Computational Statistics and Data Analysis 52 (2008) 5296–5304

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

journal homepage: www.elsevier.com/locate/csda

Regularized simultaneous model selection in multiple quantiles regression

Hui Zou [a,*], Ming Yuan [b]

[a] University of Minnesota, United States
[b] Georgia Institute of Technology, United States

ARTICLE INFO

Article history:
Received 6 September 2007
Received in revised form 16 May 2008
Accepted 17 May 2008
Available online 27 May 2008

ABSTRACT

Simultaneously estimating multiple conditional quantiles is often regarded as a more appropriate regression tool than the usual conditional mean regression for exploring the stochastic relationship between the response and covariates. When multiple quantile regressions are considered, it is of great importance to share strength among them. In this paper, we propose a novel regularization method that explores the similarity among multiple quantile regressions by selecting a common subset of covariates to model multiple conditional quantiles simultaneously. The penalty we employ is a matrix norm that encourages sparsity in a column-wise fashion. We demonstrate the effectiveness of the proposed method using both simulations and an application of gene expression data analysis.

1. Introduction

Consider a general regression setting where y represents the response and $\mathbf{x} = (x_1, \ldots, x_p)^{\mathrm{T}}$ represents a set of predictors. Classical regression methods focus on recovering the conditional expectation $E(Y \mid X)$. Quantile regression (Koenker and Bassett, 1978) estimates the conditional quantile functions. Suppose we want to infer the $100\tau\%$ quantile (say $\tau = 0.5$) of the conditional distribution of the response (y) given covariates ($\mathbf{x}$) based on n independent observations $(\mathbf{x}_i, y_i)_{i=1}^n$. Koenker and Bassett (1978) showed that one can estimate the conditional τ-quantile by minimizing

$$\sum_{i=1}^{n} \rho_\tau(y_i - \beta_0 - \mathbf{x}_i^{\mathrm{T}}\beta), \tag{1}$$

where $\rho_\tau(t) = \tau t_+ + (1-\tau)t_-$ is the so-called check function where subscripts '+' and '−' stand for the positive and negative parts, respectively. Quantile regression has been widely used in many different areas such as economics (Koenker and Hallock, 2001) and survival analysis (Koenker and Geling, 2001) among others. Nonlinear estimates can be obtained by the same method, except that we replace the covariates $\mathbf{x}$ with some basis functions such as splines (He et al., 1998; Koenker et al., 1994; Yuan, 2006). In this paper we consider the variable selection problem in the linear quantile regression model.

Variable selection in conditional mean regression has received a lot of attention in recent years. Several regularization techniques have been invented for doing automatic variable selection, including the lasso (Tibshirani, 1996) and the SCAD (Fan and Li, 2001). Similar to the conditional mean regression, variable selection is also crucial in quantile regression when the number of predictors is large. A sparse model is much more interpretable in practice and often enjoys improved

* Corresponding author.
E-mail addresses: hzou@stat.umn.edu (H. Zou), myuan@isye.gatech.edu (M. Yuan).

0167-9473/$ – see front matter © 2008 Elsevier B.V. All rights reserved.
doi:10.1016/j.csda.2008.05.013

estimation accuracy by eliminating irrelevant variables. Koenker (2004) considered using the L_1 quantile regression to automatically select significant predictors in quantile regression. The L_1 quantile regression model is estimated by

$$\widehat{\boldsymbol{\beta}}(L_1\text{norm}) = \arg\min_{\beta_0,\boldsymbol{\beta}} \sum_{i=1}^{n} \rho_\tau(y_i - \beta_0 - \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}) + \lambda\|\boldsymbol{\beta}\|_1, \tag{2}$$

where $\|\boldsymbol{\beta}\|_1 = \sum_{j=1}^{p} |\beta_j|$ is the L_1-norm penalty (or lasso penalty) on $\boldsymbol{\beta}$. When the tuning parameter λ is appropriately chosen, some components of $\widehat{\boldsymbol{\beta}}$ will be shrunk to exact zero, and the corresponding variables are excluded from the final model.

In this paper, we consider the so-called *simultaneous multiple quantiles regression* (SMQR, for short), where we are interested in estimating multiple conditional quantile functions simultaneously. As demonstrated in Koenker (2005), a main advantage of quantile regression over classical mean regression is its ability to examine multiple conditional quantile functions, and provide a more comprehensive description of the relationship between the response and the covariates. Another example of SMQR arises naturally when jointly modelling multiple responses.

When considering multiple regression models such as SMQR, it is of great importance to share strength among different models as illustrated in Breiman and Friedman (1997). It is of particular interest, when there is a large number of covariates, to find a common set of variables that can be used for all models under investigation (Turlach et al., 2005). In the context of mean regression, Turlach et al. (2005) considered the problem of selecting a subset of 770 wavelengths that are suitable as predictors for 14 different but correlated infra-red spectrometry measurements, and they proposed a novel regularization method to perform simultaneous variable selection. Complementary to this earlier work, we study simultaneous variable selection in multiple quantiles regression. Simultaneous model selection is actually more relevant in quantile regression than in classical mean regression, considering that estimating multiple quantiles of a single response is routinely done in practice. It is often desirable to select a common set of significant variables for modeling a sequence of quantiles of the response. More generally, it is natural, in many applications, that the set of variables used to model different conditional quantiles overlaps with each other. The goal of this paper is to develop a model selection method that is capable of exploring such similarity and performing simultaneous model selection in multiple quantiles regression whenever simultaneous model selection is desirable.

A naive approach to model selection in multiple quantiles regression will separately fit individual L_1 quantile regression models and take a union of the selected variables from each regression model. This naive approach cannot guarantee that the same set of variables are selected within each L_1 quantile regression model. Furthermore, the naive approach may also be suboptimal in terms of predictive accuracy in some problems. For example, consider the classical model underlying linear quantile regression

$$y = x_1\beta_1 + \cdots + x_p\beta_p + \epsilon. \tag{3}$$

We omit the intercept for brevity. Clearly, all conditional quantiles can be described by the same set of variables. By recognizing this fact, estimation can be greatly improved. The naive approach does not share information across the quantile regression models, hence the results might be suboptimal compared with the methods which combine strengths from multiple models. It is well known that when estimating multiple statistical models, it is beneficial to share information across them (Breiman and Friedman, 1997). The same wisdom applies to multiple quantiles regression, as demonstrated in Section 4.

To overcome the drawbacks of the naive approach, we introduce a new regularization method for performing simultaneous model selection in multiple quantiles regression. We propose to penalize the sum of the check functions of multiple quantile regression models by a norm of the coefficient matrix that encourages column-wise sparsity. As the regularization parameter varies, the penalty does simultaneous variable selection via continuous shrinkage. The rest of the paper is organized as follows. In Section 2 we present methodological details of the penalized multiple quantiles regression. Section 3 discusses the implementation details of the proposed method. Simulation results are presented in Section 4 and we also demonstrate the utility of the proposed method on the cardiomyopathy data in Section 5.

2. Penalized multiple quantiles regression

To fix the idea, consider first estimating multiple quantiles of a single response. Suppose we want to estimate the $\tau_1, \ldots, \tau_G$ quantiles of the conditional distribution $y|\mathbf{x}$. Denote by $\boldsymbol{\beta}^{(k)} = (\beta_1^{(k)}, \ldots, \beta_p^{(k)})^{\mathrm{T}}$ the coefficients of the covariates in the τ_k conditional quantile function of y given $\mathbf{x}$, where $k = 1, 2, \ldots, G$. We also write $\boldsymbol{\beta}_{(j)} = (\beta_j^{(1)}, \ldots, \beta_j^{(G)})^{\mathrm{T}}$ for each $j = 1, 2, \ldots, p$. We say $\boldsymbol{\beta}_{(j)}$ the coefficient vector of variable x_j. Let $\boldsymbol{\beta}$ be the coefficient matrix whose (k, j) element is $\beta_j^{(k)}$. With such a notation we define a norm of $\boldsymbol{\beta}$ as follows

$$\|\boldsymbol{\beta}\|_{1\infty} = \sum_{j=1}^{p} \max_k\{|\beta_j^{(k)}|\}, \tag{4}$$

where the subscript reflects the fact that we take the vector L_1 norm of the column-wise L_∞ norms. The new matrix norm (4) was considered in Turlach et al. (2005). It is similar in spirit to the F_∞-norm penalty used by Zou and Yuan (2008), which is a

vector norm rather than matrix norm. It is interesting to note that in the context of variable selection for multi-class support vector machines, Zhang et al. (2008) also considered using a similar penalty function as ours to select the same variables for discriminating all classes.

Instead of estimating $\beta^{(k)}$ separately, we propose estimating them simultaneously by minimizing

$$\widehat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \sum_{k=1}^{G} \left\{ \sum_{i=1}^{n} \rho_{\tau_k}(y_i - \mathbf{x}_i^{\mathrm{T}} \boldsymbol{\beta}^{(k)}) \right\} + \lambda \|\boldsymbol{\beta}\|_{1\infty}. \tag{5}$$

The estimated conditional τ_k-quantile of $y^{(k)}$ is $\mathbf{x}^{\mathrm{T}}\widehat{\boldsymbol{\beta}}^{(k)}$.

When $G = 1$, then (5) reduces to the L_1 quantile regression. Just as the lasso penalty does automatic variable selection, the new matrix norm penalty enables one to perform simultaneous variable selection. If the regularization parameter λ is appropriately chosen, some $\widehat{\boldsymbol{\beta}}_j$ will be exact zero, that is, every component of $\widehat{\boldsymbol{\beta}}_j$ is exactly zero. Hence variable x_j is excluded from all G quantile regression models. This nice property comes from the singular nature of the penalty. As pointed out in Fan and Li (2001), singularity (at the origin) of the penalty function plays a central role in automatic feature selection.

The idea can be naturally generalized to more complicated situations with multiple quantiles and/or multiple responses. Given n training samples $\{(x_i, y_i^{(1)}, \ldots, y_i^{(G)})\}_{i=1}^n$, we can estimate the τ_k quantile of $y^{(k)}$ conditioning on $\mathbf{x}$, $1 \le k \le G$ by

$$\widehat{\beta} = \arg\min_{\boldsymbol{\beta}} \sum_{k=1}^{G} \left\{ \sum_{i=1}^{n} \rho_{\tau_k}(y_i^{(k)} - \mathbf{x}_i^{\mathrm{T}} \boldsymbol{\beta}^{(k)}) \right\} + \lambda \|\boldsymbol{\beta}\|_{1\infty}. \tag{6}$$

Here the responses $y^{(k)}$, $k = 1, \ldots, G$ can be the same or different depending on the problem. For instance, if we want to study the relationship between the height ($y^{(1)}$) and weight ($y^{(2)}$) of infants and a set of medical/biological/genetical measurements (x), then it is quite reasonable to use the same variables to model the height and weight. To that end, we can employ the above SMQR method.

In some applications, we may just want to penalize a subset of the variables. For instance, $\mathbf{x}$ includes 1 and we do not penalize the intercept. This is actually a common practice, as done in our numerical examples. The proposed method can be easily modified to meet this requirement. Let S denote the subset of variables which are to be penalized. Similar to before, we estimate $\boldsymbol{\beta}$ by

$$\widehat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \sum_{k=1}^{G} \left\{ \sum_{i=1}^{n} \rho_{\tau_k}(y_i^{(k)} - \mathbf{x}_i^{\mathrm{T}} \boldsymbol{\beta}^{(k)}) \right\} + \lambda \|\boldsymbol{\beta}_S\|_{1\infty}. \tag{7}$$

We should point out here that (4) is not the only penalty function that can be used to do simultaneous model selection. One possible alternative is the *grouped lasso* penalty proposed by Yuan and Lin (2006). Consider $(\beta_j^1, \ldots, \beta_j^G)$ as a group of coefficients. The grouped lasso penalty on β_j is defined as $\sqrt{\sum_{k=1}^{G}(\beta_{(j)}^k)^2} = \|\beta_{(j)}\|_2$. We consider

$$\arg\min_{\boldsymbol{\beta}} \sum_{k=1}^{G} \left\{ \sum_{i=1}^{n} \rho_{\tau_k}(y_i^{(k)} - \mathbf{x}_i^{\mathrm{T}} \boldsymbol{\beta}^{(k)}) \right\} + \lambda \sum_{j=1}^{p} \|\beta_{(j)}\|_2. \tag{8}$$

Yuan and Lin (2006) suggested using an iterative procedure to solve the grouped lasso, which could be expensive for high dimensional problems. In contrast, we show in the next section that using the new matrix norm penalty allows us to take advantage of efficient linear programming tools to fit the penalized multiple quantiles regression model. Computational efficiency is a major reason for us to employ the new matrix norm penalty in penalized multiple quantiles regression.

3. Implementation

In this section we show that (5) can be solved efficiently by linear programming techniques. We also consider data-driven methods for selecting the regularization parameter λ.

3.1. Computing algorithm

To derive an equivalent linear program for (5), we introduce the following slack variables

$$\xi_i^{(k)} = \rho_{\tau_k}(y_i^{(k)} - \mathbf{x}_i^{\mathrm{T}} \boldsymbol{\beta}^{(k)}), \tag{9}$$

$$M_j = \max_k \{|\beta_j^{(k)}|\}. \tag{10}$$

It is clear that M_j is the maximum absolute value of the coefficient vector of variable x_j. Write $\beta_j^{(k)}$ as $(\beta_j^{(k)})^+ - (\beta_j^{(k)})^-$ where $(\beta_j^{(k)})^+$ and $(\beta_j^{(k)})^-$ denote the positive and negative parts of $\beta_j^{(k)}$, respectively. Using this notation, (5) amounts to minimizing

$$\sum_{k=1}^{G}\sum_{i=1}^{n}\xi_i^{(k)} + \lambda\sum_{j=1}^{p}M_j \tag{11}$$

subject to

$$\xi_i^{(k)} \geq -(1-\tau_k)\left(y_i^{(k)} - \sum_{j=1}^{p}\mathbf{x}_{i,j}[(\beta_j^{(k)})^+ - (\beta_j^{(k)})^-]\right),$$

$$\xi_i^{(k)} \geq \tau_k\left(y_i^{(k)} - \sum_{j=1}^{p}\mathbf{x}_{i,j}[(\beta_j^{(k)})^+ - (\beta_j^{(k)})^-]\right),$$

$$M_j \geq (\beta_j^{(k)})^+ + (\beta_j^{(k)})^-,$$

$$\xi_i^{(k)} \geq 0,\ (\beta_j^{(k)})^+ \geq 0,\ (\beta_j^{(k)})^- \geq 0$$

for all $1 \leq i \leq n,\ 1 \leq k \leq G$ and $1 \leq j \leq p$. Note that (11) is a linear program and can be solved using a standard linear programming solver. For example, in R the *lpSolve* package provides functions for solving linear programs. We have implemented the procedure in R by using *lpSolve* package and our R code is available for interested readers upon request.

3.2. Regularization parameter selection

The choice of the tuning parameter plays an important role in determining the performance of the proposed estimate. Koenker et al. (1994) suggested using Schwartz Information Criterion (*SIC*) as the tuning method for quantile regression. *SIC* can be written as

$$SIC(\lambda) = \log\left(\frac{1}{nG}\sum_{k=1}^{G}\sum_{i=1}^{n}\rho_{\tau_k}(y_i^{(k)} - \hat{f}_\lambda^{(k)}(\mathbf{x}_i))\right) + \frac{\log(nG)}{2nG}p_\lambda, \tag{12}$$

where p_λ is the effective degrees of freedom of the fitted model, which is similar to the effective degrees of freedom in ordinary regression (Meyer and Woodroofe, 2000). For a given tuning parameter λ, we define a set $\mathcal{E}_\lambda$ as

$$\mathcal{E}_\lambda = \left\{(k,i) : y_i^{(k)} - \hat{f}_\lambda^{(k)}(\mathbf{x}_i) = 0\right\}. \tag{13}$$

Following Li and Zhu (2008) we call $\mathcal{E}_\lambda$ the elbow set. Let $|\mathcal{E}_\lambda|$ denote the size of the set $\mathcal{E}_\lambda$. Koenker et al. (1994) conjectured that $|\mathcal{E}_\lambda|$ is the effective degrees of freedom in the quantile regression. Li et al. (2007) and Li and Zhu (2008) provided a rigorous proof to that conjecture. Following their results, p_λ should be taken as the size of elbow set. Therefore, we have

$$SIC(\lambda) = \log\left(\frac{1}{nG}\sum_{k=1}^{G}\sum_{i=1}^{n}\rho_{\tau_k}(y_i^{(k)} - \hat{f}_\lambda^{(k)}(\mathbf{x}_i))\right) + \frac{\log(nG)}{2nG}|\mathcal{E}_\lambda|. \tag{14}$$

Let $\hat{\lambda}$ be the minimizer of $SIC(\lambda)$, i.e.,

$$\hat{\lambda} = \arg\min_\lambda SIC(\lambda).$$

Then our fit is $\{\hat{f}_{\hat{\lambda}}^{(k)}(\mathbf{x})\}_{k=1}^{G}$.

Alternatively, we can also use nonparametric model assessment and selection methods such as cross-validation and bootstrap (Hastie et al., 2001) to select the tuning parameter. According to our experience, SIC generally has comparable performance with these methods but is much faster to compute.

4. Simulation

In this section we conduct a Monte Carlo simulation to check the performance of the proposed method. Two criteria are considered: model error and model selection performance. For any fit $\{\hat{f}^{(k)}\}_{k=1}^{G}$ its model error is defined as

$$ME(\hat{f}) = E_{y,\mathbf{x}}\left[\frac{1}{G}\sum_{k=1}^{G}\left(\frac{1}{n}\sum_{i=1}^{n}E_{z^{(k)}}\left[\rho_{\tau_k}(z_i^{(k)} - \hat{f}^{(k)}(\mathbf{x}_i))\right]\right)\right]. \tag{15}$$

In all simulated examples the underlying model has a sparse representation. Model selection performance is measured by the sparsity of the fitted model.

Table 1
Simulation model 1 ($n = 100$)

ErrDist	Oracle	SMQR			L_1		
	ME	ME	NC	NIC	ME	NC	NIC
N	0.522	0.545(0.002)	3(0)	1.78(0.12)	0.554(0.002)	3(0)	3.26(0.12)
DE	0.401	0.420(0.001)	3(0)	1.60(0.12)	0.427(0.001)	3(0)	3.30(0.12)
MN	0.447	0.471(0.001)	3(0)	1.48(0.12)	0.479(0.002)	3(0)	3.05(0.12)

"ErrDist" denotes the error distribution. "N" stands for normal, "DE" stands for double exponential and "MN" stands for a mixture of normals. The numbers in parenthesis are standard errors.

As a comparison we also include the L_1 quantile regression model (Koenker, 2004) in the simulation. For response $y^{(k)}$ we estimate its conditional τ_k quantile by

$$\widehat{\boldsymbol{\beta}}^{(k)}(L_1\text{norm}) = \arg\min \sum_{i=1}^{n} \rho_{\tau_k}(y_i^{(k)} - b^{(k)} - \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}) + \lambda_k\|\boldsymbol{\beta}\|_1. \tag{16}$$

As discussed before, each $\widehat{\boldsymbol{\beta}}^{(k)}(L_1\text{norm})$ can adopt a sparse representation, but the L_1-norm estimates do not necessarily conduct the simultaneous selection. Let $A^{(k)} = \{j : \widehat{\boldsymbol{\beta}}_j^{(k)}(L_1\text{norm}) \neq 0\}$ denote the selected variables in the τ_k quantile function. Then $\cup_{k=1}^{G} A^{(k)}$ is the set of variables that are used in the L_1 quantile regression.

In the first set of simulations we considered the following model

$$y = 3x_1 + 1.5x_2 + 2x_5 + \epsilon,$$

where the predictors $\mathbf{x} = (x_1, x_2, x_3, x_4, x_5, {}_x 6, x_7, x_8)^{\mathrm{T}}$ follow a multivariate normal distribution $N(0, \boldsymbol{\Sigma})$ and $\boldsymbol{\Sigma}_{i,j} = 0.5^{|i-j|}$ for $1 \leq i, j \leq 8$. This is a classical model for testing variable selection methods. Note that we can derive the exact conditional quantiles of y given $\mathbf{x}$, which is called the oracle

$$\hat{f}^{(k)}_{\text{oracle}}(\mathbf{x}) = 3x_1 + 1.5x_2 + 2x_5 + b_k,$$

where b_k is the τ_k quantile of the error distribution. We want to simultaneously estimate the 10%, 25%, 50%, 75%, 90% conditional quantiles of y. Since the error follows a symmetric distribution, it suffices to estimate the 50%, 75% and 90% quantiles. We use the model error of the oracle as the benchmark for measuring the prediction performance of an estimator.

In this simulated model we know the true underlying model has a sparse representation. Let *NC* denote the number of covariates in $\{x_1, x_2, x_5\}$ that have nonzero coefficient vectors and let *NIC* denote the number of covariates in $\{x_3, x_4, x_6, x_7, x_8\}$ that have nonzero coefficient vectors. Hence *NC* is the number of correctly selected variables and *NIC* is the number of incorrectly selected variables. We can use the pair (*NC*, *NIC*) to describe the model selection performance of a procedure.

We generated 100 training samples from the above model. Model fitting and tuning were done only using the training data. SIC was used to select the regularization parameters in the two methods. We also collected an independent test data set of 10,000 samples to compute the model error of each fit. We considered three types of error distributions: $N(0, 3)$ (N), a double exponential (DE) and a mixture of normals (MN) $0.1N(0, 25) + 0.9N(0, 1)$. Table 1 presents the model errors of the oracle estimator for the three error distributions. For a given error distribution we repeated the simulation 100 times. Table 1 also shows the mean and standard errors of the interested quantities over the 100 replications.

From Table 1, we see that both the proposed method and the L_1 quantile regression produce accurate models that are close to the ground truth. Nevertheless, SMQR is significantly more accurate than the three L_1 quantile regression models. SMQR has its biggest advantage in variable selection. We see from Table 1 that SMQR produces a significantly sparser model than the L_1 quantile regression, because the former has the ability to conduct simultaneous elimination. To more explicitly reveal this property, we took the output of one simulation with double exponential errors and plotted the estimated coefficients of each variable against λ (in logarithmic scale) in Fig. 1. Each variable has three coefficient curves. Let us look at the plots for variables x_3, x_4, x_6, x_7 and x_8. As λ reaches the value which gives the smallest SIC, all three coefficient curves collapse to zero. This phenomenon is simultaneous elimination.

In the second set of simulations we considered estimating quantiles of different responses. We generated three responses from the following model

$$y^{(1)} = 4x_1 + 4x_2 + \epsilon^{(1)},$$
$$y^{(2)} = 4x_1 + 4x_2 + 2x_5 + \epsilon^{(2)},$$
$$y^{(3)} = 3x_1 + 3x_2 + \epsilon^{(3)},$$

where $\epsilon^{(1)}$, $\epsilon^{(2)}$ and $\epsilon^{(3)}$ are independent samples from $N(0, 3)$, the double exponential and the mixture of normals $(0.9N(0, 25) + 0.1N(0, 1))$, respectively.

We generated predictors $\mathbf{x} = (x_1, x_2, x_3, x_4, x_5, x_6, x_7, x_8, x_9, x_{10}, x_{11}, x_{12})^{\mathrm{T}}$ from a multivariate normal distribution $N(0, \boldsymbol{\Sigma}_{\mathbf{x}})$. In the experiment we considered three types of covariance:

(1) Uncorrelated predictors: $\boldsymbol{\Sigma}_{\mathbf{x}} = \boldsymbol{\Sigma}_1$, where $\boldsymbol{\Sigma}_1$ is the 12×12 identity matrix.

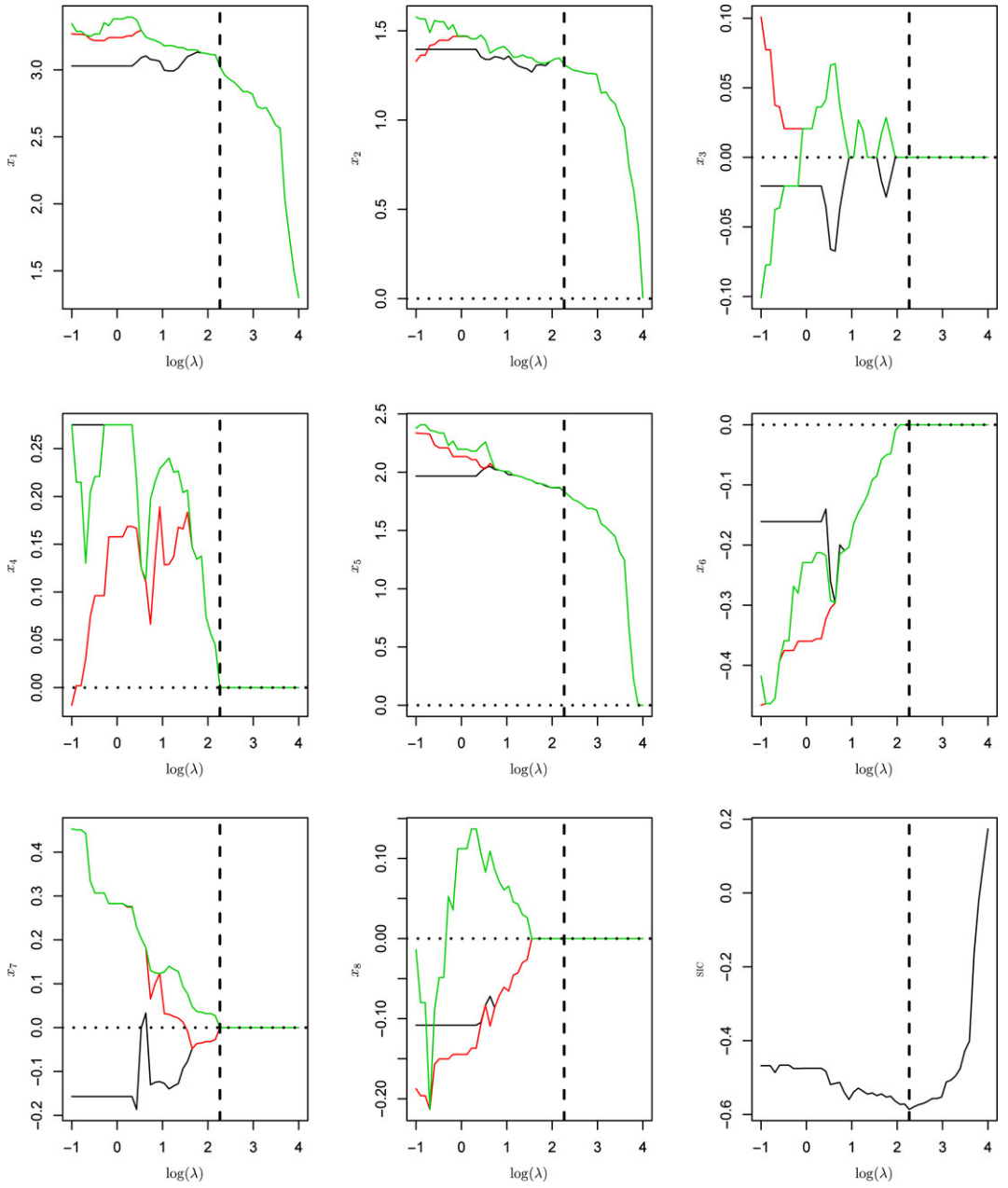

Fig. 1. SMQR in simulation model 1 ($n = 100$). We plot the coefficients of each variable. In each panel the thick broken line indicates the optimal λ value selected by SIC. The SIC curve is shown in the last panel (right corner). When λ is small, the three coefficient curves in each panel are different. That is expected since a small λ does not provide enough regularization through the penalty function. When λ is greater than a certain threshold, the three coefficient curves merge into one, because the penalty tends to shrink the coefficients of variables within a group to a common value.

(2) Pairwise constant correlation: $\Sigma_{\mathbf{x}} = \Sigma_2$, where $\Sigma_{2i,j} = 0.5$ for $i \neq j$.

(3) AR(1): $\Sigma_{\mathbf{x}} = \Sigma_3$, where $\Sigma_{3i,j} = 0.5^{|i-j|}$.

For each covariance structure, we generated 100 training samples from the above model. We estimated the 50% quantiles of $y^{(1)}$ and $y^{(2)}$ and the 90% quantile of $y^{(3)}$. Model fitting and tuning were done only using the training data. SIC was used to select the regularization parameters. We also collected an independent test data set of 10,000 samples to compute the model error of each fit. The model error of the oracle is 0.497 and is independent of Σ. We repeated the simulation 100 times. Table 2 summarizes the results.

Two interesting observations can be made from Table 2. First, both SMQR and the L_1 quantile regression produce accurate models that are comparable to the oracle. SMQR is still significantly more accurate than the three L_1 quantile regression

Table 2
Simulation model 2 ($n = 100$)

$\Sigma_{\mathbf{x}}$	SMQR			L_1		
	ME	NC	NIC	ME	NC	NIC
Σ_1	0.515(0.001)	2(0)	1.87(0.17)	0.526(0.001)	2(0)	5.27(0.20)
Σ_2	0.512(0.001)	2(0)	2.23(0.18)	0.524(0.001)	2(0)	5.43(0.21)
Σ_3	0.512(0.001)	2(0)	1.40(0.14)	0.523(0.001)	2(0)	4.44(0.23)

The numbers in parenthesis are standard errors. The model error of the oracle is 0.497.

Table 3
Simulation model 3 ($n = 100$)

ρ	SMQR			L_1		
	ME	NC	NIC	ME	NC	NIC
0	0.754(0.002)	6(0)	6(0.10)	0.725(0.001)	6(0.01)	3(0.15)
0.5	0.756(0.002)	6(0.01)	5(0.12)	0.723(0.001)	6(0)	2(0.14)
0.8	0.753(0.002)	6(0.014)	6(0.15)	0.719(0.001)	6(0.03)	2(0.13)

The numbers in parenthesis are standard errors. The model error of the oracle is 0.69.

models. More importantly, SMQR produces a much sparser model than the L_1 quantile regression without missing any truly significant variable. This clearly shows the advantage of SMQR in simultaneous model selection.

As discussed in Section 3.1, simultaneous elimination in SMQR is closely connected with shrinking the maximum absolute value of the coefficient vector to exact zero. Here we show a graphical illustration of this viewpoint. In Fig. 2 we took the outputs of one simulation and plotted M_j as a function of λ (in logarithmic scale) for variables $x_3, \ldots, x_{12}$. In each panel, increasing λ drives M_j towards zero, although the curve is not always monotone. M_j becomes zero when λ reaches $\hat{\lambda}$, the tuning parameter chosen by SIC, then variable x_j is excluded from all three quantile functions.

We have seen from the previous two examples that if these multiple quantile functions share the same significant variables, then SMQR can do much better than the naive L_1 method. To see a complete picture, we also want to see the performance of SMQR when the assumption is violated. For that, we present the third simulation example. We generated three responses from the following model

$$y^{(1)} = 3x_1 + 2x_2 + \epsilon^{(1)},$$
$$y^{(2)} = 2x_4 + 3x_5 + \epsilon^{(2)},$$
$$y^{(3)} = 2x_3 + x_6 + \epsilon^{(3)},$$

where $\epsilon^{(1)}$, $\epsilon^{(2)}$ and $\epsilon^{(3)}$ are independent samples from $N(0, 1)$. Obviously, the true models do not have any common significant variable at all.

The used predictors are $\mathbf{x} = (x_1, x_2, x_3, x_4, x_5, x_6, x_7, x_8, x_9, x_{10}, x_{11}, x_{12})^{\mathrm{T}}$ generated from a multivariate normal distribution such that the pairwise correlation between x_i and x_j is $0.5^{|i-j|}$ and each x_j has variance 1. We generated 100 training samples from the above model. We estimated the 50% quantiles of all three responses. We also collected an independent test data set of 10,000 samples to compute the model error of each fit. We repeated the simulation 100 times. Table 3 summarizes the results. We see that fitting individual L_1 models has better prediction accuracy. SMQR correctly selects the six true variables needed for explaining the three quantile functions, but it tends to select more insignificant variables than the L_1 method does. Overall, fitting multiple L_1-penalized quantile regression models work better in this simulation example.

5. Real data

In this section we apply the developed penalized multiple quantiles regression method to analyze the cardiomyopathy data. The response variable in this study is a G protein-coupled receptor, designated Ro1. When the receptor is over-expressed in the heart of adult mice, the mice develop a lethal dilated cardiomyopathy that has many hallmarks of human disease. The mice recover when the expression of the receptor is turned off (Segal et al., 2003). The goal of the study is to investigate the association between the changes in gene expression and the expression of Ro1. Thirty-two mice were tested in the study (Redfern et al., 2000). To determine which changes in gene expression were due to the expression of the Ro1 transgene, Segal et al. (2003) suggested to first find the top 50 genes that correlate with the Ro1 expression profile and then use these genes to fit a linear model to predict the gene expression of Ro1. The rationale is that genes that can explain Ro1 expression profile are potential candidates to provide additional therapeutic targets and clues to the mechanism of the disease. Although simple and ad hoc, this filtering strategy has been widely used in practice. A recent paper by Fan and Lv (2008) provided some theoretical justification for filtering. They showed that for the gene selection purpose it is appropriate to do the filtering step before fitting the linear model, because the filtering will only eliminate irrelevant genes and keep informative genes with very high probability. In this work we do not argue whether the filtering is the best strategy in

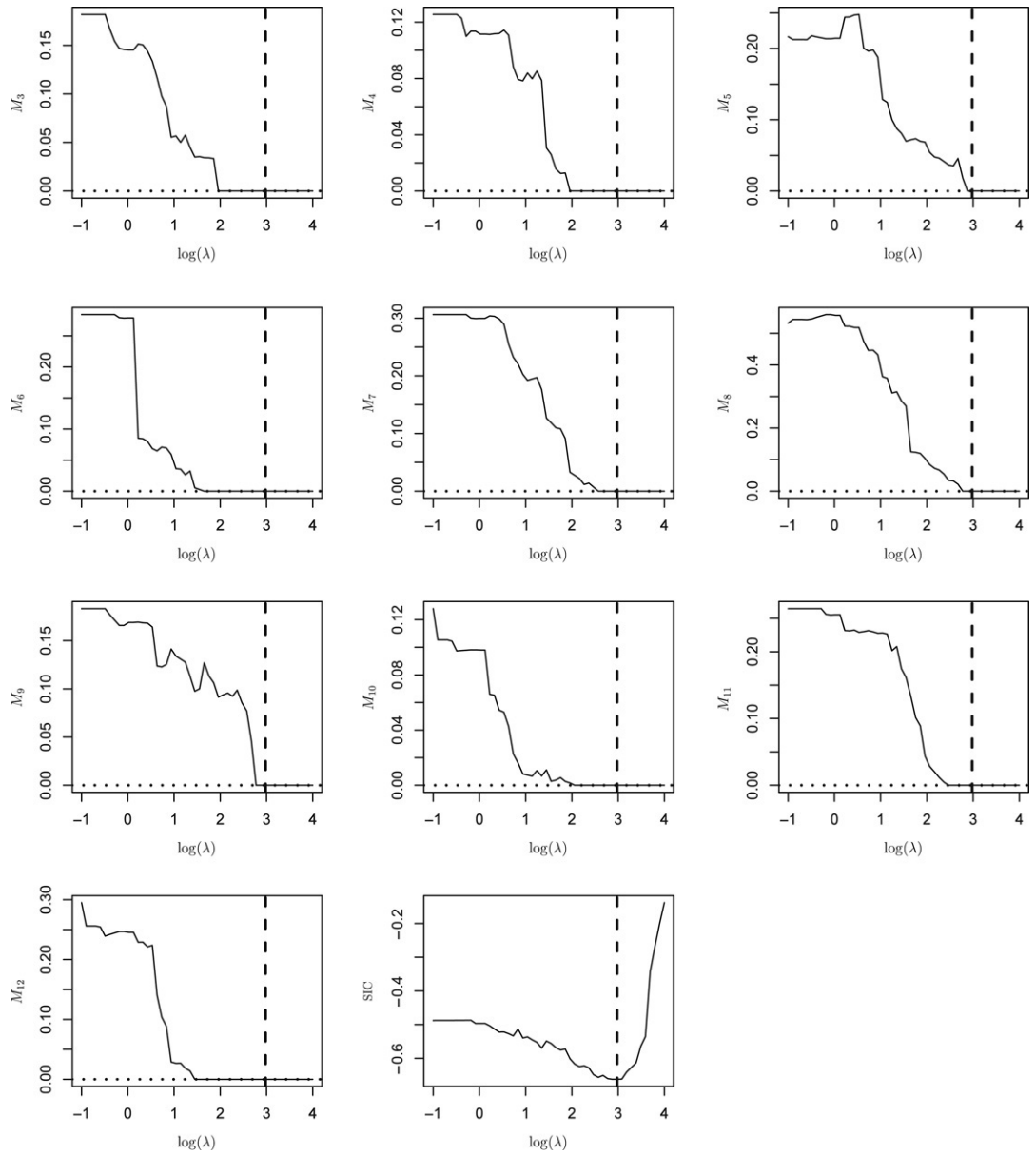

Fig. 2. SMQR in simulation model 2 (independent predictors). We plot the maximum absolute value of the coefficients of each variable. In each panel the thick broken line indicates the optimal λ value selected by SIC.

analyzing the cardiomyopathy data. Instead, we shall focus on the choice of regression analysis methods. In the following discussion we try to build a linear model based on these 50 genes. In our analysis we used 5-fold cross-validation to select the regularization parameter of each regression method.

We first fitted a lasso regression model and selected 17 genes. We then followed the quantile regression philosophy to estimate multiple conditional quantile functions of Ro1 in order to obtain a more complete description of the stochastic relationship between Ro1 and these 50 genes. Following the usual practice in quantile regression we considered simultaneously estimating the 10%, 25%, 50%, 75% and 90% conditional quantiles of Ro1. Using the L_1 quantile regression, we separately fitted the 10%, 25%, 50%, 75% and 90% conditional quantiles. The selected genes were different in different models and in total 38 genes were selected. Our simulation has shown that under the linear regression model assumption, separately fitting multiple L_1 quantile regression models can be very conservative in variable selection. Therefore, we tend to think the variable selection result by the lasso mean regression is more trustworthy. On the other hand, we still wish to find a small set of genes to model the 10%, 25%, 50%, 75% and 90% conditional quantiles of Ro1. To this end, we fitted the SMQR model and selected 12 genes. It is interesting to see that the 12 genes selected by SMQR are a subset of the 38 genes selected by the five L_1 quantile regression models and the 17 genes selected by the lasso mean regression model. Table 4

Table 4
Genes selected by SMQR

ID	Mu6500probeset	GeneBank	Description
128	Msa.10108.0	W46084	Homologous to sp P07814: MULTIFUNCTIONAL AMINOACYL
194	Msa.1043.0	L14751	Mouse ATP receptor (P2u) mRNA, complete cds
702	Msa.1293.0	L04961	Mouse Xist (X inactive specific transcript) mRNA
1210	Msa.15442.0	AA044561	Homologous to sp P07379: PHOSPHOENOLPYRUVATE CARBO
2375	Msa.2134.0	U25708	Murine mRNA for 4F2 antigen heavy chain
3178	Msa.2546.0	X55126	M.musculus Zfp-29 gene for zinc finger protein
3409	Msa.26751.0	AA064467	Homologous to sp Q07244: HETEROGENEOUS NUCLEAR RIB
3758	Msa.2877.0	D31717	Mouse MARib mRNA for ribophorin, complete cds
4015	Msa.30232.0	AA096793	Homologous to sp P05141: ADP,ATP CARRIER PROTEIN
5580	Msa.5707.0	W97077	Homologous to sp P33150: T-CADHERIN PRECURSOR
6007	Msa.778.0i	U73744	Mus musculus heat shock 70 protein (Hsc70) gene
6355	Msa.964.0	M20985	Mouse MHC class I H2-Qa-Mb1 gene, complete cds

shows the list of genes selected by SMQR and their descriptions. We also repeated the analysis using the 10%, 20%, 30%, 40%, 50%, 60%, 70%, 80%, 90% quantiles and obtained similar conclusions. Hence the SMQR analysis seems reliable. Due to space considerations, these results are not shown here.

6. Conclusion

In this paper we have proposed the a new regularization technique to perform regularized simultaneous model selection in multiple quantiles regression. We have demonstrated the promising performance of the proposed method using simulated and real data. It is also worth noting that SMQR provides a unified solution to handle three different multiple quantiles regression problems: (1) multiple quantiles of a single response; (2) the same quantile of multiple responses; and (3) multiple quantiles of multiple responses. Finally, one should only use SMQR to perform simultaneous model selection when the goal of the study is to select the same subset of variables in multiple quantile regression models or it is reasonable to assume the same subset of variables appear in the multiple quantile functions. Otherwise, fitting multiple L_1-norm quantile regression models should be considered, as suggested by the third simulation example in Section 4.

Acknowledgements

We thank Professor Mark Segal for kindly providing us the cardiomypathy data. The authors sincerely thank an AE and two referees for their helpful comments that substantially improved an earlier version of this paper. This work is supported by NSF grant DMS-0706733.

References

Breiman, L., Friedman, J., 1997. Predicting multiple responses in multiple linear regression (with discussion). Journal of the Royal Statistical Society: Series B 59, 3–54.
Fan, J., Li, R., 2001. Variable selection via nonconcave penalized likelihood and its oracle properties. Journal of the American Statistical Association 96, 1348–1360.
Fan, J., Lv, J., 2008. Sure independence screening for ultra-high dimensional feature space (with discussion). Journal of the Royal Statistical Society, Series B (in press).
Hastie, T., Tibshirani, R., Friedman, J., 2001. The Elements of Statistical Learning; Data mining, Inference and Prediction. Springer Verlag, New York.
He, X., Ng, P., Portnoy, S., 1998. Bivariate quantile smoothing splines. Journal of the Royal Statistical Society, Series B 60, 537–550.
Koenker, R., 2004. Quantile regression for longitudinal data. Journal of Multivariate Analysis 91, 74–89.
Koenker, R., 2005. Quantile Regression. In: Econometric Society Monograph Series. Cambridge University Press.
Koenker, R., Bassett, G., 1978. Regression quantiles. Econometrica 4, 33–50.
Koenker, R., Geling, R., 2001. Reappraising medfly longevity: A quantile regression survival analysis. Journal of the American Statistical Association 96, 458–468.
Koenker, R., Hallock, K., 2001. Quantile regression. Journal of Economic Perspectives 15, 143–156.
Koenker, R., Ng, P., Portnoy, S., 1994. Quantile smoothing splines. Biometrika 81, 673–680.
Li, Y., Liu, Y., Zhu, J., 2007. Quantile regression in reproducing kernel hilbert spaces. Journal of the American Statistical Association 102, 255–268.
Li, Y., Zhu, J., 2008. The l_1 norm quantile regression. Journal of Computational and Graphical Statistics 17, 163–185.
Meyer, M., Woodroofe, M., 2000. On the degrees of freedom in shape-restricted regression. The Annals of Statistics 28, 1083–1104.
Redfern, C., Degtyarev, M., Kwa, A., Salomonis, N., Cotte, N., Nanevicz, T., Fidelman, N., Desai, K., Vranizan, K., Lee, E., Coward, P., Shah, N., Warrington, J., Fishman, G., Bernstein, D., Baker, A., Conklin, B., 2000. Conditional expression of a gi-coupled receptor causes ventricular conduction delay and a lethal cardiomyopathy. PNAS 97, 4826–4831.
Segal, M., Kam, D., Bruce, C., 2003. Regression approaches for microarray data analysis. Journal of Computational Biology 10 (6), 961–980.
Tibshirani, R., 1996. Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society, Series B 58, 267–288.
Turlach, B., Venables, W., Wright, S., 2005. Simultaneous variable selection. Technometrics 47, 349–363.
Yuan, M., 2006. GACV for quantile smooting splines. Computational Statistics and Data Analysis 50, 813–829.
Yuan, M., Lin, Y., 2006. Model selection and estimation in regression with grouped variables. Journal of the Royal Statistical Society, Series B 68, 49–67.
Zhang, H., Liu, Y., Wu, Y., Zhu, J., 2008. Variable selection for multicategory svm via sup-norm regularization. Electronic Journal of Statistics 2, 149–167.
Zou, H., Yuan, M., 2008. The F_∞-norm support vector machine. Statisitca Sinica 18, 379–398.