.t. } \forall k, \quad \beta^{(k)} \in \mathcal{R}^{(k)}(r_k) = \left\{ \beta \in \mathbb{R}^p : \ \frac{1}{n}\sum_{i=1}^{n} \rho_{\tau_k}(y_i - {x_i}^T\beta) \le r_k \right\}, \quad (2.4)$$

where $\rho_\tau(t) = t(\tau - 1\{t \le 0\})$ is the τ-th quantile loss function (Koenker and Basset (1978)).

Let $\hat{\mathcal{B}} = [\hat{\beta}^{(1)}, \cdots, \hat{\beta}^{(K)}]$ be any optimal solution to (2.3) and (2.4). The $\hat{\mathcal{B}}$ is used to estimate true parameter $\mathcal{B}^o = [\beta(\tau_1), \ldots, \beta(\tau_{K_n})]$. In (2.3), two types of penalties are required to simultaneously provide sparse and stable models. The first one, a sparsity penalty, aims to obtain a sparse model. The second one, a weighted total variation penalty (WTV), controls the rate of changes in quantile coefficient vectors; see related work by Rudin, Osher and Fatemi (1992) and Tibshirani et al. (2005). The feasible set of the optimization problem (2.3) is non-empty for any choice of positive r_k's because there always exists $\beta \in \mathbb{R}^p$ satisfying $Y = X\beta$ provided the column space of X spans $\mathbb{R}^n$.

2.1. Notations

Throughout the paper, it is to be understood that the design matrix X is normalized to have column ℓ_2 norm $\sqrt{n}$, and is non-stochastic. The quantities p, s_0, and K_n depend on the sample size n. Given a vector $\delta = (\delta_1, \cdots, \delta_p)^T \in \mathbb{R}^p$ and a set of indices $S \subset \{1, \ldots, p\}$, denote by $\delta_S \in \mathbb{R}^p$ the vector with the jth component $\delta_{S,j} = \delta_j I(j \in S)$. Let $\|\delta\|_0$, $\|\delta\|_\infty$, and $\|\delta\|_q$ for any positive integer q be the number of non-zero components, the maximum absolute value, and the ℓ_q norm of δ, respectively. Let S^c be the complement set of S. For p-dimensional vectors $\beta^{(1)}, \cdots, \beta^{(K)}$, let $[\beta^{(1)}, \cdots, \beta^{(K)}]$ be the $p \times K$ matrix whose kth column is $\beta^{(k)}$ for $k = 1, \ldots, K$. For numbers a and b, we also use the notation $a \vee b = \max\{a, b\}$, $a \wedge b = \min\{a, b\}$ and $x_+ = xI(x > 0)$ for $x \in \mathbb{R}$. For sequences $\{a_n\}$ and $\{\zeta_n\}$, we write $a_n = O(\zeta_n)$ to mean that $a_n \le C\zeta_n$ for a universal constant $C > 0$. Similarly, $a_n = \Omega(\zeta_n)$ when $a_n \ge C'\zeta_n$ for some universal constant $C' > 0$. We summarize notations used in the theorems in Table 1.

Table 1. Notations used in the paper.

Parameters	Definitions
$\lambda =$	A regularization parameter in (2.3)
$d_{\min} =$	$\min_{k\geq 2} \lvert\tau_k - \tau_{k-1}\rvert$
$W_0 =$	$\max_k \lVert w^{(k)}{}_{(T^{(k)})^c}\rVert_\infty \bigvee \max_{k\geq 2} \lVert v^{(k)}{}_{(B^{(k)})^c}\rVert_\infty$
$W_1 =$	$\max_k \lVert w^{(k)}{}_{T^{(k)}}\rVert_\infty \bigvee \max_{k\geq 2} \lVert v^{(k)}{}_{B^{(k)}}\rVert_\infty$
$W_2 =$	$\min_k \min_{j\in\{T^{(k)}\}^c} w_j^{(k)} \bigwedge \min_{k\geq 2} \min_{j\in\{B^{(k)}\}^c} v_j^{(k)}$
$c_0 =$	$(d_{\min}W_1 + 2\lambda(W_0 \vee W_1))/(d_{\min}W_2 - 2\lambda(W_0 \vee W_1))$
$\psi_\lambda =$	$(d_{\min} + 2\lambda)/(d_{\min} - 2\lambda)$
$M_n =$	$\max_i \lVert x_{i,\cup_k T^{(k)}}\rVert_\infty$
$d_0 =$	$\lvert T^{(1)}\rvert + \sum_{k=2}^K \lvert B^{(k)} \setminus T^{(k)}\rvert$
$\mathbb{M}(S) =$	Median of a sequence of real number S

3. Theoretical Properties

We first define a cone constraint: for any set $J \subset \{1, \cdots, p\}$ and any positive number c,

$$C(J, c) = \{x \in \mathbb{R}^p \mid x \neq 0,\ \|x_{J^c}\|_1 \leq c\|x_J\|_1\}.$$

Consider a restricted eigenvalue (RE) condition (Bickel, Ritov and Tsybakov (2009); van de Geer and Bühlmann (2009)): for any integer $0 < s < p$ and any positive number $c > 0$, RE(s, c) means

$$k^2(s, c) := \min_{\substack{J\subseteq\{1,\ldots,p\},\\ |J|\leq s}} \min_{\delta\in C(J,c)} \frac{\delta^T X^T X\delta}{n\|\delta_J\|_2^2} > 0, \tag{3.1}$$

as imposed on the $p \times p$ sample covariance matrix $X^T X/n$. The RE condition is needed to guarantee consistency of the Lasso and Dantzig selectors (Bickel, Ritov and Tsybakov (2009)). This condition also implies that the gram matrix $X^T X/n$ behaves like a positive definite matrix over the cone $C(J, c)$ for any J such that $|J| \leq s$. See Raskutti, Wainwright and Yu (2010) and Rudelson and Zhou (2013) for examples of random design for which the restricted eigenvalue (RE) condition holds in the high-dimensional setting.

Similarly, we introduce a restricted nonlinear impact (RNI) condition, as in Belloni and Chernozhukov (2011): For any integer $0 < s < p$ and any positive number $c > 0$, RNI(s, c) means

$$q(s, c) := \min_{\substack{J\subseteq\{1,\ldots,p\},\\ |J|\leq s}} \min_{\delta\in C(J,c)} \frac{\|X\delta\|_2^3}{n^{1/2}\|X\delta\|_3^3} > 0. \tag{3.2}$$

This controls the norm $\|X\delta\|_3$ by $\|X\delta\|_2$ over the cone $C(J,c)$ for any J such that $|J| \leq s$. RNI(s,c) can be equivalently written as, for $\delta \in C(J,c)$,

$$\left(\frac{1}{n}\sum_{i=1}^{n}|x_i^T\delta|^2\right)^3 \geq q^2(s,c)\left(\frac{1}{n}\sum_{i=1}^{n}|x_i^T\delta|^3\right)^2,$$

which implies that the third sample moment is controlled by the second sample moment. This condition is necessary to control the quantile regression objective function by quadratic terms (Belloni and Chernozhukov (2011)).

Condition 1. [On the conditional density] For each $i = 1, \ldots, n$, let $f_i(\cdot)$ denote the probability density function of y_i given x_i. The function $f_i(\cdot)$ has a continuous derivative $f_i'(\cdot)$. For each i, $f_i(\cdot) \leq \overline{f}$, $|f_i'(\cdot)| \leq \overline{f}$ and $\min_k f_i\left({x_i}^T\beta(\tau_k)\right) \geq \underline{f}$ for some positive numbers $\overline{f}$ and $\underline{f}$.

Condition 2. [On the weights] Let W_0 and W_1 be the maximum weight imposed on the zero components and non-zero components, respectively, and W_2 be the minimum weight imposed on zero components. The weights satisfy

$$\frac{W_2}{W_0 \vee W_1} \geq \frac{2.5\lambda}{\min_k |\tau_k - \tau_{k-1}|}.$$

Condition 3. [On the growth rate of the sparsity] The maximal sparsity s_0 satisfies the growth condition, $s_0 \log p = o(n)$.

Condition 1 is the same as Condition D.1 in Belloni and Chernozhukov (2011). For the location model and the location-scale model, Belloni and Chernozhukov (2011, Lemmas 1 and 2) derive sufficient conditions that guarantee that Condition 1 holds. Condition 2 implies that W_2 must not be too small. In Sections 4 and 5, we demonstrate that W_0, W_1, and W_2 can be constructed from some initial estimates such that W_0 and W_1 are upper bounded and W_2 is lower bounded by some constants.

Remark 1. The regular adaptive lasso weights are used in Jiang, Wang and Bondell (2013) where, for $q > 0$, $w_j^{(k)} = 1/|\tilde{\beta}_j^{(k)}|^q$ and $v_j^{(k)} = 1/|\tilde{\beta}_j^{(k)} - \tilde{\beta}_j^{(k-1)}|^q$ with initial estimates $\tilde{\beta}^{(k)}$ $(k = 1, \cdots, K_n)$. Condition 2 may not be satisfied given these weights because $W_0 \vee W_1$ can be arbitrarily large. This motivates us to use a different type of weights.

3.2. Main results

Throughout this section, for any $\eta \geq 0$, let

$$\mathbb{E}_\eta = \left\{ 0 \le r_k - \frac{1}{n}\sum_{i=1}^{n} \rho_{\tau_k}\left(y_i - {x_i}^T\beta(\tau_k)\right) \le \eta \quad (k = 1, \ldots, K_n) \right\}. \tag{3.3}$$

Theorem 1. *Suppose that Conditions* 1-2, $\mathrm{RE}(2s_0, c_0)$, *and* $\mathrm{RNI}(2s_0, c_0)$ *hold. Let* $\hat{\mathcal{B}} = [\hat\beta^{(1)}, \cdots, \hat\beta^{(K_n)}]$ *be the solution to* (2.3) *and* (2.4). *Let* $\eta_n = o(1)$ *be any sequence of positive numbers with* $0 \le \eta_n < 9\underline{f}^3 q^2(2s_0, c_0)/(32\overline{f}^2)$. *Then, we have with probability at least* $1 - 1/n - \mathbb{P}(\mathbb{E}^c_{\eta_n})$,

$$\max_k \|\hat\beta^{(k)} - \beta(\tau_k)\|_2 \le \xi_1\sqrt{\frac{s_0 \log p}{n} + \eta_n}, \tag{3.4}$$

$$\sum_{k=1}^{K_n} \|\hat\beta^{(k)}_{\{T^{(k)}\}^c}\|_1 \bigvee \lambda \sum_{k=2}^{K_n} \left\| \frac{\{\hat\beta^{(k)} - \hat\beta^{(k-1)}\}_{\{B^{(k)}\}^c}}{|\tau_k - \tau_{k-1}|} \right\|_1 \le \xi_3 K_n\sqrt{s_0}\left(\sqrt{\frac{s_0 \log p}{n}} + \sqrt{\eta_n}\right), \tag{3.5}$$

where for some absolute constant $C_1 > 0$,

$$\xi_1 = \frac{2(1+c_0)^2}{k(2s_0, c_0)\sqrt{\underline{f}}}\left\{1 + \frac{2C_1}{k(s_0, c_0)}\right\} \quad \text{and} \quad \xi_3 = \xi_1\frac{W_1}{W_2}. \tag{3.6}$$

The results in Theorem 1 hold even if the weights $w_j^{(k)}$ and $v_j^{(k)}$ in (2.3) are data-dependent, provided that they satisfy Condition 2. The upper bound in (3.4) implies that the estimates $\hat\beta^{(k)}$ for $k = 1, \cdots K_n$ are uniformly consistent when $\eta_n = o(1)$ and $n = \Omega(s_0 \log p)$. The upper bound in (3.4) has two components, where the first component $\sqrt{s_0 \log p/n}$ is within a factor of $\sqrt{\log p}$ of the oracle rate, and the second component $\sqrt{\eta_n}$ characterizes the bias induced by the use of the feasible region $\mathcal{R}^{(k)}(r_k)$ in (3.3). To obtain the consistency rate $\sqrt{s_0 \log p/n}$ for $\hat\beta^{(k)}$ in (3.4), which is an expected bound for high dimensional models (Belloni and Chernozhukov (2011); Fan, Fan and Barut (2014); Zheng, Peng and He (2015)), $\eta_n = O(s_0 \log p/n)$ is required. By using a consistent initial estimate, we can choose such η_n with r_k, such that the event $\mathbb{E}_{\eta_n}$ holds with a high probability; see (4.6) for details.

As can be seen in (3.4), as η_n increases, the bound on the estimation error is looser while the probability $\mathbb{P}(\mathbb{E}^c_{\eta_n})$ becomes smaller. The optimal r_k is $1/n\sum_{i=1}^n \rho_{\tau_k}\left(y_i - {x_i}^T\beta(\tau_k)\right)$, which provides the best possible rate given (2.3) and (2.4). Using r_k near this optimal value in (2.4) is a key part of implementations. We use a proper initial estimate of $\beta^{(k)}$ to estimate the optimal value r_k.

Inequality (3.5) shows that the ℓ_1 norm of the quantile coefficients estimates for inactive predictors (with true zero coefficients) converge to zero provided

that $W_1/W_2 = o(1)$, $K_n^2 s_0 \eta_n = o(1)$, and $n = \Omega(K_n^2 s_0^2 \log p)$. Moreover, the ℓ_1 norm is decreasing as W_1/W_2 becomes smaller, which implies that choosing smaller weights W_1 and larger weights W_2 would improve the rate of convergence; This is consistent with the idea used in adaptive Lasso (Zou (2006)). Later in Theorem 3, we will discuss exact model selection by using (3.5) with an additional beta–min condition.

Remark 2. Our formulation (2.3) and (2.4) enable us to utilize r_k as a tuning parameter, and the scale of r_k is more interpretable than a tuning parameter in the Lagrangian formulation. Under the fixed p setting, Jiang, Wang and Bondell (2013) set the weights of the quantile loss functions for all quantile levels to be equal in the dual problem, which includes fewer regularization parameters. It is not clear whether model selection consistency holds for such estimators in the high dimensional setting.

4. Implementation

We provide a specific realization for the Dantzig–type joint quantile regression introduced in Section 3. This procedure involves the derivative of the SCAD penalty function (Fan and Li (2001)):

$$P_\zeta(x) = I(x \le \zeta) + \frac{(3.7\zeta - x)_+}{2.7\zeta} I(x > \zeta)$$

with a regularization parameter $\zeta \ge 0$.

Step 1. Obtain initial estimates. (Belloni and Chernozhukov (2011)) Let $\tilde{\lambda} = 1.1\ \Pi(0.9)$ be a regularization parameter, where $\Pi(0.9)$ is defined in Remark 3,

$$\tilde{\beta}^{(k)} = \arg\min_{\beta \in \mathbb{R}^p} \frac{1}{n} \sum_{i=1}^{n} \rho_{\tau_k} \left(y_i - {x_i}^T \beta \right) + \tilde{\lambda} \|\beta\|_1 \quad (k = 1, \ldots, K_n). \tag{4.1}$$

Step 2. Solve the Dantzig-type optimization.

- **Step 2a** For the parameters in (2.3), let $\tilde{s} = \max_k \|\tilde{\beta}^{(k)}\|_0$,

$$\zeta_n = 0.1 \sqrt{\tilde{s} \frac{\log p}{n}}, \tag{4.2}$$

$$w_j^{(k)} = P_{\zeta_n}(|\tilde{\beta}_j^{(k)}|)\ (k = 1, \ldots, K_n), \tag{4.3}$$

$$v_j^{(k)} = P_{\zeta_n}\left(|\tilde{\beta}_j^{(k)} - \tilde{\beta}_j^{(k-1)}| \right)\ (k = 2, \ldots, K_n), \tag{4.4}$$

$$\lambda = 0.4 \min_{k \ge 2} |\tau_k - \tau_{k-1}|. \tag{4.5}$$

- **Step 2b** Let $h > 0$ denote a scaling parameter to be chosen and $\Lambda_k^{(h)} \ge$

0 ($k = 1, \ldots, K_n$) be regularization parameters taken to be $\Lambda_k^{(h)} = \mathbb{M}(R_k)h$, where $\mathbb{M}$ is defined in Table 1 and $R_k = \left\{ |y_i - x_i{}^T \tilde{\beta}^{(k)}| : i = 1, \ldots, n \right\}$. For the parameter r_k in (2.4), use

$$r_k^{(h)} = \frac{1}{n} \sum_{i=1}^{n} \rho_{\tau_k} \left(y_i - x_i{}^T \tilde{\beta}^{(k)} \right) + \Lambda_k^{(h)} \frac{\tilde{s} \log p}{n} \quad (k = 1, \ldots, K_n). \tag{4.6}$$

Step 3. Choose h. Randomly split the data into five roughly equal parts $X^{(1)}$, $\cdots$, $X^{(5)} \in \mathbb{R}^{[n/5] \times p}$ and $y^{(1)}, \cdots, y^{(5)} \in \mathbb{R}^{[n/5] \times 1}$, respectively. For $t = 1, \cdots, 5$, let $X^{(t)} = [x_1^{(t)}, \cdots, x_{[n/5]}^{(t)}]^T$. Let $\hat{\beta}_t^{(k)}(h)$ ($k = 1, \cdots, K_n$) be the solution to the (2.3) and (2.4) following Step 1 and Step 2 for the data X and Y excluding the tth fold. Let the CV score function

$$\text{score}(h) := \sum_{t=1}^{5} \sum_{k=1}^{K_n} \sum_{i=1}^{[n/5]} \rho_{\tau_k} \left(y_i^{(t)} - (x_i^{(t)})^T \hat{\beta}_t^{(k)}(h) \right).$$

Choose h^0 from the set $S := \{0.01, 0.02, \cdots, 4\}$, so $h^o := \arg\min_{h \in S} \text{ score}(h)$.

The Dantzig-type estimate $\hat{\beta}^{(k)}$ is the solution to (2.3) and (2.4) using the aforementioned specifications with $h = h^o$, $\Lambda_k := \Lambda_k^{(h^o)}$, and $r_k := r_k^{(h^0)}$.

In Step 2 (b), $\Lambda_k^{(h)}$ plays the role of scaling to achieve scale equivariance of the method. Those choices of the regularization parameters do not give the best results for any given models, but they lead to good empirical results in a variety of settings and can help us understand how the proposed Dantzig–type penalization performs with reasonable choices of these tuning parameters.

Remark 3. Following Belloni and Chernozhukov (2011), take

$$\Pi := \max_{1 \le k \le K_n} \max_{1 \le j \le p} \frac{1}{n} \left| \sum_{i=1}^{n} \frac{x_{ij} \left(\tau_k - I(u_i \le \tau_k) \right)}{\sqrt{\tau_k (1 - \tau_k)}} \right|,$$

where $u_1, \cdots, u_n$ are independent and identically distributed from the uniform distribution on $(0, 1)$ and independent of x_i's; x_{ij} is the jth component of the design x_i for $i = 1, \ldots, n$ and $j = 1, \ldots, p$. Let $\Pi(0.9)$ be the 0.9th quantile of Π; it can be computed using simulated Π. We use $\tilde{\lambda} = 1.1\ \Pi(0.9)$, where the constant factor 1.1 differs from the recommendation made in Belloni and Chernozhukov (2011), giving us initial estimates with low false negative rates.

5. Theoretical Properties (continued)

Let $\hat{\mathcal{B}} = [\hat{\beta}^{(1)}, \cdots, \hat{\beta}^{(K_n)}]$ be any optimum of (2.3) and (2.4), where $w_j^{(k)}$s, $v_j^{(k)}$s, and r_ks are defined in (4.3), (4.4), and (4.6), respectively. Define an event for the initial estimates $\tilde{\beta}^{(k)}$s for $k = 1, \ldots, K_n$ as follows: for some positive

constants C_2, C_3, and C_4,

$$E_1 = \left\{ \tilde{\lambda} \le C_2 \sqrt{\frac{\log p}{n}},\ \max_k \|\tilde{\beta}^{(k)} - \beta(\tau_k)\|_2 \le C_3 \sqrt{\frac{s_0 \log p}{n}},\ \max_k \|\tilde{\beta}^{(k)}\|_0 \le C_4 s_0 \right\}, \tag{5.1}$$

Denote by $\gamma_n := \mathbb{P}(E_1^c)$ the probability that the event E_1 does not occur.

Belloni and Chernozhukov (2011) prove that their estimators and the corresponding regularization parameters, as stated in (4.1), satisfy condition E_1 with probability close to 1. We need further conditions.

Condition 4. [On the regularization parameters]

$$\min_k \Lambda_k \ge 6\sqrt{C_4 + 1} C_3 \quad \text{and} \quad \zeta_n \ge 2C_3 \sqrt{\frac{s_0 \log p}{n}}.$$

Condition 5. [On the non-zero coefficients] For some positive constants C_5 and C_6,

$$\min_k \min_{j \in T^{(k)}} |\beta_j(\tau_k)| > C_5 \sqrt{\frac{s_0 \log p}{n}}, \tag{5.2}$$

$$\min_{k \ge 2} \min_{j \in B^{(k)}} \frac{|\beta_j(\tau_k) - \beta_j(\tau_{k-1})|}{|\tau_k - \tau_{k-1}|} > C_6 K_n \sqrt{\frac{s_0 \log p}{n}}, \tag{5.3}$$

with $n = \Omega(K_n^2 s_0 \log p)$.

Theorem 2. *Suppose Conditions* 1, 3, 4, $\mathrm{RE}(2s_0, \psi_\lambda)$, *and* $\mathrm{RNI}(2s_0, \psi_\lambda)$ *hold. Then, with probability at least* $1 - 2/n - \gamma_n$, $\hat{\mathcal{B}}$ *satisfies*

$$\max_k \|\hat{\beta}^{(k)} - \beta(\tau_k)\|_2 \le \xi_2 \sqrt{\frac{s_0 \log p}{n}},$$

where for some absolute constant $C > 0$, $\xi_2 = C/k(2s_0, \psi_\lambda) \sqrt{(1 + \max_k \Lambda_k)/\underline{f}}$.

Theorem 3. *If the conditions of Theorem* 2 *and Condition* 5 *hold, then*

$$\mathbb{P}\left(\left\{ \hat{T}^{(k)} = T^{(k)} \text{ and } \hat{B}^{(k)} = B^{(k)} \text{ for all } k \right\} \right) \ge 1 - \frac{2}{n} - \gamma_n.$$

Theorem 2 follows from (3.4) in Theorem 1 and demonstrates that our multi-step Dantzig–type joint quantile estimator $\hat{\mathcal{B}}$ is consistent when $n = \Omega(s_0 \log p)$ under appropriate conditions. Theorem 2 requires the lower bound of Λ_k for the feasible regions (2.4) to include the true parameter $\mathcal{B}^o$ with high probability. In our simulations, the estimator still worked quite well even with Λ_k set to zero. Theorem 3 implies that $\hat{\mathcal{B}}$ recovers the true model structure with high probability, which also satisfies the exact model selection property used in Zhao and Yu (2006), Wainwright (2009), and Fan, Fan and Barut (2014).

Remark 4. The beta-min condition (5.2) imposes a lower bound of the non-zero coefficients. While Condition (5.2) has been studied in high-dimensional analysis to establish the exact model selection property (Meinshausen and Bühlmann (2006)), the beta-min condition (5.3) has not been considered elsewhere.

The beta–min condition (5.3) can be illustrated as follows. Consider equally-spaced quantile levels τ_k $(k = 1, \ldots, K_n)$ with $\tau_k - \tau_{k-1} \asymp 1/K_n$. Consider the location–scale model $y_i = {x_i}^T\beta + {x_i}^T r\epsilon_i$, where the design x_i and the vector $r \in \mathbb{R}^p$ have non-negative components with $x_i^T r > 0$ for all i. Then, (5.3) holds as long as the components of r satisfy $r_j 1\{r_j \neq 0\} \succ K_n\sqrt{s_0 \log p/n}$ $(j = 1, \ldots, p)$, where r_j is the jth component of r.

6. Post–Selection Joint Quantile Regression

We consider a post–selection joint quantile regression that minimizes the sum of quantile loss functions over all quantiles of interest based on the model structure $\hat{T}^{(k)}$ $(k = 1, \ldots, K_n)$ and $\hat{B}^{(k)}$ $(k = 2, \ldots, K_n)$ of the multi-step Dantzig–type joint quantile estimator, as described in Section 4. The post–selection joint quantile estimator (POST JQR) denoted by $\hat{\mathcal{B}}^{po}$ is a minimizer of

$$\min_{\mathcal{B}=[\beta^{(1)},\cdots,\beta^{(K_n)}]\in G} \sum_k \sum_i \rho_{\tau_k}\left(y_i - {x_i}^T\beta^{(k)}\right), \quad \text{where} \tag{6.1}$$

$$G = \left\{\mathcal{B} = [\beta^{(1)}, \ldots, \beta^{(K_n)}] \in \mathbb{R}^{p\times K_n} : {\beta^{(k)}}_{\{\hat{T}^{(k)}\}^c} = 0,\ {\beta^{(k)}}_{\{\hat{B}^{(k)}\}^c} = {\beta^{(k-1)}}_{\{\hat{B}^{(k)}\}^c}\right\}$$

is a set of matrices whose induced model structure is the same as the structure of $\hat{\mathcal{B}}$. Throughout, we assume that $\hat{T}^{(k)} = T^{(k)}$ $(k = 1, \ldots, K_n)$ and $\hat{B}^{(k)} = B^{(k)}$ $(k = 2, \ldots, K_n)$, which holds with probability tending to 1. As can be seen in the proof of Theorem 4 in the Supplementary material, there is a one-to-one mapping T between G and $\mathbb{R}^{d_0}$, where d_0 is the effective dimension of the parameter for the selected model, as defined in Table 1. Thus the set $G \subset \mathbb{R}^{p\times K_n}$ in (6.1) can be embedded in $\mathbb{R}^{d_0}$. We use $T(\hat{\mathcal{B}}^{po})$ to estimate $T(\mathcal{B}^o)$, which is a d_0–dimensional vector that consists of the active components of $\mathcal{B}^o$.

To establish the theoretical properties of $T(\hat{\mathcal{B}}^{po})$, we redefine POST JQR. As given in the proof of Theorem 4 in the Supplementary material, there exist new design variables $z_i^{(k)}$ $(i = 1, \ldots, n;\ k = 1, \ldots, K_n)$ such that

$$T(\hat{\mathcal{B}}^{po}) = \arg\min_{\beta\in\mathbb{R}^{d_0}} \sum_k \sum_i \rho_{\tau_k}\left(y_i - (z_i^{(k)})^T\beta\right). \tag{6.2}$$

To establish the asymptotic convergence rate and asymptotic normality of $T(\hat{\mathcal{B}}^{po})$, we use a sparse eigenvalue condition: for $0 < s < p$,

$$\text{Sparse}(s):\ \phi(s) = \max_{\|\delta\|_0 \le s} \frac{\|X\delta\|_2^2}{n\|\delta\|_2^2} < \infty. \tag{6.3}$$

Sparse(s) means that the maximal s-sparse eigenvalue of the gram matrix X^TX/n is bounded by some constant (Rudelson and Zhou (2013); Belloni, Chernozhukov and Kato (2015); Zheng, Peng and He (2015)). We need further conditions.

Condition 6(a). [On the sample size]

$$n = \Omega\left(d_0 s_0^3 (\log n)^6 \vee M_n^4 d_0 (\log n)^2\right).$$

Condition 6(b). $n = \Omega\left(d_0^5 s_0^3 (\log n)^6 \vee M_n^2 d_0^3 s_0\right)$.

These conditions involve d_0, s_0, M_n, and n. If the entries in x_i are uniformly bounded, and d_0 and s_0 increase slowly with n, then Conditions 6(a) and 6(b) are quite mild. The POST JQR exhibits an asymptotic oracle consistency rate as follows.

Theorem 4. *If the conditions of Theorem* 3*, Condition* 6(a)*, and* Sparse(s_0) *hold, then*

$$\|T(\hat{\mathcal{B}}^{po}) - T(\mathcal{B}^o)\|_2 = O_p\left(\frac{\sqrt{d_0}}{n}\right). \tag{6.4}$$

Theorem 5. *If the conditions of Theorem* 4 *and Condition* 6(b) *hold, then, for any sequence of vectors* $\alpha_n \in R^{d_0}$ *with* $\|\alpha_n\|_2 = 1$,

$$\alpha_n^T \sqrt{n}({A_n}^{-1} B_n {A_n}^{-1})^{-1/2}\left(T(\hat{\mathcal{B}}^{po}) - T(\mathcal{B}^o)\right) \to^d N(0,1),$$

where

$$A_n = \sum_{k=1}^{K_n}\sum_{i=1}^{n} \frac{1}{n} f_i\left({x_i}^T \beta(\tau_k)\right) z_i^{(k)}\left(z_i^{(k)}\right)^T,$$

$$B_n = \sum_{i=1}^{n}\sum_{k,k'=1,\ldots,K_n} \frac{1}{n} z_i^{(k)}\left(z_i^{(k')}\right)^T (\tau_k \wedge \tau_{k'} - \tau_k\tau_{k'}).$$

The exact model selection, as defined in Theorem 3, is typically fragile without a beta-min condition, and is not uniformly valid (Leeb and Pötscher (2005)). Leeb and Pötscher (2003) and Belloni, Chernozhukov and Kato (2015) consider the post-model-selection estimator conditional on selecting an incorrect model, and establish a uniform asymptotic distribution of the estimator. Establishing an asymptotic distribution without a beta-min condition in our setting is of interest in follow-up work.

7. Numerical Studies

7.1. Experiments and setup

Solving (2.3) is equivalent to a linear programming problem with the assistance of slack variables, and can be done by existing optimization packages in a way that is similar to the problem of Jiang, Wang and Bondell (2013). For the other estimators, we use $\tilde{\beta}^{(k)}$ as an initial estimate at the τ_k-th quantile. More specifically, ALasso at τ_k is

$$\arg\min_{\beta\in\mathbb{R}^p} \frac{1}{n}\sum_{i=1}^{n} \rho_{\tau_k}(y_i - {x_i}^T\beta) + \lambda_{ad,k}\sum_{j=1}^{p}\frac{|\beta_j|}{|\tilde{\beta}_j^{(k)}|},$$

where $\lambda_{ad,k}$ is the regularization parameter chosen by five-fold cross validation to minimize the τ_k-th quantile loss function, and FAL (Jiang, Wang and Bondell (2013)) uses five-fold cross validation to choose the tuning parameter, as well. Our proposed estimator Dantzig is described in Section 4.

To assess the performances of the methods, the following performance measures were calculated based on 100 Monte Carlo replications:

1. "FP_k", the number of false positives in the selected model at τ_k;
2. "FN_k", the number of false negatives in the selected model at τ_k;
3. "SD_k", the size of set differences of the selected models for adjacent quantile levels, τ_k and τ_{k-1}, i.e., $|\hat{T}^{(k)}\triangle\hat{T}^{(k-1)}|$ for $k = 2,\ldots,K_n$;
4. "FP_U", the number of false positives in the union of the selected models across all quantile levels, i.e., $|\cup_k \hat{T}^{(k)}\setminus\cup_k T^{(k)}|$;
5. "FN_U", the number of false negatives in the union of the selected models across all quantile levels, i.e., $|\cup_k T^{(k)}\setminus\cup_k \hat{T}^{(k)}|$.

7.2. Simulation results

We considered a location model, a location-scale model, and a random coefficient model.

Example 1. Consider the linear regression model with $(n, p, K_n, s_0) = (100, 500, 5, 6)$ and $(\tau_1,\tau_2,\tau_3,\tau_4,\tau_5) = (0.30, 0.40, 0.50, 0.60, 0.70)$:

$$y_i = {x_i}^T\beta + \epsilon_i, \quad \beta = (1.0, 0.8, 0.0, 0.9, 0.5, 0.0, 0.3, 0.7, 0.0, \cdots, 0.0)^T,$$

where the ϵ_i's are independent and identically distributed standard normals and independent of x_i's. The regressors are $x_i = (1, z_i)^T$, where $z_{ij} \sim N(0,\Sigma)$ is generated with $\Sigma_{(i,j)} = 0.5^{|i-j|}$.

Example 2. Consider the location-scale model with $(n, p, K_n, s_0) = (100, 500, 5,$

7) and $(\tau_1, \tau_2, \tau_3, \tau_4, \tau_5) = (0.30, 0.40, 0.50, 0.60, 0.70)$:

$$y_i = x_1 + 0.8x_2 + 0.9x_4 + 0.5x_5 + 0.3x_7 + 0.75x_8 + (0.5x_2 + x_3 + 0.5x_8)\epsilon_i,$$

where the ϵ_i's are independent and identically distributed standard normals and independent of x_i's. The regressors were generated in two steps, following Wang, Wu and Li (2012): generate $\tilde{x}_{ij} \sim N(0, \Sigma_x)$ from the AR(1) model, with correlation 0.5, and take $x_{ij} = \Phi(\tilde{x}_{ij})$ $(j = 2, 3, 8)$ and $x_{ij} = \tilde{x}_{ij}$ $(j \neq 2, 3, 8)$, where Φ is the cumulative distribution function of the standard normal.

Example 3. Consider the random coefficient model with $(n, p, K_n, s_0) = (100, 500, 5, 6)$ and $(\tau_1, \tau_2, \tau_3, \tau_4, \tau_5) = (0.70, 0.75, 0.80, 0.85, 0.90)$:

$$y_i = {x_i}^T \beta(u_i), \quad \beta(u_i) = (\beta_1(u_i), \cdots, \beta_p(u_i))^T,$$

where $u_1, \cdots, u_n$ are independent and identically distributed from the uniform distribution on $(0, 1)$ and independent of x_i, and $\beta_1(u) = 1.7 + \Phi^{-1}(u)$, $\beta_2(u) = 0.35$, $\beta_3(u) = 3(u - 0.8)_+$, $\beta_5(u) = 0.5 + 0.5 \times 2^u$, $\beta_6(u) = 0.5 + u$, $\beta_{10}(u) = 0.4 + \sqrt{u}$, and $\beta_j(u) = 0$ $(j \neq 1, 2, 3, 5, 6, 10)$. The regressors were generated as in Example 2.

Figure 1 gives the performance measures of Section 7.1 for Examples 1–3. Across all figures, the largest standard errors for the false positives, the false negatives, and the size of set differences are less than 0.9, 0.1, and 0.5, respectively. As seen in Figure 1, Dantzig includes a smaller number of false positives with more false negatives compared to the other methods. This increase in false negatives is relatively small considering the decrease in false positives.

Dantzig has a smaller size of set difference for two neighboring quantiles, and fewer false positives than other methods for the union of the selected variables across the five quantile levels. This suggests that Dantzig shares many common variables across different quantiles, and provides more stable models. Overall, at each quantile, Dantzig provides a sparser model than other competitors in each of the examples, and outperforms the other methods in terms of the stability of the selected models across quantiles.

8. An Application

We applied the proposed Dantzig–type joint quantile regression method to a genetic data set used in Scheetz et al. (2006). This data set consists of the expression values of 31, 042 probe sets for 120 rats. As in Huang, Ma and Zhang (2008), Kim, Choi and Oh (2008), and Wang, Wu and Li (2012), we are interested in finding genes that are related to gene TRIM32, which is known to cause

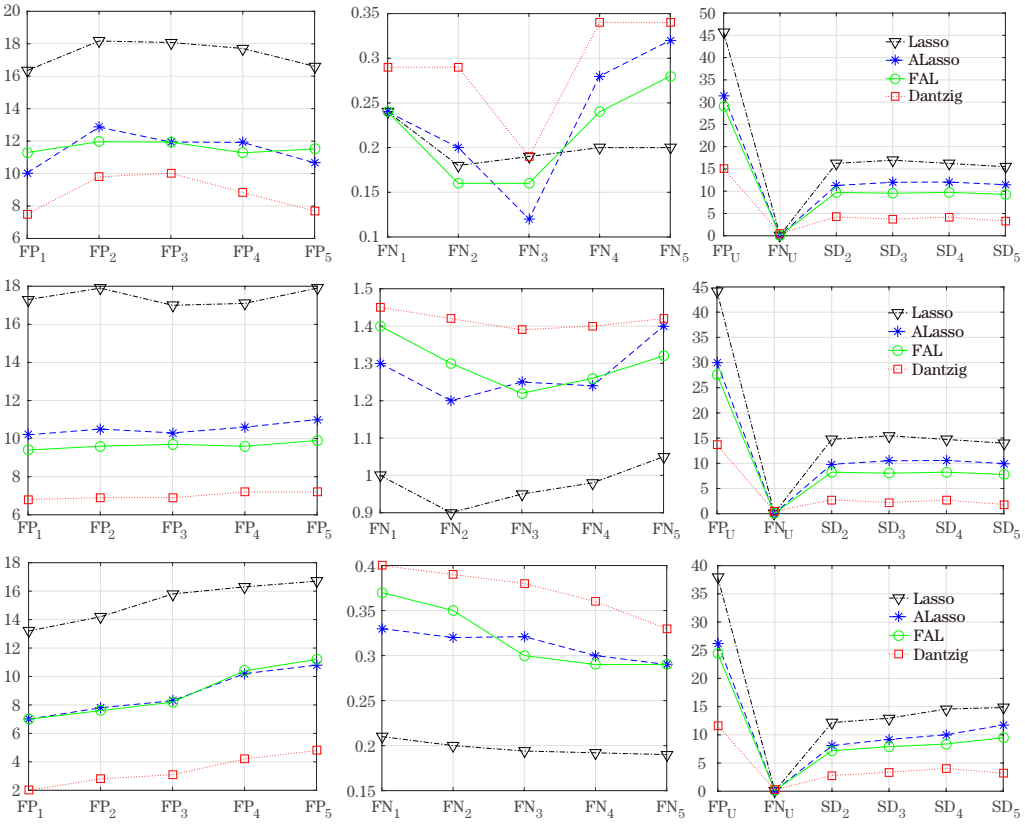

Figure 1. Results for Example 1 (top), 2 (middle) and 3 (below): Each plot shows the false positives (left), the false negatives (middle), and the stability measures (right). Four competing procedures are evaluated: Lasso, ALasso, FAL, and Dantzig.

Bardet-Biedl syndrome.

The model selection approach was applied to 300 probe sets that pass an initial screening (see Huang, Ma and Zhang (2008) for details of the screening steps). We applied Dantzig, Lasso, ALasso, FAL, and SCAD (Wang, Wu and Li (2012)) on these 300 probe sets ($p = 300$) with 120 rats ($n = 120$). SCAD is a single quantile regression method that uses the SCAD penalty function to penalize quantile coefficients. We considered two sets of five quantile levels $(\tau_1, \tau_2, \tau_3, \tau_4, \tau_5)$ as $(0.48, 0.49, 0.50, 0.51, 0.52)$ and $(0.81, 0.82, 0.83, 0.84, 0.85)$, representing interests in the middle and the upper tail of the distribution of the target gene expressions, respectively. To select the tuning parameter for each method, we employed five-fold cross validation (see Sections 4 and 7.1 for details).

We report the number of non-zero coefficients ("SIZ") selected by each

Table 2. Performance results of the whole dataset.

Method	SIZ	DIF	TOT	Method	SIZ	DIF	TOT
Lasso (0.48)	37			Lasso (0.81)	37		
Lasso (0.49)	38	9		Lasso (0.82)	41	8	
Lasso (0.50)	35	15		Lasso (0.83)	39	8	
Lasso (0.51)	36	5		Lasso(0.84)	36	7	
Lasso (0.52)	37	3	45	Lasso (0.85)	38	4	49
SCAD (0.48)	24			SCAD (0.81)	20		
SCAD (0.49)	24	0		SCAD (0.82)	25	9	
SCAD (0.50)	20	6		SCAD (0.83)	16	9	
SCAD (0.51)	14	7		SCAD (0.84)	29	13	
SCAD (0.52)	18	5	25	SCAD (0.85)	27	4	35
ALasso (0.48)	27			ALasso (0.81)	25		
ALasso (0.49)	17	14		ALasso (0.82)	24	3	
ALasso (0.50)	20	7		ALasso (0.83)	22	4	
ALasso (0.51)	14	6		ALasso (0.84)	21	3	
ALasso (0.52)	15	1	29	ALasso (0.85)	28	7	34
FAL (0.48)	21			FAL (0.81)	25		
FAL (0.49)	21	0		FAL (0.82)	25	1	
FAL (0.50)	22	2		FAL (0.83)	26	2	
FAL (0.51)	21	3		FAL (0.84)	25	2	
FAL (0.52)	21	2	25	FAL (0.85)	25	2	26
Dantzig (0.48)	21			Dantzig (0.81)	21		
Dantzig (0.49)	19	2		Dantzig (0.82)	20	1	
Dantzig (0.50)	20	1		Dantzig (0.83)	21	1	
Dantzig (0.51)	21	3		Dantzig (0.84)	22	2	
Dantzig (0.52)	20	1	22	Dantzig (0.85)	21	1	24

method at each quantile level. The size of set differences of the selected models at adjacent quantile levels ("DIF") and the size of the union of the selected covariates over five quantile levels ("TOT") were considered to determine the stability of the selected models. As can be seen in Table 2, Dantzig consistently provides a sparser model than the other methods, and affords the most stable model.

We randomly divided the data set into a training set and a test set with the training set including 80 rats and the test set including 40 rats. We estimated the models with each method using the training set, and recorded "SIZ", "DIF", and "TOT". The prediction error ("PRE") was calculated over the test set as the quantile loss for each quantile level τ_k. We repeated this random experiment 100 times and report the average value of "SIZ", "DIF", "TOT", and "PRE" over the 100 repetitions for each method in Table 3. As seen in Table 3, all of

Table 3. Performance results of 100 random partitions of the data.

Method	SIZ	DIF	PRE	TOT	Method	SIZ	DIF	PRE	TOT
Lasso (0.48)	30.94		1.79		Lasso (0.81)	32.94		1.33	
Lasso (0.49)	31.10	3.35	1.79		Lasso (0.82)	33.04	4.22	1.30	
Lasso (0.50)	31.76	4.38	1.78		Lasso (0.83)	33.00	6.36	1.26	
Lasso (0.51)	31.92	4.66	1.78		Lasso (0.84)	32.88	4.20	1.23	
Lasso (0.52)	32.60	4.78	1.78	37.73	Lasso (0.85)	32.78	4.34	1.21	40.28
SCAD (0.48)	22.04		1.79		SCAD (0.81)	20.90		1.32	
SCAD (0.49)	22.82	5.10	1.78		SCAD (0.82)	20.32	6.46	1.27	
SCAD (0.50)	21.86	5.44	1.78		SCAD (0.83)	21.02	6.62	1.27	
SCAD (0.51)	21.38	4.52	1.78		SCAD (0.84)	22.10	6.12	1.23	
SCAD (0.52)	21.66	5.40	1.79	28.74	SCAD (0.85)	20.50	5.16	1.20	28.86
ALasso (0.48)	19.96		1.82		ALasso (0.81)	19.98		1.34	
ALasso (0.49)	19.70	2.98	1.79		ALasso (0.82)	19.22	4.04	1.31	
ALasso (0.50)	19.32	3.46	1.80		ALasso (0.83)	20.04	5.34	1.26	
ALasso (0.51)	19.08	3.40	1.80		ALasso (0.84)	19.92	3.36	1.25	
ALasso (0.52)	19.64	3.76	1.80	24.56	ALasso (0.85)	19.44	3.60	1.21	25.78
FAL (0.48)	19.75		1.85		FAL (0.81)	20.95		1.37	
FAL (0.49)	20.70	1.28	1.82		FAL (0.82)	21.71	2.34	1.32	
FAL (0.50)	20.72	1.94	1.88		FAL (0.83)	20.33	3.90	1.27	
FAL (0.51)	20.18	2.40	1.82		FAL (0.84)	20.18	2.76	1.25	
FAL (0.52)	19.94	2.59	1.83	23.63	FAL (0.85)	21.74	2.20	1.22	24.55
Dantzig (0.48)	20.20		1.84		Dantzig (0.81)	21.94		1.33	
Dantzig (0.49)	20.06	0.98	1.84		Dantzig (0.82)	21.72	1.02	1.31	
Dantzig (0.50)	19.98	1.82	1.82		Dantzig (0.83)	21.98	2.70	1.27	
Dantzig (0.51)	20.70	2.01	1.81		Dantzig (0.84)	21.60	1.78	1.25	
Dantzig (0.52)	21.02	2.52	1.80	22.90	Dantzig (0.85)	21.42	1.18	1.22	23.86

the five methods are similar in terms of prediction error. Regarding the sparsity of the selected models, all of the methods except Lasso are similar. In terms of the stability of the models, Dantzig outperforms other competitors. In Table 3, the largest standard errors for the columns corresponding to SIZ, DIF, PRE, and TOT are less than 0.7, 0.3, 0.05, and 1.2, respectively.

9. Conclusion

Model selection stability across quantile levels adds credibility and interpretability of the selected models in applications. Selecting models that vary significantly from one quantile to the next when the quantile levels used are very close to each other is undesirable. The proposed Dantzig–type approach is a more stable selection without a noticeable sacrifice in prediction error. A simulation study and data analysis demonstrate that the proposed method consistently

provides sparse and stable models, while reducing the noisy component in model selection at single quantile levels for both homogeneous and heterogeneous cases.

Supplementary Materials

Proofs and additional simulation results can be found in the Supplementary material.

Acknowledgment

The authors thank the Editor, an associate editor, and two referees for their valuable comments and suggestions, which led to an improved presentation. The authors also express gratitude to Timothy Johnson, Kristjan Greenewald, and Alexandre Belloni for helpful discussions. The research is partially supported by NSF Grants DMS-1307566 and DMS-1316731, and the Elizabeth Caroline Crosby Research Award from the Advance Program at the University of Michigan.

References

Bang, S. and Jhun, M. (2012). Simultaneous estimation and factor selection in quantile regression via adaptive sup-norm regularization. *Computational Statistics and Data Analysis* **56**, 813–826.

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

Belloni, A., Chernozhukov, V. and Kato, K. (2015). Uniform post-selection inference for least absolute deviation regression and other Z-estimation problems. *Biometrika* **102**, 77–94.

Bickel, P. J., Ritov, Y. and Tsybakov, A. B. (2009). Simultaneous analysis of Lasso and Dantzig selector. *Annals of Statistics* **37**, 1705–1732.

Bondell, H. D., Reich, B. J. and Wang, H. (2010). Non-crossing quantile regression curve estimation. *Biometrika* **97**, 825–838.

Candes, E. and Tao, T. (2007). The Dantzig selector: Statistical estimation when p is much larger than n. *Annals of Statistics* **35**, 2313–2351.

Dette, H. and Volgushev, S. (2008). Non-crossing non-parametric estimates of quantile curves. *Journal of the Royal Statistical Society B* **70**, 609–627

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.

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

Fan, J., Xue, L. and Zou, H. (2014). Strong oracle optimality of folded concave penalized estimation. *Annals of Statistics* **42**, 819–849.

He, X. (1997). Quantile curves without crossing. *The American Statistician* **51**, 186–192.

Huang, J., Ma, S. and Zhang, C. H. (2008). Adaptive lasso for sparse high-dimensional regression

models. *Statist. Sinica* **18**, 1603–1618.

Jang, W. and Wang, H. (2015). A semiparametric bayesian approach for joint-quantile regression with clustered data. *Journal of Computational and Graphical Statistics* **84**, 99–115.

Jiang, L., Wang, H. and Bondell, H. (2013). Interquantile shrinkage in regression models. *Journal of Computational and Graphical Statistics* **69**, 208–219.

Kim, Y., Choi, H. and Oh, H. S. (2008). Smoothly clipped absolute deviation on high dimensions. *Journal of the American Statistical Association* **103**, 1665–1673.

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

Koenker, R. (2005). *Quantile Regression*. Cambridge Univ. Press, Cambridge.

Ledoux, M. and Talagrand, M. (1991). *Probability in Banach Spaces: Isoperimetry and Processes*. Springer, Berlin.

Leeb, H. and Pötscher, B. M. (2003). The finite-sample distribution of post-model-selection estimators and uniform versus nonuniform approximations. *Economic Theory* **19**, 100–142.

Leeb, H. and Pötscher, B. M. (2005). Model selection and inference: Facts and fiction. *Economic Theory* **21**, 21–59.

Li, Y. and Zhu, J. (2008). l_1-norm quantile regression. *Journal of Computational and Graphical Statistics* **17**, 163–185.

Meinshausen, N. and Bühlmann, P. (2006). High-dimensional graphs and variable selection with the Lasso *Annals of Statistics* **3**, 1436–1462.

Meinshausen, N. and Yu, B. (2009). Lasso-type recovery of sparse representations for highdimensional data. *Annals of Statistics* **37**, 246–270.

Peng, L., Xu, J. and Kutner, N. (2014). Shrinkage estimation of varying covariate effects based on quantile regression. *Statistics and Computing* **24**, 853–869.

Raskutti, G., Wainwright, M. and Yu, B. (2010). Restricted eigenvalue properties for correlated gaussian designs. *Journal of Machine Learning Research* **11**, 2241–2259.

Rudelson, M. and Zhou, S. (2013). Reconstruction from anisotropic random measurements. *IEEE Transactions on Information Theory* **59**, 3434–3447.

Rudin, L., Osher, S. and Fatemi, E. (1992). Nonlinear total variation based noise removal algorithms. *Physica D* **60**, 259-268.

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

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 B* **67**, 91–108.

van de Geer, S. and Bühlmann, P. (2009). On the conditions used to prove oracle results for the lasso. *Electronic Journal of Statistics* **3**, 1360–1392.

Volgushev, S., Wagener, J. and Dette, H. (2014). Censored quantile regression processes under dependence and penalization. *Electronic Journal of Statistics* **8**, 2405–2447.

Wainwright, M. (2009). Sharp thresholds for high-dimensional and noisy sparsity recovery using ℓ_1–constrained quadratic programming. *IEEE Trans. Inform. Theory* **55**, 2183–2202.

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

Wang, L. (2013). l_1 penalized LAD estimator for high dimensional linear regression. *Journal of Multivariate Analysis* **120**, 135–151.

Wu, Y. and Liu, Y. (2009). Variable selection in quantile regression. *Statist. Sinica* **19**, 801–817.

Zhang, C. H. and Huang, J. (2008). The sparsity and bias of the LASSO selection in high-dimensional linear regression. *Annals of Statistics* **36**, 1567–1594.

Zhao, P. and Yu, B. (2006). On model selection consistency of Lasso. *Journal of Machine Learning Research* **7**, 2541–2567.

Zheng, Q., Gallagher, C. and Kulasekera, K. B. (2013). Adaptive penalized quantile regression for high-dimensional data. *Journal of Computational and Graphical Statistics* **143**, 1029–1038.

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

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). Regularized simultaneous model selection in multiple quantiles regression. *Journal of Computational and Graphical Statistics* **52**, 5296–5304.

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

Department of Biostatistics, School of Public Health, Yale, New Haven, CT 06520-8034, USA

E-mail: seyoung.park@yale.edu

Department of Statistics, University of Michigan, Ann Arbor, MI 48109-1107, USA

E-mail: xmhe@umich.edu

Department of Statistics, University of Michigan, Ann Arbor, MI 48109-1107, USA

E-mail: shuhengz@umich.edu

(Received February 2016; accepted August 2016)