Communications of the Korean Statistical Society Vol. 15, No. 5, 2008, pp. 753–764

The Doubly Regularized Quantile Regression

Hosik Choi[1)], Yongdai Kim[2)]

Abstract

The L_1 regularized estimator in quantile problems conduct parameter estimation and model selection simultaneously and have been shown to enjoy nice performance. However, L_1 regularized estimator has a drawback: when there are several highly correlated variables, it tends to pick only a few of them. To make up for it, the proposed method adopts doubly regularized framework with the mixture of L_1 and L_2 norms. As a result, the proposed method can select significant variables and encourage the highly correlated variables to be selected together. One of the most appealing features of the new algorithm is to construct the entire solution path of doubly regularized quantile estimator. From simulations and real data analysis, we investigate its performance.

Keywords: Quantile regression; regularization; LASSO; elastic net.

1. Introduction

Suppose that $\{(\mathbf{x}_i, y_i)_{i=1}^n | \mathbf{x}_i \in \mathbb{R}^p, y_i \in \mathbb{R}\}$ is a set of independent observations from a distribution. Let $F(y \,|\, \mathbf{x})$ be the conditional distribution function of y given $\mathbf{x}$. Then for any $\tau \in (0,1)$, the τ^{th} quantile of y given $\mathbf{x}$ is defined as follows:

$$f(\mathbf{x};\tau) = F^{-1}(\tau \,|\, \mathbf{x}) = \inf\{y : F(y \,|\, \mathbf{x}) \geq \tau\}. \tag{1.1}$$

The objective of quantile regression is to estimate the τ^{th} conditional quantile defined by (1.1). A popular way of estimating of conditional quantiles is to use the check function, $\rho_\tau(z) = \tau z I_{[0,\infty)}(z) - (1-\tau) z I_{(-\infty,0)}(z)$ (See, Figure 1.1). Under linear model, $f(\mathbf{x};\tau) = \beta_0(\tau) + \mathbf{x}'\beta(\tau)$, Koenker and Bassett (1978) proposed the method minimizing

$$\sum_{i=1}^{n} \rho_\tau\big[y_i - \{\beta_0(\tau) + \mathbf{x}_i'\beta(\tau)\}\big] \tag{1.2}$$

and solved the problem by linear programming. The regularized quantile estimators of it are investigated by several researchers via minimizing

$$\sum_{i=1}^{n} \rho_\tau\big(y_i - f(\mathbf{x};\tau)\big) + \lambda \|f\|_{\mathcal{H}}^2 \tag{1.3}$$

1) Full Time Lecturer, Department of Informational Statistics and Institute of Basic Science, Hoseo University, Asan Campus, San 165, Sechul-ri, Baebang-myun, Asan, Chungnam 336-795, Korea. Correspondence: choi.hosik@gmail.com.

2) Associate Professor, Department of Statistics, Seoul National University, San 56-1, Sillim-dong, Gwanak-gu, Seoul 151-742, Korea. E-mail: ydkim0903@gmail.com

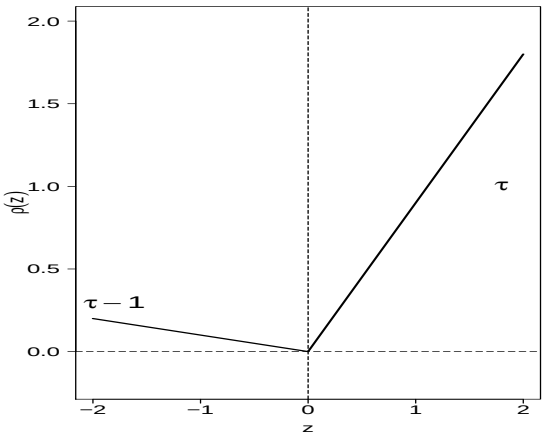


Figure 1.1: The check function is $\rho_\tau(z) = \tau z I_{[0,\infty)}(z) - (1-\tau) z I_{(-\infty,0)}(z)$, where $\tau = 0.75$

for the general function $f \in \mathcal{H}$ where $\mathcal{H}$ is a Reproducing Kernel Hilbert Space(RKHS). Here, $\lambda > 0$ controls the balance between the fidelity and the smoothness or complexity of f (Koenker *et al.*, 1994; Li *et al.*, 2007; Yuan, 2006; Oh *et al.*, 2004). Under such a framework, Tibshirani laid down important groundwork on variable selection so called Least Absolute Shrinkage and Selection Operator(LASSO). Recently, Li and Zhu (2008) proposed L_1 regularized quantile regression estimator.

For the simplicity of notations, given τ, we use $f(\mathbf{x})$, β_0 and β instead of $f(\mathbf{x};\tau)$, $\beta_0(\tau)$ and $\beta(\tau)$ respectively. Then the L_1-norm penalized linear quantile regression estimator is defined by

$$\arg \min_{\beta_0, \beta \in \mathbb{R}^p} \sum_{i=1}^{n} \rho_\tau\big(y_i - (\beta_0 + \mathbf{x}_i'\beta)\big) + \lambda \sum_{j=1}^{p} |\beta_j|. \tag{1.4}$$

The L_1-norm penalty shrinks to fitted coefficients toward zero and benefits from the reduction of the variance of fitted coefficients. But, it is known that the LASSO estimator with the L_1 norm penalty suffers from three limitations (Zou and Hastie, 2005):

1. In the $p > n$ case, the LASSO select at most n variables.

2. For example, in microarray experiments the highly correlated genes are called homogeneous functional genes, showing similar patterns along time or samples. Biologists want to treat homogeneous functional genes simultaneously and want to include them together in a prediction model, because they are expected to have similar functions. If we apply LASSO for analyzing microarray data, it tends to select only one gene among homogeneous functional genes(variables). This is mainly because several components induced from homogeneous functional genes are put into competition in order to be retained in the final prediction model (Choi, 2007).

3. For usual $n > p$ situations, if there are high correlations between predictors, it

has been empirically observed that the prediction performance of the LASSO is dominated by ridge regression (Tibshirani, 1996).

To overcome limitations of the L_1-norm estimator, Zou and Hastie (2005) proposed the *elastic net* using the mixture of L_1 and L_2 norm penalties. As a classification setting, Wang *et al.* (2006) developed the doubly regularized Support Vector Machine(DrSVM). In this paper, our objective is to extend such a doubly regularized framework to quantile regression.

The paper is organized as follows. In Section 2, the doubly regularized quantile regression estimator is proposed and the optimization algorithm is given. Numerical results are presented in Section 3. Concluding remarks follow in Section 4.

2. The Doubly Regularized Quantile Regression

2.1. Method

Given the check loss function ρ_τ and L_1 and L_2 norm penalties, we are to minimize the following regularized cost functional

$$\sum_{i=1}^{n} \rho_\tau\big(y_i - (\beta_0 + \mathbf{x}_i'\beta)\big) + \frac{\lambda_2}{2}\sum_{j=1}^{p}\beta_j^2 + \lambda_1\sum_{j=1}^{p}|\beta_j|, \tag{2.1}$$

where $\rho_\tau(z_i) = [\tau z_i I_{[0,\infty)}(z_i) - (1-\tau)z_i I_{(-\infty,0)}(z_i)]$. When λ_1 equals to zero, then (2.1) reduces to ridge-type quantile estimator and when λ_2 equals to zero, then (2.1) reduces to the LASSO-type quantile estimator. Thus, we call the estimator minimizing (2.1) as **doubly regularized quantile regression estimator(DrQR)**. For a comparison, denote the L_1-norm quantile regression estimator by L1QR.

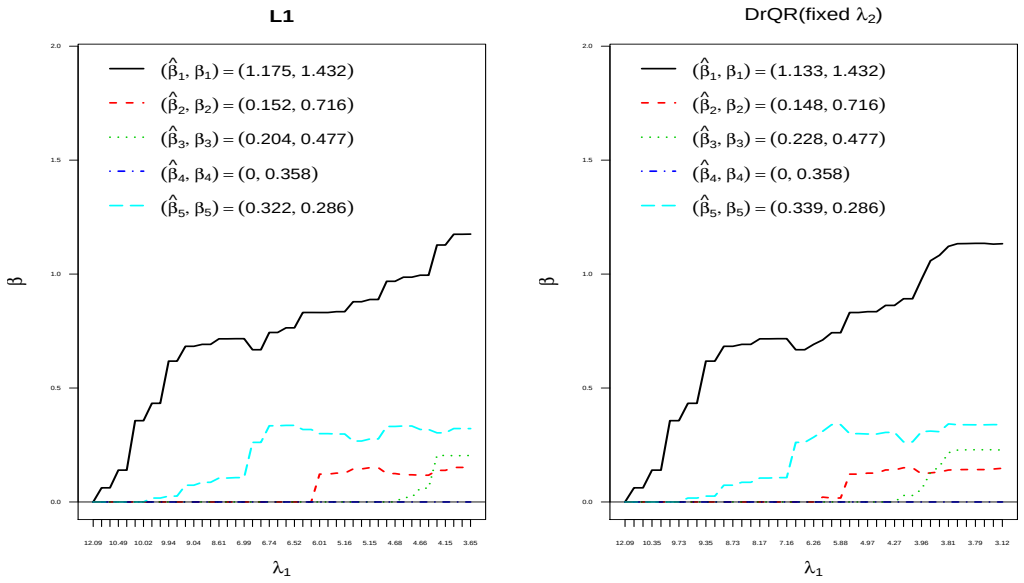
For illustration, we give a simple example which provides the path of solutions for simulated data. Figure 2.1 draws the entire piecewise linear solution paths of the DrQR and L1QR estimates of the first 5 regression coefficients from the simulated model (3.1) with $q = 5(p = 10)$, $r = 0$. Let $\beta(\lambda_1)$ be the solution given λ_1. x-axis indicates λ_1 values and y-axis indicates the path of $\beta(\lambda_1)$. The solution paths of 2.1(b) have much shrinkage effects than those of 2.1(a) because of large λ_2. Note that the rightmost 5 values of Figure 2.1(a) are the solution of (1.2) without the regularization.

Main advantage of path-following algorithms (See, Efron *et al.*, 2004) is to save computational cost in selecting the appropriate regularization parameter λ_1 because only one fitting procedure can provide the entire solution paths according to regularization parameters.

Now, before specific explanation of the proposed method, we note the grouping effect of the proposed method via following theorem. Following the spirit of Wang *et al.* (2006) and Theorem 2.1 is simply verified in the quantile regression setting.

Theorem 2.1 If $\rho_\tau(t), t \in \mathbb{R}$ is Lipschitz continuous, *i.e.*, $|\rho_\tau(t_1) - \rho_\tau(t_2)| \le M|t_1 - t_2|$, then for any pair $(j,\ l)$ and some positive M, we have

$$|\beta_j - \beta_l| \le \frac{M}{\lambda_2}\|x_j - x_l\|_1 = \frac{M}{\lambda_2}\sum_{i=1}^{n}|x_{ij} - x_{il}|, \tag{2.2}$$



(a) When $\lambda_2 = 0$ (b) When $\lambda_2 = 1$

Figure 2.1: The entire solution paths of L1QR and DrQR methods

where the covariate vector x_j and $x_l \in \mathbb{R}^p$. Furthermore, if the covariate vector x_j and x_l are standardized, then

$$|\beta_j - \beta_l| \leq \frac{\sqrt{n}M}{\lambda_2}\sqrt{2(1-r)}, \tag{2.3}$$

where r is the sample correlation between covariate vectors x_j and x_l.

Note that if we set M to $\max(\tau, 1-\tau)$, then $\rho_\tau(\cdot)$ satisfies the Lipschitz continuity condition. From the upper bound of (2.3), as the corresponding variables are more highly correlated variables, their estimates become closer.

2.2. DrQR algorithm and its computation

In this section, we present the proposed method. Let the covariate set with nonzero coefficient be $\mathcal{V} = \{j : \beta_j \neq 0\}$, $f(\mathbf{x}_i) = \beta_0 + \sum_{j=1}^p \beta_j x_{ij}$, $i = 1, \ldots, n$ and $f_i = f(\mathbf{x}_i)$. And we divide samples into three sets, left elbow set $\mathcal{L} = \{i : y_i - f_i < 0\}$, elbow set $\mathcal{E} = \{i : y_i - f_i = 0\}$ and right elbow set $\mathcal{R} = \{i : y_i - f_i > 0\}$.

For a computational convenience, we transfer the problem (2.1) to constraint form:

$$\sum_{i=1}^n \rho_\tau\big(y_i - (\beta_0 + \mathbf{x}_i'\beta)\big) + \frac{\lambda_2}{2}\sum_{j=1}^p \beta_j^2, \tag{2.4}$$

subject to $\sum_{j=1}^{p} |\beta_j| \leq s$ for some positive s. Then for given $s(>0)$, the primal problem for (2.4) is

$$\tau \sum_{i=1}^{n} \xi_i + (1-\tau) \sum_{i=1}^{n} \zeta_i,$$

subject to $\sum_{j=1}^{p} |\beta_j| \leq s$ and $-\zeta_i \leq y_i - f_i \leq \xi_i,\ \zeta_i \geq 0,\ \xi_i \geq 0$. Then the primal Lagrangian function is

$$\begin{aligned} L = \tau \sum_{i=1}^{n} \xi_i + (1-\tau) \sum_{i=1}^{n} \zeta_i + \sum_{i=1}^{n} \alpha_i (y_i - f_i - \xi_i) + \sum_{i=1}^{n} \gamma_i (-y_i + f_i - \zeta_i) \\ - \sum_{i=1}^{n} \kappa_i \xi_i - \sum_{i=1}^{n} \rho_i \zeta_i + \frac{\lambda_2}{2} \sum_{j=1}^{p} \beta_j^2 + \eta \left(\sum_{j=1}^{p} |\beta_j| - s \right), \end{aligned}$$

where $\alpha_i \geq 0,\ \delta_i \geq 0,\ \gamma_i \geq 0$ and $\eta \geq 0$ are Lagrange multipliers. Taking derivatives with respect to β_0, β, ξ_i and ζ_i, we have

$$\begin{aligned} &\frac{\partial L}{\partial \beta_0} : \sum_{i=1}^{n} (\alpha_i - \gamma_i) = 0, \\ &\frac{\partial L}{\partial \beta_j} : -\sum_{i=1}^{n} (\alpha_i - \gamma_i) x_{ij} + \lambda_2 \beta_j + \eta \text{sign}(\beta_j) = 0, \quad j \in \mathcal{V}, \\ &\frac{\partial L}{\partial \xi_i} : \tau - \alpha_i - \kappa_i = 0, \quad i = 1, \ldots, n, \\ &\frac{\partial L}{\partial \zeta_i} : (1-\tau) - \gamma_i - \rho_i = 0, \quad i = 1, \ldots, n. \end{aligned}$$

Then Karush-Kuhn-Tucker(KKT) conditions from the optimization problem are

$$\begin{aligned} &\alpha_i (y_i - f_i - \xi_i) = 0, \qquad \gamma_i (y_i - f_i + \zeta_i) = 0, \\ &\kappa_i \xi_i = 0, \qquad \rho_i \zeta_i = 0, \\ &\eta \left(\sum_{j=1}^{p} |\beta_j| - s \right) = 0. \end{aligned}$$

And then the corresponding KKT conditions imply

1. $\mathcal{L} = \{i : y_i - f_i < 0\} \longrightarrow\ \zeta_i > 0,\ \rho_i = 0,\ \gamma_i = 1-\tau,\ \alpha_i = 0,\ \kappa_i = \tau$ and $\xi_i = 0$,
2. $\mathcal{E} = \{i : y_i - f_i = 0\} \longrightarrow\ 0 \leq \alpha_i \leq \tau,\ \xi_i = 0,\ 0 \leq \gamma_i \leq 1-\tau$ and $\zeta_i = 0$,
3. $\mathcal{R} = \{i : y_i - f_i > 0\} \longrightarrow\ \xi_i > 0,\ \kappa_i = 0,\ \alpha_i = \tau,\ \gamma_i = 0,\ \rho_i = 1-\tau$ and $\zeta_i = 0$.

Therefore, without loss of generality we let $\tilde{\alpha}_i = \alpha_i - \gamma_i$. Then $\tilde{\alpha}_i = -(1-\tau)$ for $i \in \mathcal{L}$ and $\tilde{\alpha}_i = \tau$ for $i \in \mathcal{R}$. And also, $\tilde{\alpha}_i$ is in $[-(1-\tau),\ \tau]$ for $i \in \mathcal{E}$. From the above conditions,

we acquire following linear equations:

$$\begin{cases} \lambda_2\beta_j - \sum_{i=1}^{n}\tilde{\alpha}_i x_{ij} + \eta\text{sign}(\beta_j) = 0, \quad \text{for } j \in \mathcal{V}, \\ \sum_{i=1}^{n}\tilde{\alpha}_i = 0, \\ \beta_0 + \sum_{j\in\mathcal{V}}\beta_j x_{ij} = y_i, \quad i \in \mathcal{E}, \\ \|\beta\|_1 = \sum_{j=1}^{p}\text{sign}(\beta_j)\beta_j = s. \end{cases} \tag{2.5}$$

Let $\Delta\beta_0/\Delta s$, $\Delta\beta_j/\Delta s$, $\Delta\tilde{\alpha}_i/\Delta s$ and $\Delta\eta/\Delta s$ be the right derivative with respect to s. The right derivatives are obtained by solving the following linear equations

$$\begin{cases} \lambda_2\dfrac{\Delta\beta_j}{\Delta s} - \sum_{i\in\mathcal{E}}\dfrac{\Delta\tilde{\alpha}_i}{\Delta s}x_{ij} + \dfrac{\Delta\eta}{\Delta s}\text{sign}(\beta_j) = 0, \quad \text{for } j \in \mathcal{V}, \\ \sum_{i\in\mathcal{E}}\dfrac{\Delta\tilde{\alpha}_i}{\Delta s} = 0, \\ \dfrac{\Delta\beta_0}{\Delta s} + \sum_{j\in\mathcal{V}}\dfrac{\Delta\beta_j}{\Delta s}x_{ij} = 0, \quad i \in \mathcal{E}, \\ \sum_{j\in\mathcal{V}}\text{sign}(\beta_j)\dfrac{\Delta\beta_j}{\Delta s} = 1. \end{cases} \tag{2.6}$$

Then when s increases, by continuity of solutions, the set $\mathcal{L}, \mathcal{E}, \mathcal{R}$ and $\mathcal{V}$ will not change and hence, the right derivatives will not change. The key idea of the proposed algorithm is that we update the current solution to the right derivative direction which must satisfy KKT condition (2.5) where updating formula is on (2.7). For more details on path-following algorithm, see Wang *et al.* (2006).

Note that in the above linear system (2.6), there is $|\mathcal{V}| + |\mathcal{E}| + 2$ unknowns and $|\mathcal{V}| + |\mathcal{E}| + 2$ equations. But if $|\mathcal{E}| > |\mathcal{V}|$ then the solution can not be found because of the third equation of (2.6). Therefore, it is important to satisfy $|\mathcal{E}| \leq |\mathcal{V}|$ to get new right derivatives for the next step.

To sum up, the code of DrQR including L1QR is below.

DrQR Algorithm

1. Given τ, let
$$Q(\beta) = \sum_{i=1}^{n}\rho_\tau\big(y_i - (\beta_0 + \mathbf{x}_i'\beta)\big) + \frac{\lambda_2}{2}\sum_{j=1}^{p}\beta_j^2.$$
2. Find initial β_0 and specify initial $\mathcal{E}$ and $\mathcal{V}$.
3. Repeat until current η reduces to 0 or step size.

(a) Let the generalized correlation be $c_j = -\sum_{i=1}^{n} \tilde{\alpha}_i x_{ij} + \lambda_2 \beta_j$, compute the derivative of the current generalized correlation value for variable x_j.

$$\frac{\Delta c_j}{\Delta s} = -\sum_{i \in \mathcal{E}} \frac{\Delta \tilde{\alpha}_i}{\Delta s} x_{ij}, \quad \text{for } j \notin \mathcal{V}.$$

(b) Let the current residual be $r_i = y_i - f(\mathbf{x}_i)$, compute the derivative of the residual for every points

$$\frac{\Delta r_i}{\Delta s} = -\left(\frac{\Delta \beta_0}{\Delta s} + \sum_{j \in \mathcal{V}} \frac{\Delta \beta_j}{\Delta s} x_{ij} \right).$$

(c) Compute how much increase of s is needed to get to each type of event:

i. An elbow point leaves the elbow set, $\delta_1 = \min_{i \in \mathcal{E}^-} \max\big((-(1-\tau) - \tilde{\alpha}_i)/(\Delta\tilde{\alpha}_i/\Delta s), (\tau - \tilde{\alpha}_i)/(\Delta\tilde{\alpha}_i/\Delta s)\big)$, where $\mathcal{E}^- = \{i : i \in \mathcal{E}, (-(1-\tau) - \tilde{\alpha}_i)/(\Delta\tilde{\alpha}_i/\Delta s) > 0 \text{ or } (\tau - \tilde{\alpha}_i)/(\Delta\tilde{\alpha}_i/\Delta s) > 0\}$.

ii. A nonelbow point becomes elbow point, $\delta_2 = \min_{i \in \mathcal{E}^+} \big((0 - r_i)/(\Delta r_i/\Delta s)\big)$, where $\mathcal{E}^+ = \{i : i \notin \mathcal{E}, (0 - r_i)/(\Delta r_i/\Delta s) > 0\}$.

iii. An active variable becomes inactive, $\delta_3 = \min_{j \in \mathcal{V}^-} \big((0 - \beta_j)/(\Delta\beta_j/\Delta s)\big)$, where $\mathcal{V}^- = \{j : j \in \mathcal{V}, (0 - \beta_j)/(\Delta\beta_j/\Delta s) > 0\}$.

iv. An inactive variable joins the active set, $\delta_4 = \min_{j \in \mathcal{V}^+} \min\big((-\eta - c_j)/(\Delta c_j/\Delta s + \Delta\eta_j/\Delta s), (\eta - c_j)/(\Delta c_j/\Delta s - \Delta\eta_j/\Delta s)\big)$, where $\mathcal{V}^+ = \{j : j \notin \mathcal{V}, (-\eta - c_j)/(\Delta c_j/\Delta s + \Delta\eta_j/\Delta s) > 0, (\eta - c_j)/(\Delta c_j/\Delta s - \Delta\eta_j/\Delta s) > 0\}$.

v. The generalized correlation of active variables reduces to 0, $\delta^5 = -\eta/(\Delta\eta/\Delta s)$.

(d) Find which event happens first. Set $\delta_s = \min(\delta_1, \delta_2, \delta_3, \delta_4, \delta_5)$ and update

$$\begin{cases} \tilde{\alpha}_i \leftarrow \tilde{\alpha}_i + \delta_s \Delta\tilde{\alpha}_i/\Delta s, & i \in \mathcal{E}, \\ \beta_0 \leftarrow \beta_0 + \delta_s \Delta\beta_0/\Delta s, & \\ \beta_j \leftarrow \beta_j + \delta_s \Delta\beta_j/\Delta s, & j \in \mathcal{V}, \\ \eta \leftarrow \eta + \delta_s \Delta\eta/\Delta s, & \\ s \leftarrow s + \delta_s. & \end{cases} \tag{2.7}$$

(e) Update $\mathcal{L}$, $\mathcal{E}$, $\mathcal{R}$ and $\mathcal{V}$.

(f) Compute the new right derivatives using linear system (2.6).

Note that in path-following algorithm, it is important to find initial β_0, $\mathcal{E}$ and $\mathcal{V}$. Our proposed algorithm also uses the strategy of Li *et al.* (2007).

2.3. Tuning parameter selection

For any regularization methods, an important issue is to find a good choice of the regularization parameter such that the corresponding model is optimal according to

some criterion. Yuan (2006) presented the generalized approximate cross-validation criterion(GACV) in selecting the tuning parameter for quantile spline method. This measure is defined by

$$\text{GACV}(\lambda) = \frac{\sum_{i=1}^{n} \rho_\tau \left(y_i - \hat{f}(\mathbf{x}_i) \right)}{n - df},$$

where $df = |\mathcal{V}|(|\cdot|$: size of a set). We select λ_1 which minimizes GACV in the training set for the L1QR method. For the DrQR, we select (λ_1, λ_2) which minimizes GACV in the training set.

3. Numerical Studies

In this section, we investigate the finite sample performance of the DrQR estimator via simulation experiments as well as real data analysis. In particular, we compare the DrQR estimator with the L1QR estimator in terms of selectivity(power of selecting of true nonzero coefficients) and prediction accuracy. All simulations is conducted by **R** programm and by the pentium laptop with cpu 2.0GHz and ram 1GB.

3.1. Simulation: Prediction accuracy and selectivity

We now compare the performance of the DrQR and L1QR in terms of selectivity and prediction accuracy through simulation. The simulation model is

$$y = \sum_{k=1}^{p} \beta_k x_k + \epsilon, \tag{3.1}$$

where all the βs except the first 8 βs are set to zero and first 8 βs are set to ($\underbrace{2, 1.5}_{g_1}$, $\underbrace{-1, -1, -1}_{g_2}$, $\underbrace{-0.5, -0.5, -0.5}_{g_3}$). In order, denote three groups variables by g_1, g_2 and g_3. And $\mathbf{x} = (x_1, \ldots, x_p)'$ is a multivariate Gaussian random vector with mean 0 and blockwise covariances matrix

$$\begin{pmatrix} \Sigma^1_{2\times 2} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \Sigma^2_{3\times 3} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \Sigma^3_{3\times 3} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \Sigma^4_{(p-8)\times(p-8)} \end{pmatrix},$$

where each Σ^j has elements of being $r^{|k-l|}$, correlations between x_k and x_l contained in a each block for some $r \in [0, 1)$. The ϵ is a Gaussian random variable with mean 0 and variance 2. We fix the sample size to 100 and we investigate the performances with $p = (20, 150)$ and $r = (0.0, 0.8)$, respectively. The results about prediction accuracy are the averages based on 20 repetitions of the simulation. The regularization parameters are selected by the GACV. The prediction errors and mean absolute deviation(MAD)

Table 3.1: Simulation results when $\tau = 0.5$ and 0.9 the prediction accuracy of DrQR and L1QR: mean prediction errors(standard errors), mean absolute deviation(standard errors), $\#g_j$(average of size of j^{th} group variable being selected)

τ	p	r	Method	Test Error	Test MAD	$\#g_1$	$\#g_2$	$\#g_3$
0.5	20	0.0	L1QR	0.949 (0.012)	1.038 (0.043)	2	2.95	2.75
			DrQR	0.943 (0.011)	1.019 (0.040)	2	2.95	2.85
		0.8	L1QR	0.950 (0.012)	1.046 (0.043)	2	2.85	2.25
			DrQR	0.929 (0.013)	0.959 (0.048)	2	2.90	2.65
	150	0.0	L1QR	1.765 (0.087)	3.119 (0.195)	2	3.00	2.60
			DrQR	1.687 (0.062)	2.947 (0.139)	2	3.00	2.70
		0.8	L1QR	1.853 (0.126)	3.305 (0.285)	2	2.70	2.40
			DrQR	1.622 (0.065)	2.809 (0.150)	2	2.75	2.65
0.9	20	0.0	L1QR	0.678 (0.011)	1.795 (0.052)	2	3.00	2.80
			DrQR	0.670 (0.011)	1.779 (0.052)	2	3.00	2.80
		0.8	L1QR	0.677 (0.012)	1.790 (0.054)	2	2.75	2.30
			DrQR	0.662 (0.012)	1.781 (0.059)	2	2.90	2.60
	150	0.0	L1QR	1.835 (0.120)	3.299 (0.237)	2	3.00	2.50
			DrQR	1.811 (0.092)	3.196 (0.203)	2	3.00	2.70
		0.8	L1QR	1.866 (0.113)	3.394 (0.240)	2	2.75	2.50
			DrQR	1.762 (0.094)	3.114 (0.214)	2	2.90	2.70

are measured on independent test samples of size 1,000. These criterions are based on the following:

$$\text{Test Error} = \frac{1}{1,000}\sum_{i=1}^{1,000}\rho_\tau\big(y_i - \hat{f}(\mathbf{x}_i)\big), \quad \text{Test MAD} = \frac{1}{1,000}\sum_{i=1}^{1,000}\big|f(\mathbf{x}_i) - \hat{f}(\mathbf{x}_i)\big|$$

for the test sample, where the fitted quantile function is $\hat{f}(\mathbf{x})$ and the true quantile function is $f(\mathbf{x})$. From Table 3.1, we can see that the DrQR estimator tends to be better in both the test error and the MAD than the L1QR estimator when the correlation between predictive variables is large, but those are similar when the correlation is small or p is small. Also, DrQR has more power in selecting nonzero coefficients in the highly correlated setting($r = 0.8$). Additionally, similar conclusion is observed in results of the high quantile $\tau = 0.9$ case.

3.2. Real data analysis: Boston Housing data

In this section, we analyzed Boston Housing data obtained from the library(mlbench) in **R** program. This data set has 506 census tracts of Boston from the 1970 census where one target variable and 13 explanatory variables. The target value is median value of owner-occupied homes in USD 1,000's. Given each $\tau(0.1, 0.5, 0.9)$, we fit the L1QR and DrQR from training data and then, measure their performances from test data. As shown in Table 3.2, DrQR performs better than L1QR in test error.

Figure 3.1 describes nonzero coefficients of selected models according to low quantile($\tau = 0.1$), median($\tau = 0.5$) and high quantile($\tau = 0.9$). Except zn(proportion of residential land zoned for lots over 25,000 ft), other 4 variables(dis: weighted distances to five Boston employment centres, tax: full-value property-tax rate per 10,000 USD, b: where B is the

Table 3.2: Results of Boston Housing data

τ	Method	Test Error	Nonzeros
0.1	L1QR	1.693 (0.02)	10.9
	DrQR	1.681 (0.02)	11.1
0.5	L1QR	1.768 (0.02)	9.8
	DrQR	1.753 (0.02)	10.7
0.9	L1QR	1.267 (0.04)	10.3
	DrQR	1.258 (0.04)	10.6

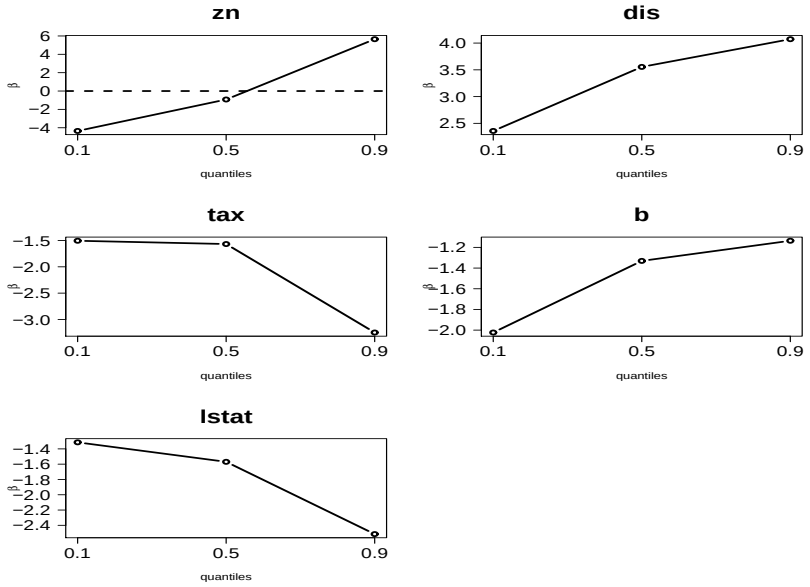


Figure 3.1: Nonzero coefficients of selected models according to three quantiles

proportion of blacks by town, lstat: percentage of lower status of the population) have coefficients with same sign though the scales are different.

3.3. Real data analysis: Microarray data

In this section, we apply the DrQR to a quantile regression problem of gene microarrays in high dimensional settings. We employ the data set used in Scheetz *et al.* (2006), which consists of gene expression levels of 18,975 genes obtained from 120 rats. The main objective of the analysis is to find genes that are correlated with gene TRIM32 known to cause Bardet-Biedl syndrome. As was done by Huang *et al.* (2006), we first select 3000 genes with the largest variance in expression level and then choose the top 500 genes that have the largest absolute correlation with gene TRIM32 among the selected 3000 genes.

Each data set is divided into two parts, training and test data sets, by randomly selecting 2/3 observations and 1/3 observations, respectively. The optimal values of the regularization parameters are chosen by GACV. Results of 20 replicated experiments are

Table 3.3: Results of Microarray Data

τ	Method	Test Error	Nonzeros
0.1	L1QR	0.379 (0.03)	72.71
	DrQR	0.339 (0.02)	146.80
0.5	L1QR	0.421 (0.02)	75.35
	DrQR	0.413 (0.03)	105.65
0.9	L1QR	0.365 (0.09)	69.94
	DrQR	0.347 (0.08)	151.89

summarized in Table 3.3 according to τ quantiles.

As shown in Table 3.3, the DrQR performs best in terms of Test Error. Meanwhile, the number of nonzero coefficients of the DrQR is much larger than those of the L1QR. because some signal variables are highly correlated.

4. Discussion

In this paper, we presented the doubly regularized quantile estimator and its computational algorithm which provides entire solution path following algorithm along λ_1 and also, we showed in highly correlated setting, the proposed method(DrQR) performs better than L1QR of Li and Zhu (2008) in view of accuracy and selectivity.

But, since the estimated quantile functions was independently fitted, they can overlap the order of true quantile function. To resolve this, that is, to maintain the order of conditional quantile functions, we can consider the following composite quantile regression functional

$$\arg \min_{\beta_{0k},\beta\in\mathbb{R}^p} \sum_{k=1}^{K}\sum_{i=1}^{n} \rho_\tau\big(y_i - (\beta_{0k} + \mathbf{x}_i'\beta)\big) + \frac{\lambda_2}{2}\sum_{j=1}^{p}\beta_j^2 + \lambda_1 \sum_{j=1}^{p}|\beta_j|,$$

where $0 < \tau_1 < \tau_2 < \cdots < \tau_K < 1$. However, one of drawbacks of composite quantile regression is to use only one β across given quantiles, that is, use only one slope β. In this case, the corresponding computation is very complex. We remain this topic as future work.

References

Choi, H. (2007). An extension of COSSO algorithm by combining variables, *Journal of the Korean Data Analysis Society,* **9**, 2117–2125.

Efron, B., Hastie, T., Johnstone, I. and Tibshirani, R. (2004). Least angle regression, *The Annals of Statistics,* **32**, 407–499.

Huang, J., Ma, S. and Zhang, C. H. (2006). *Adaptive Lasso for sparse high-dimensional regression models*, Technical Paper 374, University of Iowa, Dept. of Statistics and Actuarial Science.

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

Koenker, R., Ng, P. and Portnoy, S. (1994). Quantile smoothing splines, *Biometrika,* **81**, 673–680.

Li, Y., Liu, Y. and Zhu, J. (2007). Quantile regression in reproducing kernel Hilbert spaces, *Journal of the American Statistical Association,* **102**, 255–268.

Li, Y. and Zhu, J. (2008). L_1-norm Quantile Regression, *Journal of Computational & Graphical Statistics,* **17**, 163–185.

Oh, H. S., Nychka, D., Brown, T. and Charbonneau, P. (2004). Period analysis of variable stars by robust smoothing, *Journal of the Royal Statistical Society, Series C,* **53**, 15–30.

Scheetz, T. E., Kim, K. Y., 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,* **103**, 14429–14434.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso, *Journal of the Royal Statistical Society, Series B,* **58**, 267–288.

Wang, L., Zhu, J. and Zou, H. (2006). The doubly regularized support vector machine, *Statistica Sinica,* **16**, 589–615.

Yuan, M. (2006). GACV for quantile smoothing splines, *Computational Statistics & Data Analysis,* **50**, 813–829.

Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net, *Journal of Royal Statistical Society, Series B,* **67**, 301–320.

[Received July 2008, Accepted August 2008]