Statistics and Its Interface Volume 2 (2009) 299–310

Stepwise multiple quantile regression estimation using non-crossing constraints*

Yichao Wu[†,§] and Yufeng Liu[‡]

Quantile regression is an important statistical tool for statistical modeling. It has been widely used in various fields including econometrics, medicine, and bioinformatics. Despite its popularity in practice, individually estimated quantile regression functions often cross each other and consequently violate the basic properties of quantiles. In this paper we propose a new method for estimating multiple quantile regression functions without crossing. Both linear and kernel quantile regression models are considered. Several numerical examples are presented to illustrate competitive performance of the proposed method.

AMS 2000 subject classifications: Primary 62J05, 62G08; secondary 62H12.
Keywords and phrases: Constraints, non-crossing, quantile regression, RKHS, variable selection.

1. INTRODUCTION

Quantile regression is a very useful statistical tool for estimating conditional quantile regression functions. It has been intensively studied after its introduction by Koenker and Bassett (1978). Examples include Koenker and Hallock (2001), Yu, Lu and Stander (2003). Its wide applications vary from medicine (Cole and Green 1992, Heagerty and Pepe 1999), to survival analysis (Yang 1999, Koenker and Geling 2001), and to economics (Hendricks and Koenker 1992, Koenker and Hallock 2001). We refer the readers to a recent book on this subject by Koenker (2005) to get a more complete review on quantile regression.

In many situations, it is useful to estimate multiple quantile regression functions. Despite the flexibility of individual estimation of these curves, an embarrassing phenomenon of quantile crossing may occur. Such a kind of quantile crossing violates the basic principle of distribution functions so that their associated inverse functions should be monotone increasing. Although this phenomenon typically only occurs in outlying regions of the input space when the observations are scarce, it is nevertheless an undesirable phenomenon for utilization and interpretation of these quantile regression functions. He (1997) proposed the location-scale shift model to impose monotonicity across the quantile functions. However, as noted by Neocleousa and Portnoy (2007), even for linear regression quantiles, corresponding models can be much more general. Thus, a more general development of non-crossing regression quantiles is needed. For kernel quantile regression, Takeuchi, Le, Sears and Smola (2006) proposed to impose non-crossing constraints on the data points. Although the approach can help to reduce the chance of crossing, the dimension of the optimization problem for simultaneous multiple quantile estimation can be large for certain applications.

In this paper, we propose a new method to estimate multiple quantile regression functions without crossing. Our estimation scheme is in a stepwise fashion to ensure non-crossing of the regression functions. In particular, with the current quantile regression function at a particular given level, we add constraints in the estimation procedure to ensure the next quantile regression function does not cross the current one. The procedure continues till quantile regression functions at all desired levels are obtained.

Both linear and kernel quantile regression models are considered. Our numerical examples show that the non-crossing constraints can not only help to obtain more interpretable quantile functions, but also help to improve the estimation accuracy of the resulting regression functions.

The remainder of this article is organized as follows. In Section 2, we give a brief review of quantile regression. In Section 3, we illustrate our proposed non-crossing estimation scheme for multiple quantile regression functions in a stepwise fashion. An extension for the setting of regularization is given in Section 4. Several simulated examples in different settings are presented in Section 5, followed by a real data example in Section 6. Some discussion is given in Section 7.

2. QUANTILE REGRESSION

Suppose we have a set of training data $(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_n, y_n)$ with the input $\boldsymbol{x}_i \in \boldsymbol{\mathcal{X}} \subset \mathbb{R}^p$ and the output $y_i \in \mathbb{R}$. We would like to recover the $100\tau\%$ quantile of the conditional distribution of Y given $\boldsymbol{X} = \boldsymbol{x}$ for $\tau \in (0, 1)$.

*The authors would like to thank the editor, the associate editor, and reviewers for their constructive comments and suggestions.
†Wu's research was support in part by NSF grant DMS-0905561.
‡Liu's research was supported in part by NSF grants DMS-0606577 and DMS-0747575.
§Corresponding author.

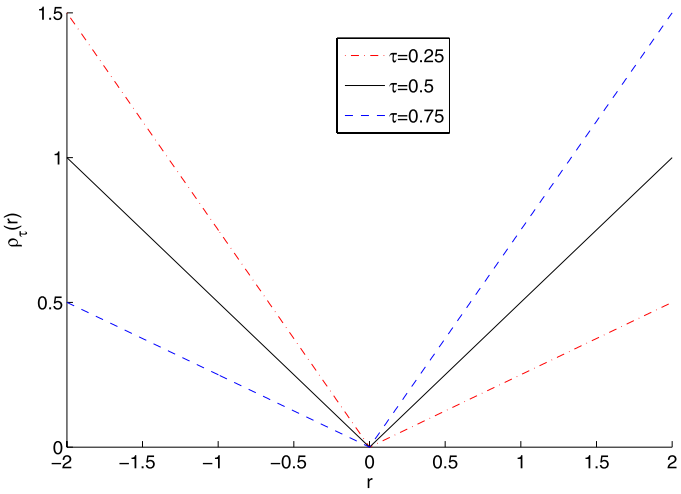

Figure 1. Plot of the check function for three different values of τ.

The conditional τ-th quantile function $f_\tau(\boldsymbol{x})$ is defined such that $P(Y \le f_\tau(\boldsymbol{X})|\boldsymbol{X} = \boldsymbol{x}) = \tau$. By tilting the absolute loss function, Koenker and Bassett (1978) introduced the check function which is defined by

$$\rho_\tau(r) = \begin{cases} \tau r & \text{if } r > 0 \\ -(1-\tau)r & \text{otherwise.} \end{cases}$$

An illustrating plot with several different values of τ is given in Fig. 1. Note that the check function generalizes the absolute loss used in least absolute deviation regression from $\tau = 0.5$ to any value in $(0, 1)$.

In their seminal paper, Koenker and Bassett (1978) demonstrated that the τ-th conditional quantile function can be estimated by solving the following minimization problem

$$\min_{f_\tau \in \mathcal{F}} \sum_{i=1}^{n} \rho_\tau(y_i - f_\tau(\boldsymbol{x}_i)). \tag{1}$$

To avoid over-fitting and improve generalization ability, as in Koenker, Ng, and Portnoy (1994), one can consider the penalized version of (1) in the following regularization framework

$$\min_{f_\tau \in \mathcal{F}} \sum_{i=1}^{n} \rho_\tau(y_i - f_\tau(\boldsymbol{x}_i)) + \lambda J(f_\tau), \tag{2}$$

where $\lambda \ge 0$ is the regularization parameter and $J(f_\tau)$ denotes the roughness penalty of the function $f_\tau(\cdot)$.

Computation of (1) can be carried out by standard linear programming (LP). For the regularized version (2), the optimization tool depends on the choice of the penalty function $J(f_\tau)$. For instance, when we use the L_1 penalty as in the LASSO (Tibshirani, 1996), we can implement it using LP as for (1). In the context of least absolute deviation regression (a special case of quantile regression with $\tau = 0.5$), a similar consideration was given in Wang, Li, and Jiang (2007). When we use the L_2 penalty as in the ridge regression, (2) can be solved using quadratic programming (QP).

To further improve the computation efficiency and facilitate the choice of the tuning parameter λ in (2), Li and Zhu (2008) and Li et al. (2007) developed entire solution paths of linear and kernel regularized quantile regression with respect to λ for any fixed $\tau \in (0, 1)$ respectively. The path algorithm helps to speed up the computation and simplifies the tuning parameter selection.

Despite the success of quantile regression for estimating individual conditional quantile regression functions, problems may occur when multiple quantile functions are needed at the same time. In particular, the separately estimated multiple quantile functions may cross each other. In Section 3, we propose to estimate multiple quantile functions using non-crossing constraints. Our proposed method can not only help to obtain non-crossing quantile functions, but also help to improve the estimation accuracy of the resulting functions.

3. MULTIPLE NON-CROSSING QUANTILE REGRESSION ESTIMATION

Of particular interest is to estimate simultaneous quantile functions. Out goal is to estimate multiple non-crossing quantile functions with different values of τ. For simplicity, we first focus on linear quantile functions. An extension to nonlinear functions using kernel methods is discussed in Section 4.

Suppose we want to estimate quantile functions simultaneously at $0 < \tau_1 < \tau_2 < \cdots < \tau_K < 1$. Then we need to estimate K sets of coefficients b_k and $\boldsymbol{\beta}_k = (\beta_{1k}, \beta_{2k}, \ldots, \beta_{pk})^T$ for $k = 1, 2, \ldots, K$. The theoretical quantile functions $f_k(\boldsymbol{x}) = b_k + \boldsymbol{\beta}_k^T \boldsymbol{x}$ should satisfy that

$$f_k(\boldsymbol{x}) < f_{k+1}(\boldsymbol{x}) \quad \text{for } k = 1, 2, \ldots, K-1 \text{ and } \forall \boldsymbol{x} \in \boldsymbol{\mathcal{X}}. \tag{3}$$

To incorporate this constraint into our estimation scheme to ensure non-crossing, we assume that each predictor variable has a bounded support and without loss of generality consider $X_j \in [0, 1]$, for $j = 1, 2, \ldots, p$. Here our predictor vector is $\boldsymbol{X} = (X_1, X_2, \ldots, X_p)^T$. In this case (3) is automatically satisfied if the constraint is satisfied at all vertices, i.e.,

$$f_k(\boldsymbol{x}) < f_{k+1}(\boldsymbol{x}) \quad \text{for } k = 1, 2, \ldots, K-1 \text{ and } \forall \boldsymbol{x} \in \{0, 1\}^p. \tag{4}$$

3.1 Naive constrained estimation

Our estimation scheme is in a stepwise fashion. Specifically, given the current quantile function, we estimate the

next quantile function so that it does not cross with the existing quantile. Naturally, there are several different ways to proceed with the estimation depending on the direction of the stepwise procedure.

Complete Up CU(k):

Denote our current estimated coefficients for the $100\tau_k$-th quantile function by $\hat{b}_k$ and $\hat{\boldsymbol{\beta}}_k$. Our non-crossing quantile regression solves the following optimization problem to estimate the coefficients for the $100\tau_{k+1}$-th quantile function:

$$\min_{b_{k+1},\boldsymbol{\beta}_{k+1}} \sum_{i=1}^{n} \rho_{\tau_{k+1}}(y_i - b_{k+1} - \boldsymbol{\beta}_{k+1}^T \boldsymbol{x}_i) \tag{5}$$
$$\text{s.t.} \quad \hat{b}_k + \hat{\boldsymbol{\beta}}_k^T \boldsymbol{x} + \delta_0 \le b_{k+1} + \boldsymbol{\beta}_{k+1}^T \boldsymbol{x}, \forall \boldsymbol{x} \in \{0,1\}^p,$$

where δ_0 is some pre-specified small positive number introduced to ensure strict inequality in (4) and can be chosen as the numerical precision level. In our numerical study, we set $\delta_0 = 10^{-4}$.

Complete Down CD(k):

Similar to the complete up version, we can estimate b_{k-1} and $\boldsymbol{\beta}_{k-1}$ based on $\hat{b}_k$ and $\hat{\boldsymbol{\beta}}_k$ by solving

$$\min_{b_{k-1},\boldsymbol{\beta}_{k-1}} \sum_{i=1}^{n} \rho_{\tau_{k-1}}(y_i - b_{k-1} - \boldsymbol{\beta}_{k-1}^T \boldsymbol{x}_i) \tag{6}$$
$$\text{s.t.} \quad b_{k-1} + \boldsymbol{\beta}_{k-1}^T \boldsymbol{x} \le \hat{b}_k + \hat{\boldsymbol{\beta}}_k^T \boldsymbol{x} - \delta_0 \text{ for } \forall \boldsymbol{x} \in \{0,1\}^p.$$

The number of constraints in optimization problems (5) and (6) is 2^p, which can be large for moderate p. For more efficient implementation in practice, we need to reduce the number of constraints.

3.2 Improved constrained estimation

To improve the naive constrained estimation scheme, we note that most of the 2^p vertex constraints are redundant. Thus, we can greatly reduce the number of constraints. Here we propose a more efficient iterative method.

Simplified Up SU(k):

Without loss of generality, we consider the estimation of b_{k+1} and $\boldsymbol{\beta}_{k+1}$ based on $\hat{b}_k$ and $\hat{\boldsymbol{\beta}}_k$. First solve the regular quantile function estimation

$$\min_{b_{k+1},\boldsymbol{\beta}_{k+1}} \sum_{i=1}^{n} \rho_{\tau_{k+1}}(y_i - b_{k+1} - \boldsymbol{\beta}_{k+1}^T \boldsymbol{x}_i)$$

and denote the current solution by $\tilde{b}_{k+1}$ and $\tilde{\boldsymbol{\beta}}_{k+1}$. The vertex that most likely violates the non-crossing constraints is given by $\tilde{\boldsymbol{x}} = (\tilde{x}_1, \tilde{x}_2, \ldots, \tilde{x}_p)^T$ with $\tilde{x}_j = (1 - \text{sign}(\tilde{\beta}_{(k+1)j} - \hat{\beta}_{kj}))/2$, $j = 1, 2, \ldots, p$. Set $C = \{\tilde{\boldsymbol{x}}\}$. If the constraint is violated at $\tilde{\boldsymbol{x}}$, i.e., $\hat{b}_k + \hat{\boldsymbol{\beta}}_k^T \tilde{\boldsymbol{x}} > \tilde{b}_{k+1} + \tilde{\boldsymbol{\beta}}_{k+1}^T \tilde{\boldsymbol{x}}$, we solve the following optimization problem

$$\min_{b_{k+1},\boldsymbol{\beta}_{k+1}} \sum_{i=1}^{n} \rho_{\tau_{k+1}}(y_i - b_{k+1} - \boldsymbol{\beta}_{k+1}^T \boldsymbol{x}_i) \tag{7}$$
$$\text{s.t.} \quad \hat{b}_k + \hat{\boldsymbol{\beta}}_k^T \boldsymbol{x} + \delta_0 \le b_{k+1} + \boldsymbol{\beta}_{k+1}^T \boldsymbol{x}, \forall \boldsymbol{x} \in C,$$

and denote its solution by $\check{b}_{k+1}$ and $\check{\boldsymbol{\beta}}_{k+1}$. Define $\check{\boldsymbol{x}}$ to be the vertex that most likely violates the non-crossing constraint. If $\check{\boldsymbol{x}}$ does violate the non-crossing constraint, we add it to the set C, i.e., $C = C \cup \check{\boldsymbol{x}}$ and solve (7) with the updated set C. We continue the iteration until the non-crossing constraints are satisfied.

A similar scheme can be carried out for estimating b_{k-1} and $\boldsymbol{\beta}_{k-1}$ based on $\hat{b}_k$ and $\hat{\boldsymbol{\beta}}_k$ and is called *Simplified Down* SD(k) accordingly.

3.3 Why from the middle

An important issue is the choice of the starting quantile function without any constraint. Our proposal is to start from the middle, that is the quantile function with $\tau = 0.5$. To further demonstrate this choice, we note that the asymptotic variance of quantile estimator is proportional to $\tau(1-\tau)/f(F^{-1}(\tau))$, where $f(\cdot)$ and $F(\cdot)$ are the pdf and cdf of the error distribution respectively. For the normal distribution, we can show that $\tau(1-\tau)/f(F^{-1}(\tau))$ is minimized at $\tau = 0.5$. Thus, the estimated quantile function with $\tau = 0.5$ is relatively more accurate than other estimated quantiles. It is reasonable to begin with $\tau = 0.5$ and use constraints to estimate other non-crossing quantiles.

3.4 Full estimation scheme

Assume that, out of $\tau_1, \tau_2, \ldots,$ and τ_K, τ_{k_0} is the one that is closest to $\tau = 0.5$.

Scheme 1:

Use the standard quantile regression to estimate b_{k_0} and $\boldsymbol{\beta}_{k_0}$ and denote them by $\hat{b}_{k_0}^{(1)}$ and $\hat{\boldsymbol{\beta}}_{k_0}^{(1)}$. Use SU to sequentially estimate b_k and $\boldsymbol{\beta}_k$ one after another for $k = k_0 + 1, \ldots, K$ and use SD to sequentially estimate b_k and $\boldsymbol{\beta}_k$ one after another for $k = k_0 - 1, \ldots, 1$. These estimates are denoted by $\hat{b}_k^{(1)}$ and $\hat{\boldsymbol{\beta}}_k^{(1)}$ for $k = 1, 2, \ldots, K$.

One may use the solution obtained by Scheme 1 as the final solution. Alternatively, one can use the solutions in Scheme 1 and perform additional updating. According to our limited numerical experience, Scheme 1 can be improved using the following scheme.

Scheme 2(U):

Set $\hat{b}_1^{(2U)} = \hat{b}_1^{(1)}$ and $\hat{\boldsymbol{\beta}}_1^{(2U)} = \hat{\boldsymbol{\beta}}_1^{(1)}$. Beginning with $\hat{b}_1^{(2U)}$ and $\hat{\boldsymbol{\beta}}_1^{(2U)}$, we apply SU sequentially to get updated estimates $\hat{b}_k^{(2U)}$ and $\hat{\boldsymbol{\beta}}_k^{(2U)}$ for $k = 2, \ldots, K$.

Scheme 2(D):

Set $\hat{b}_K^{(2D)} = \hat{b}_K^{(1)}$ and $\hat{\boldsymbol{\beta}}_K^{(2D)} = \hat{\boldsymbol{\beta}}_K^{(1)}$. Beginning with $\hat{b}_K^{(2D)}$ and $\hat{\boldsymbol{\beta}}_K^{(2D)}$, we apply SD sequentially to get updated estimates $\hat{b}_k^{(2D)}$ and $\hat{\boldsymbol{\beta}}_k^{(2D)}$ for $k = K-1, \ldots, 1$.

Our final estimates are given by averaging, i.e., $\hat{b}_k = (\hat{b}_k^{(2D)} + \hat{b}_k^{(2U)})/2$ and $\hat{\boldsymbol{\beta}}_k = (\hat{\boldsymbol{\beta}}_k^{(2D)} + \hat{\boldsymbol{\beta}}_k^{(2U)})/2$ for $k = 1, 2, \ldots, K$.

4. NON-CROSSING KERNEL QUANTILE FUNCTIONS

Our stepwise scheme presented in Section 3 can be directly extended to the regularized version (2). Essentially, we need to add non-crossing constraints to the optimization problem (2) as in the linear case. In this section, we discuss nonlinear quantile estimation using the kernel trick.

For any given positive definite kernel function $K(\cdot, \cdot)$ on $\boldsymbol{\mathcal{X}}$, kernel quantile regression can be carried out by solving

$$(8) \quad \min_{b, \alpha_1, \ldots, \alpha_n} \sum_{i=1}^{n} \rho_\tau(y_i - b - \sum_{j=1}^{n} \alpha_j K(\boldsymbol{x}_i, \boldsymbol{x}_j)) + \lambda \boldsymbol{\alpha}^T \mathbf{K} \boldsymbol{\alpha},$$

where $\boldsymbol{\alpha}$ is a vector of length n with the i-th element α_i, and $\mathbf{K}$ is a matrix of $n \times n$ with the ij-th element $(\boldsymbol{x}_i, \boldsymbol{x}_j)$. The estimated quantile function is then given by $\hat{f}_\tau(\boldsymbol{x}) = \hat{b} + \sum_{i=1}^{n} \hat{\alpha}_i K(\boldsymbol{x}, \boldsymbol{x}_i)$.

For our stepwise non-crossing procedure, we first assume that the estimated fitting for the $100\tau_k$-th quantile is given by $\hat{f}_{\tau_k}(\boldsymbol{x}) = \hat{b}_k + \sum_{i=1}^{n} \hat{\alpha}_{ki} K(\boldsymbol{x}, \boldsymbol{x}_i)$. The corresponding non-crossing version is to solve

$$(9) \quad \begin{aligned} \min_{b_{k+1}, \boldsymbol{\alpha_{(k+1)}}} \; & \lambda \boldsymbol{\alpha}_{(k+1)}^T \mathbf{K} \boldsymbol{\alpha}_{(k+1)} \\ & + \sum_{i=1}^{n} \rho_{\tau_{k+1}}(y_i - b_{k+1} - \sum_{j=1}^{n} \alpha_{(k+1)j} K(\boldsymbol{x}_i, \boldsymbol{x}_j)) \\ s.t. \quad & \hat{b}_k + \sum_{j=1}^{n} \hat{\alpha}_{kj} K(\boldsymbol{x}_i, \boldsymbol{x}_j) + \delta_0 \\ & \leq b_{k+1} + \sum_{j=1}^{n} \alpha_{(k+1)j} K(\boldsymbol{x}_i, \boldsymbol{x}_j) \text{ for } i = 1, 2, \ldots, n, \end{aligned}$$

where $\boldsymbol{\alpha}_{(k+1)}$ is a n-dimensional vector with its i-th element $\alpha_{(k+1)i}$. Our iterative schemes described in Section 3 can be then applied to the proposed kernel quantile regression.

It is worth noting that the constraints used in the kernel formulation (9) are different from the linear formulation. In the linear case, we use vertices of the standard input d-dimensional cube to form the constraints. This is not directly applicable to the kernel case. In fact, it is difficult to enforce non-crossing for the kernel case in the entire input space. In our formulation (9), we simplify the requirement and enforce non-crossing over the domain spanned by features $(K(\boldsymbol{x}_i, \boldsymbol{x}_1), \ldots, K(\boldsymbol{x}_i, \boldsymbol{x}_n))^T$, $i = 1, 2, \ldots, n$. In the literature, Takeuchi et al. (2006) also proposed to use non-crossing constraints on the data points for kernel quantile regression. They aim to solve for one optimization problem to obtain multiple quantile functions. When both the number of quantile functions desired and the dimension of covariates are large, the optimization problem can be computationally intensive. Our stepwise procedure is computationally simpler since it solves multiple smaller optimization problems.

5. SIMULATION STUDIES

In this section, we use simulation studies to illustrate improvement of our new non-crossing multiple quantile estimation by comparing it to the naive individual estimates, the method proposed by He (1997), and the method of Takeuchi et al. (2006). We generate training samples of size n. An identically distributed test set is generated to report numerical summaries to compare different methods.

For kernel learning, we generate an independent and identically distributed sample of size n to tune the regularization parameter. The Gaussian kernel $K(\boldsymbol{x}_1, \boldsymbol{x}_2) = \exp(- \parallel \boldsymbol{x}_1 - \boldsymbol{x}_2 \parallel_2^2 / \sigma^2)$ is used to achieve the nonlinear quantile estimation.

Five different examples are presented in this section. Examples 1 and 2 are devoted to show quantile crossing for linear and nonlinear cases, respectively. We use two different linear models (Examples 3 and 4) to compare the performance of our new non-crossing multiple quantile estimation, the naive individual estimation, and He (1997)'s method. Example 3 has independent and identically distributed errors while Example 4 involves a location-scale model. Example 5 presents a nonlinear example with independent and identically distributed errors.

Example 1. We use a simple linear example to demonstrate that naive individual estimates may suffer from quantile crossing which the new proposed method can avoid. With $p = 1$, the univariate predictor X is generated from the standard uniform distribution, i.e., $X \sim \mathit{Uniform}[0, 1]$. Response Y is generated by the simplest linear model $Y = X + \epsilon$, where $\epsilon \sim N(0, 1)$ is independent of the one dimensional predictor variable X. The sample size is fixed at $n = 100$. Quantile functions are estimated at $\tau = 0.05, 0.1, \ldots, 0.95$. The naive individually estimated quantile functions are plotted in the left panel of Fig. 2 while our non-crossing estimates are given on the right panel. We can clearly see that the naive individual estimates suffer from quantile crossing. However by enforcing our non-crossing constraint, our new estimates do not cross each other.

Example 2. A nonlinear example is presented to demonstrate quantile crossing for the naive individual estimates. The predictor X is univariate and uniformly distributed over $[0, 1]$. Conditional on X, the response is generated by

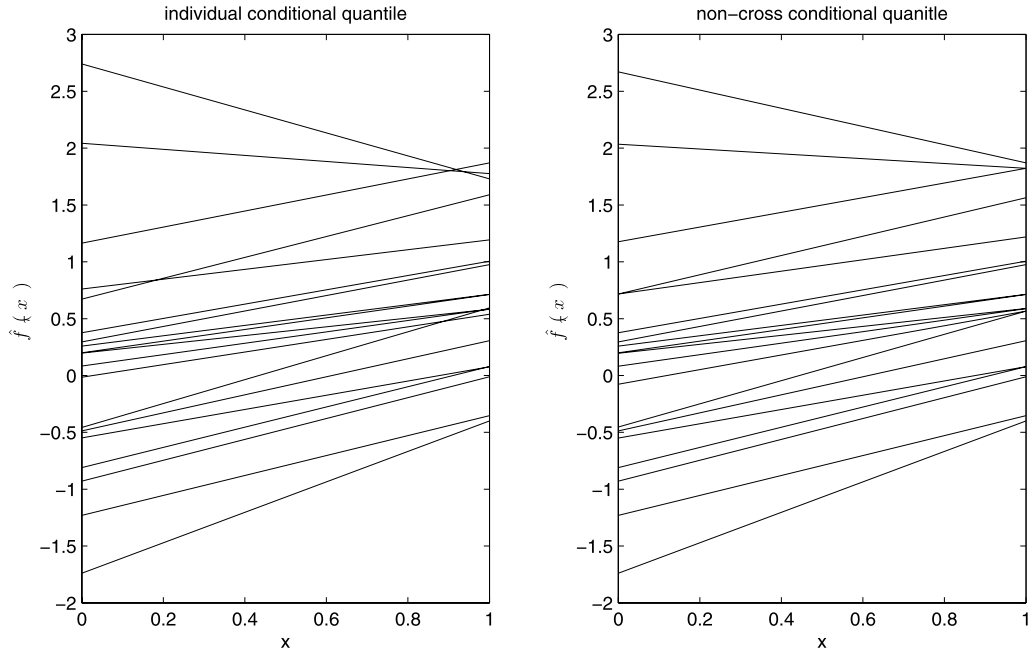

Figure 2. Comparison of naive individually estimated quantile functions and the new non-crossing quantile estimates for Example 1: the left panel plots the naive individually estimated quantile functions; the right panel shows the new non-crossing quantile estimates.

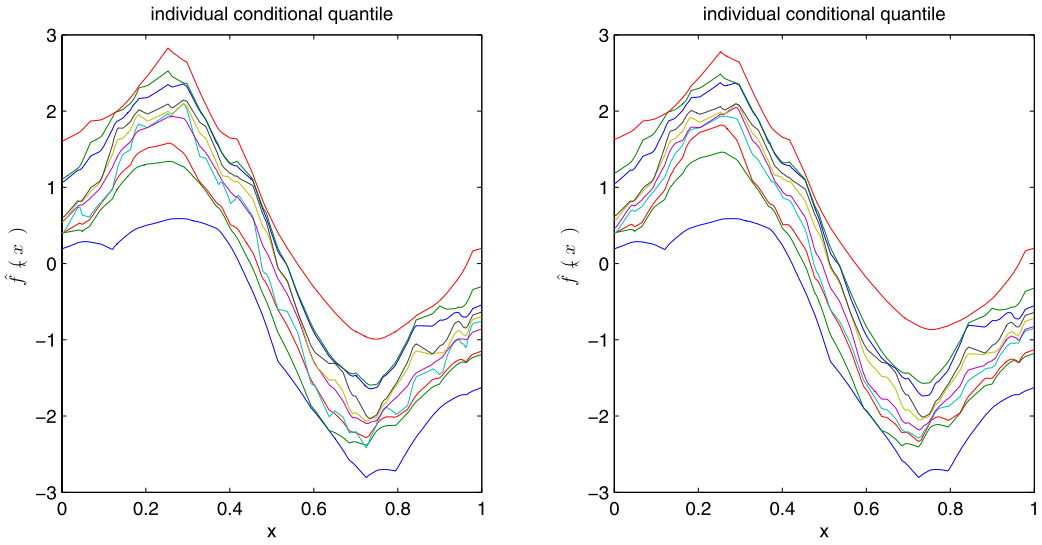

Figure 3. Comparison of naive individually estimated conditional quantile functions and our non-crossing estimates for Example 2: the left panel plots the individually estimated conditional quantile functions; the right panel shows our non-crossing estimates.

$Y = \sin(2\pi X) + 0.5\epsilon$ with independent standard normal random error ϵ. The sample size is fixed at $n = 100$. Nonlinear quantile function is estimated using the Gaussian kernel $K(x_1, x_2) = e^{-(x_1-x_2)^2/\sigma^2}$ with $\sigma^2 = 0.4$ for simplicity. An independent and identically distributed tuning set of size n is generated to select the tuning parameter λ for both our non-crossing quantile estimation and the naive individual estimation. Conditional quantile functions are estimated at $\tau = 0.05, 0.1, 0.15, \ldots, 0.95$.

Figure 3 plots estimated quantile functions for $\tau = 0.1, 0.2, \ldots, 0.9$. Similar as Example 1, quantile crossing is observed for this nonlinear example using naive estimation. However by enforcing the non-crossing constraint, we at least guarantee that our estimated conditional quantile

Table 1. Average of the differences between test errors and Bayes errors over 200 repetitions for different τ's and different estimation methods in Example 3. M1 corresponds to the standard quantile regression. M2, M3, and M4 represent our Scheme 1, Scheme 2(U), and Scheme 2(D), respectively. M5 corresponds to our final estimator by averaging. He97 refers to He (1997)'s method. Note that each table entry of error difference is enlarged by a factor of 1000. Numbers in parentheses are the corresponding standard errors

τ	M1	M2	M3	M4	M5	He97
0.05	17.87 (0.78)	10.62 (0.51)	10.62 (0.51)	10.44 (0.48)	10.46 (0.49)	22.39 (1.32)
0.10	18.67 (0.81)	12.66 (0.58)	15.78 (0.72)	12.21 (0.53)	13.36 (0.60)	24.69 (1.39)
0.15	18.58 (0.85)	13.81 (0.65)	17.59 (0.84)	13.41 (0.58)	14.45 (0.65)	24.84 (1.36)
0.20	17.81 (0.85)	14.36 (0.68)	17.75 (0.83)	13.82 (0.60)	14.49 (0.65)	23.64 (1.25)
0.25	18.49 (0.83)	15.88 (0.74)	17.78 (0.79)	14.74 (0.63)	14.88 (0.65)	22.75 (1.15)
0.30	19.31 (0.93)	16.79 (0.77)	18.02 (0.84)	15.42 (0.66)	15.23 (0.67)	21.76 (1.07)
0.35	18.81 (0.82)	17.62 (0.76)	17.67 (0.84)	15.73 (0.63)	15.26 (0.65)	20.81 (0.97)
0.40	19.15 (0.84)	17.82 (0.77)	17.64 (0.84)	15.84 (0.64)	15.34 (0.66)	20.00 (0.90)
0.45	18.90 (0.83)	18.59 (0.80)	17.34 (0.84)	16.20 (0.67)	15.35 (0.66)	19.62 (0.89)
0.50	18.59 (0.82)	18.59 (0.82)	16.42 (0.82)	16.10 (0.72)	14.79 (0.67)	18.59 (0.82)
0.55	17.93 (0.83)	17.82 (0.82)	15.80 (0.81)	16.08 (0.75)	14.51 (0.69)	18.86 (0.83)
0.60	17.78 (0.79)	16.89 (0.77)	15.21 (0.76)	16.70 (0.75)	14.56 (0.68)	18.84 (0.82)
0.65	17.85 (0.78)	16.30 (0.74)	14.67 (0.70)	16.97 (0.74)	14.48 (0.66)	19.02 (0.83)
0.70	18.02 (0.79)	15.51 (0.71)	14.44 (0.68)	16.64 (0.70)	14.25 (0.63)	19.77 (0.85)
0.75	17.84 (0.76)	14.62 (0.66)	13.71 (0.58)	16.92 (0.72)	14.13 (0.59)	20.58 (0.90)
0.80	17.22 (0.72)	14.18 (0.62)	13.62 (0.58)	16.35 (0.71)	13.94 (0.60)	22.11 (0.98)
0.85	16.82 (0.69)	13.19 (0.56)	12.94 (0.51)	16.16 (0.69)	13.69 (0.56)	22.85 (1.03)
0.90	16.89 (0.71)	12.03 (0.48)	11.92 (0.45)	14.75 (0.60)	12.77 (0.50)	23.08 (1.01)
0.95	16.24 (0.67)	10.48 (0.42)	10.35 (0.40)	10.48 (0.42)	10.33 (0.41)	21.38 (0.92)
time	.152 (.005)	33.697 (.734)				.149 (.001)

functions do not cross except potentially near the boundary at two end points 0 and 1 due to the fact that we do not have observations near the boundary in our training sample.

Example 3. We set $p = 5$ and $n = 100$ and consider the linear model

$$Y = X_1 + X_2 + \cdots + X_5 + \epsilon, \tag{10}$$

where $X_j \sim N(0,1)$ for $j = 1, 2, \ldots, 5$ and $\epsilon \sim N(0,1)$ are independent of each other. Conditional quantile functions are to be estimated at $\tau = 0.05, 0.1, \ldots, 0.95$. we generate an independent and identically distributed test set of size $100n$. For each set of quantile estimates $\hat{f}_{\tau_k}(\cdot)$, we evaluate the test error as follows

$$\text{Test Error}(\hat{f}_{\tau_k}) = \sum_{i=1}^{100n} \rho_\tau(\tilde{y}_i - \hat{f}_{\tau_k}(\tilde{\boldsymbol{x}}_i))/(100n), \tag{11}$$

where $(\tilde{y}_i, \tilde{\boldsymbol{x}}_i)$ denotes a general pair of observations in the test set for $i = 1, 2, \ldots, 100n$.

For model (10), the Bayes prediction error is given by $\text{ER}_{Bayes,\tau} = E_\epsilon \rho_\tau(\epsilon - \Phi^{-1}(\tau))$ for any $0 < \tau < 1$. Here $\Phi(\cdot)$ denotes the cumulative distribution function of the standard normal distribution and $\Phi^{-1}(\cdot)$ denotes its inverse function. We calculate $\text{ER}_{Bayes,\tau}$ by Monte Carlo simulation based on a sample of 10^6 independent and identically distributed ϵ from the standard normal distribution.

In Table 1, we report the average of the difference between the Test Error$(\hat{f}_{\tau_k})$ and the Bayes error ER_{Bayes,τ_k} over 200 repetitions for different τ_k's and different conditional quantile estimation methods. Numbers in parentheses are the corresponding standard errors. Here we enlarge by a factor of 1000 in Table 1. Note that the difference between the Test Error$(\hat{f}_{\tau_k})$ and the Bayes error ER_{Bayes,τ_k} indicates the relative performance each conditional quantile estimate $\hat{f}_{\tau_k}(\cdot)$ does comparing to the best theoretical error. Here we denote the naive individual estimation by M1. Intermediate estimates of our non-crossing estimation are denoted by M2, M3, and M4 for Scheme 1, Scheme 2(U), and Scheme 2(D), respectively, as explained in Section 3. Our final non-crossing estimate is denoted by M5. Column He97 corresponds to the method of He (1997). From Table 1, we can easily observe the improvement of our non-crossing estimates over the naive individual estimation. The last row in Table 1 reports the average CPU times (in seconds, with standard error in parentheses) for each method to estimate conditional quantile functions at all different τ's. Because our estimation scheme consists of several different intermediate steps, we report the computation time altogether. It shows that our estimation scheme costs more time to achieve non-crossing and better estimation accuracy.

To further visualize the improvement, we plot in the left panel of Fig. 4 the average of the difference between the Test Error$(\hat{f}_{\tau_k})$ and the Bayes error ER_{Bayes,τ_k} over 200

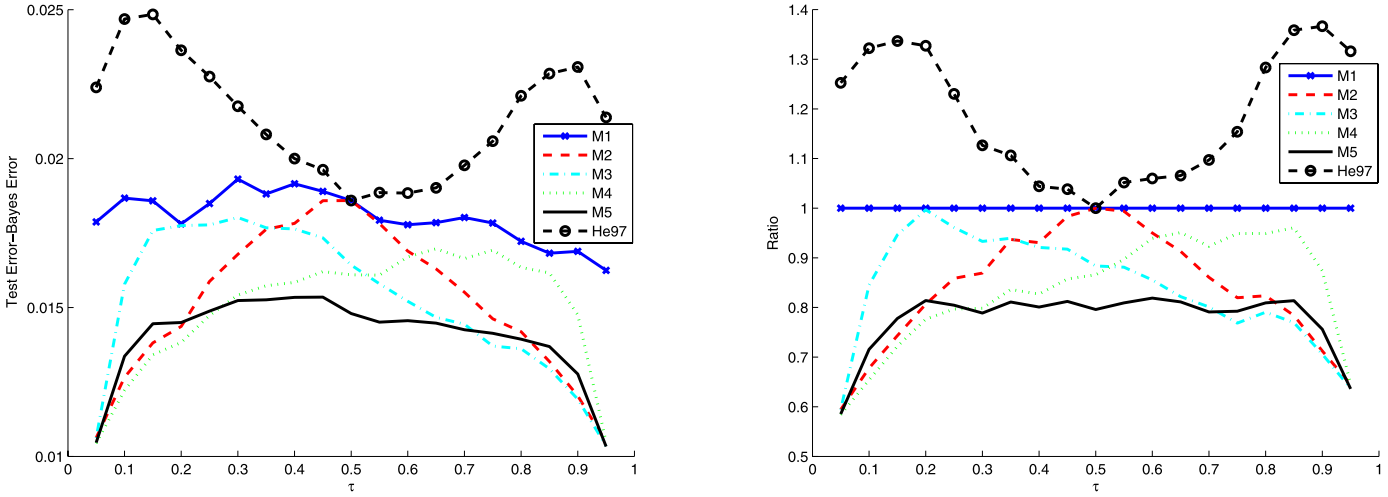

Figure 4. Plots of the average differences and the scaled average differences of the test errors and the Bayes errors over 200 repetitions for different τ's and different methods in Example 3 on the left and right panels respectively.

repetitions for different τ_k's and different conditional quantile estimation methods. Using the naive individual estimation as the baseline, we divide the average difference of each method by the average difference of the naive individual estimation and plot the ratio in the right panel of Fig. 4. It clearly demonstrates that the improvement of our non-crossing estimation because the curve corresponding to our non-crossing estimation falls way below the curve for the naive individual estimation. From Fig. 4, we can see that our non-crossing estimation improves around 20% for most τ's in the middle and even more for other very small or very large τ's. As a remark, we note that in this example, the method of He (1997) appears to be worse than other methods.

Example 4. Different from the previous example, we now consider data from a location scale family with $p = 5$. Each predictor is uniformly distributed, $X_j \sim$ *Uniform*$[0, 1]$ for $j = 1, 2, \ldots, 5$, and independent of each other. The response is generated from the model

$$Y = \sum_{j=1}^{5} X_j + (0.5X_1 + 0.5)\epsilon \tag{12}$$

with independent noise $\epsilon \sim N(0, 1)$. Training samples are of size $n = 100$. An independent and identically distributed test set of size $100n$ is generated to calculate the test error to report performance.

Note that the true conditional quantile function of model (12) is given by $f_\tau(\boldsymbol{x}) = (1 + 0.5\Phi^{-1}(\tau))x_1 + x_2 + x_3 + x_4 + x_5 + 0.5\Phi^{-1}(\tau)$. In this example, we approximate the Bayes error by $\widetilde{\mathrm{ER}}_{Bayes,\tau} = \sum_{i=1}^{100n} \rho_\tau(\tilde{y}_i - f_\tau(\tilde{\boldsymbol{x}}_i))/(100n)$. Results over 200 repetitions are summarized in Table 2 and Fig. 5 in the same format as in Example 3. Consistent improvements are observed for our non-crossing estimation scheme and similar message can be delivered as in the previous example.

Example 5. In this example, we consider a nonlinear model

$$Y = 4\sin(\pi X_1) + 4(X_2 - 0.5)^2 + \epsilon$$

with $X_1 \sim$ *Uniform*$[0, 1]$, $X_2 \sim$ *Uniform*$[0, 1]$, and $\epsilon \sim N(0, 1)$ being independent of each other. We generate independent identically distributed tuning and testing sets of size n and $10n$, respectively. The tuning set is used to select the tuning parameter and the testing set is used to evaluate performance. The Bayes Error is estimated using the test data as in Example 3. We apply the Gaussian kernel $K(\boldsymbol{x}_1, \boldsymbol{x}_2) = \exp(- \parallel \boldsymbol{x}_1 - \boldsymbol{x}_2 \parallel_2^2 / \sigma^2)$. The effect of the kernel parameter σ^2 is also investigated.

Figure 6 plots the average differences between the test error and the Bayes error over 200 repetitions for different σ^2, where different panels correspond to different τ values. This plot shows that the selection of σ^2 does affect the performance of our method. We minimize the sum of the average test errors at different τ's to select the best σ. The results indicate that $\sigma^2 = 0.4$ works well for these methods in this example. In practice, one can use cross validation or other tuning procedures to select σ^2.

Results with $\sigma^2 = 0.4$ are reported in Fig. 7 and Table 3. The results indicate that our proposed method works better than the individual quantile estimation as well as the simultaneous estimation method by Takeuchi et al. (2006). As a remark, we note that the computing time for our method is longer than that of Takeuchi et al. (2006) for this example. One possible explanation is that the dimension of this problem is relatively low.

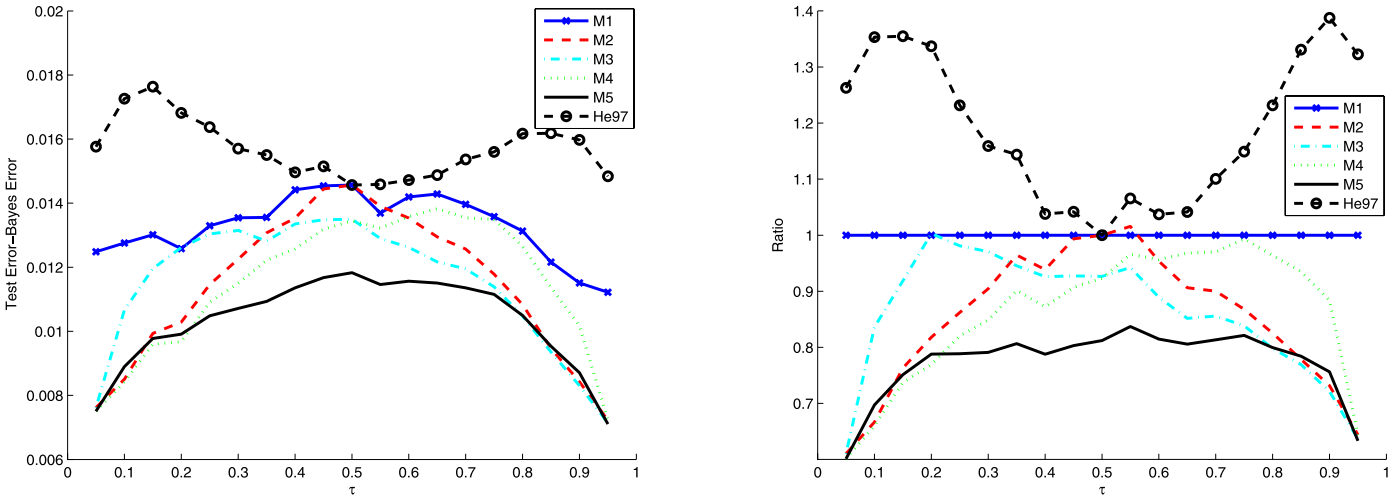

Figure 5. Plots of average differences and the scaled average differences of the test error and the Bayes error over 200 repetitions for different τ's and different methods in Example 4 on the left and right panels respectively.

Table 2. Table of the average of the test error and the Bayes error over 200 repetitions for different τ's and different estimation methods in Example 4. M1 corresponds to the standard quantile regression. M2, M3, and M4 represent our Scheme 1, Scheme 2(U), and Scheme 2(D), respectively. M5 corresponds to our final estimator by averaging. He97 refers to He (1997)'s method. Note that each table entry of error difference is enlarged by a factor of 1000. Numbers in parentheses are the corresponding standard errors

τ	M1	M2	M3	M4	M5	He97
0.05	12.48 (0.56)	7.62 (0.40)	7.62 (0.40)	7.51 (0.39)	7.50 (0.39)	15.76 (0.90)
0.10	12.76 (0.58)	8.51 (0.43)	10.67 (0.52)	8.42 (0.42)	8.89 (0.45)	17.26 (0.98)
0.15	13.02 (0.61)	9.93 (0.49)	11.96 (0.57)	9.59 (0.45)	9.78 (0.47)	17.63 (0.95)
0.20	12.58 (0.62)	10.29 (0.49)	12.62 (0.59)	9.68 (0.44)	9.91 (0.47)	16.82 (0.88)
0.25	13.29 (0.63)	11.46 (0.50)	13.04 (0.60)	10.91 (0.43)	10.48 (0.47)	16.37 (0.84)
0.30	13.55 (0.63)	12.25 (0.52)	13.15 (0.60)	11.50 (0.45)	10.71 (0.45)	15.70 (0.76)
0.35	13.55 (0.56)	13.07 (0.54)	12.81 (0.57)	12.22 (0.47)	10.93 (0.45)	15.50 (0.71)
0.40	14.42 (0.61)	13.53 (0.60)	13.35 (0.58)	12.57 (0.49)	11.36 (0.47)	14.96 (0.69)
0.45	14.54 (0.61)	14.44 (0.61)	13.48 (0.58)	13.19 (0.53)	11.68 (0.47)	15.15 (0.65)
0.50	14.57 (0.60)	14.57 (0.60)	13.50 (0.58)	13.45 (0.54)	11.83 (0.49)	14.57 (0.60)
0.55	13.69 (0.57)	13.90 (0.57)	12.90 (0.56)	13.22 (0.54)	11.46 (0.48)	14.59 (0.59)
0.60	14.19 (0.58)	13.54 (0.56)	12.62 (0.53)	13.58 (0.56)	11.56 (0.47)	14.72 (0.61)
0.65	14.28 (0.60)	12.95 (0.53)	12.17 (0.53)	13.83 (0.56)	11.51 (0.48)	14.88 (0.61)
0.70	13.96 (0.57)	12.57 (0.53)	11.95 (0.53)	13.55 (0.56)	11.36 (0.48)	15.36 (0.62)
0.75	13.58 (0.57)	11.78 (0.51)	11.39 (0.50)	13.49 (0.59)	11.15 (0.47)	15.60 (0.64)
0.80	13.13 (0.57)	10.83 (0.48)	10.48 (0.48)	12.65 (0.54)	10.50 (0.47)	16.17 (0.69)
0.85	12.16 (0.53)	9.47 (0.45)	9.35 (0.45)	11.37 (0.48)	9.54 (0.42)	16.18 (0.71)
0.90	11.51 (0.49)	8.42 (0.38)	8.30 (0.38)	10.17 (0.43)	8.71 (0.39)	15.97 (0.74)
0.95	11.22 (0.47)	7.23 (0.33)	7.14 (0.34)	7.23 (0.33)	7.10 (0.33)	14.84 (0.69)
time	.128 (.003)	28.660 (.739)				.149 (.001)

6. APPLICATION TO THE BASEBALL DATASET

In this section, we apply our stepwise approach to analyze one real data set, the Annual Salary of Baseball Players Data provided by He et al. (1998). This data set is based on $n = 263$ North American major league baseball players for the 1986 season. As in He et al. (1998), we use the number of home runs in the latest year (performance measure) and the number of years played (seniority measure) as predictor variables. The response variable is the annual salary of each player (measured in thousands of dollars). We first standardize each predictor variable to have a mean zero and variance of one. For our stepwise non-crossing non-

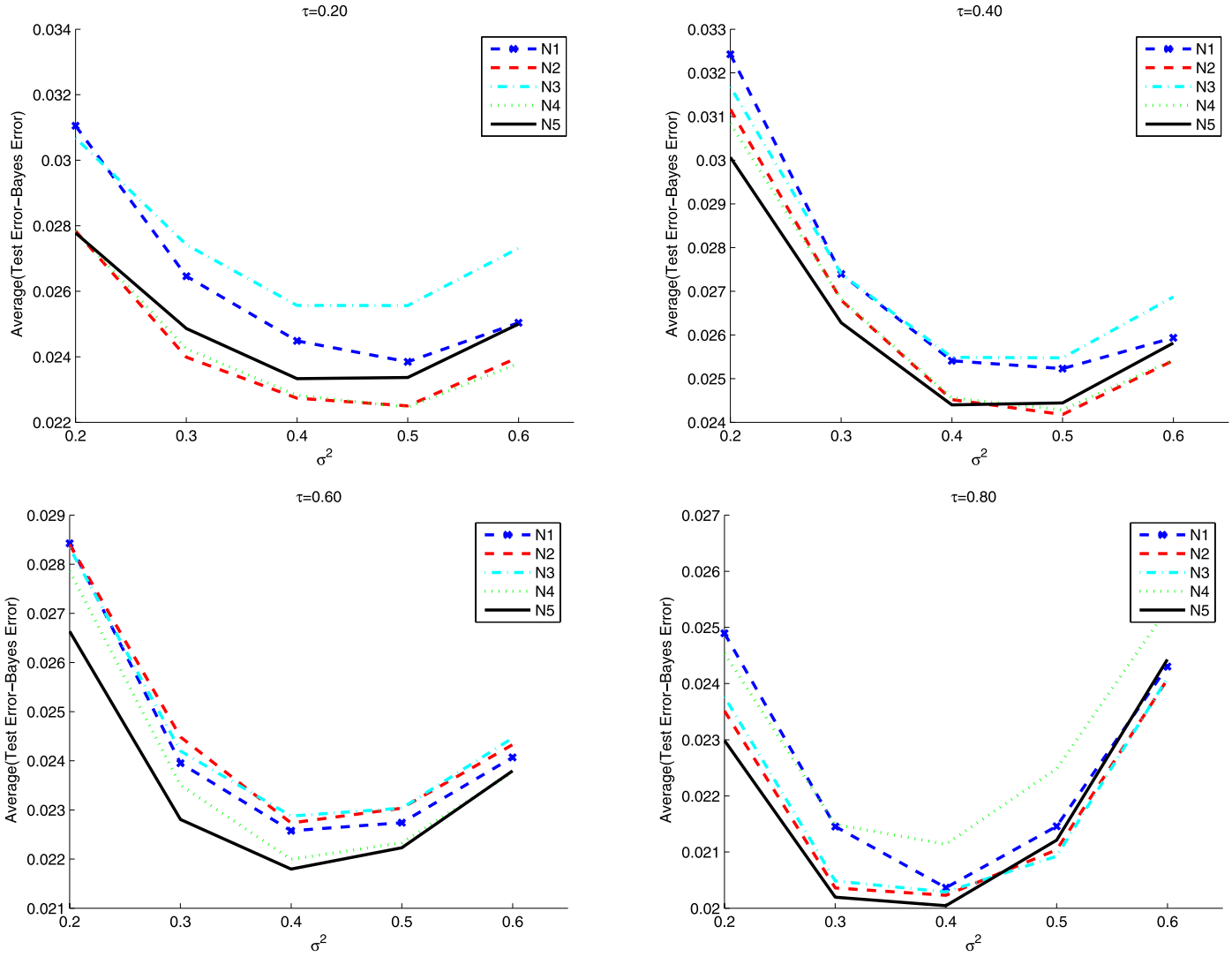

Figure 6. Plots of average differences between the test error and the Bayes error over 200 repetitions for different σ^2 and different methods in Example 5. Different panels correspond to different τ.

parametric kernel regression with restriction, we choose the Gaussian kernel with the data width parameter σ to be the median pairwise Euclidean distance of the new standardized predictor variables. Our limited experience shows that this choice of σ works reasonably well. This choice was also previously used by Brown et al. (2000) and Wu and Liu (2007). To select the regularization parameter λ, we use the 10-fold cross validation in each step of our stepwise non-crossing nonparametric kernel regression with restriction.

We estimate the conditional quantile functions at $\tau = 0.05, 0.1, \ldots, 0.95$. After the estimation is performed, we plot the estimated nonparametric quantile function at $\tau = 0.5$ on the top left panel of Fig. 8. The plot is on the original data scale by applying the inverse linear transformation of the standardization step. To compare with the standard quantile regression, we plot the difference between the new quantile function and the original individually estimated quantile function at $\tau = 0.5$ on the top right panel of Fig. 8. Furthermore, for each k, we plot the difference of the estimated conditional quantile function at τ_{k+1} and τ_k in the original scale. Two examples with $k = 8$ and 11 are displayed on the bottom role of Fig. 8. Note that although the differences are not guaranteed to be nonnegative, the minimal difference is typically positive or close to zero if it is negative.

7. DISCUSSION

In this paper, we consider the estimation problem of multiple non-crossing quantile regression functions. A stepwise procedure is introduced to ensure non-crossing. Our numerical results indicate that our non-crossing method not

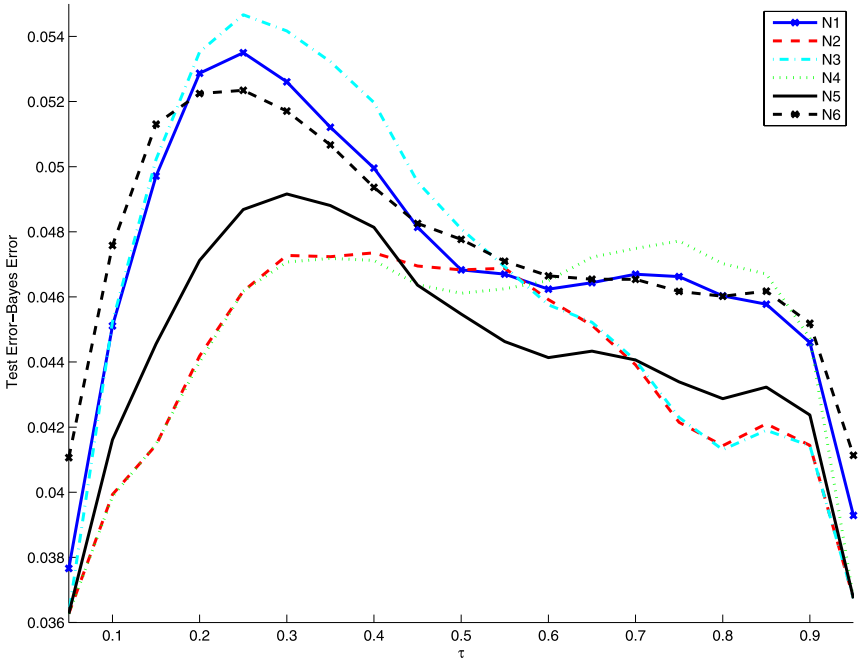

Figure 7. Plot of the average difference of the test error and the Bayes error over 200 repetitions for different τ's and different methods in Example 5. Here N1 denotes the individual estimate. N2 is our first round estimate beginning from zero. N3 and N4 are our second round estimates. N5 is our final estimates by averaging N3 and N4. N6 is the simultaneous estimates by Takeuchi et al. (2006).

Table 3. Table of the average difference of the test error and the Bayes error over 200 repetitions for different τ's and different estimation methods in Example 5. Here N1 denotes the individual estimate. N2 is our first round estimate beginning from zero. N3 and N4 are our second round estimates. N5 is our final estimates by averaging N3 and N4. N6 is the simultaneous estimates proposed by Takeuchi et al. (2006). Note that each table entry of error difference is enlarged by a factor of 1000

τ	N1	N2	N3	N4	N5	N6
0.05	19.23 (0.89)	17.13 (0.79)	17.13 (0.79)	17.14 (0.79)	17.12 (0.79)	18.91(0.87)
0.10	20.21 (0.88)	18.44 (0.83)	20.39 (0.89)	18.45 (0.83)	19.00 (0.84)	19.83(0.88)
0.15	22.49 (1.05)	20.28 (0.95)	23.64 (1.03)	20.42 (0.97)	21.25 (0.98)	21.34(1.05)
0.20	24.49 (1.13)	22.74 (1.04)	25.57 (1.11)	22.82 (1.04)	23.33 (1.05)	23.17(1.12)
0.25	25.49 (1.16)	24.40 (1.07)	26.79 (1.15)	24.40 (1.06)	24.81 (1.08)	24.44(1.11)
0.30	25.88 (1.13)	24.75 (1.05)	26.75 (1.17)	24.72 (1.06)	24.96 (1.09)	25.06(1.09)
0.35	26.31 (1.17)	25.35 (1.11)	26.47 (1.16)	25.30 (1.11)	25.15 (1.12)	25.59(1.14)
0.40	25.41 (1.12)	24.52 (1.08)	25.49 (1.10)	24.56 (1.08)	24.40 (1.08)	25.17(1.09)
0.45	24.56 (1.14)	24.11 (1.08)	24.64 (1.12)	23.89 (1.07)	23.61 (1.07)	24.30(1.12)
0.50	23.37 (1.06)	23.37 (1.06)	23.65 (1.05)	22.69 (1.01)	22.57 (1.01)	23.66(1.05)
0.55	23.10 (0.99)	23.07 (1.00)	23.20 (0.99)	22.24 (0.95)	22.02 (0.95)	22.96(0.98)
0.60	22.58 (1.01)	22.73 (1.01)	22.88 (1.02)	21.99 (0.98)	21.80 (0.98)	22.64(1.03)
0.65	21.92 (1.07)	21.55 (0.99)	21.69 (1.02)	21.21 (1.02)	20.83 (1.00)	21.64(1.04)
0.70	20.96 (1.00)	21.10 (0.96)	21.19 (0.98)	20.46 (0.98)	20.16 (0.95)	20.48(0.98)
0.75	20.65 (1.00)	21.00 (0.96)	20.93 (0.97)	20.73 (1.01)	20.18 (0.95)	20.09(0.99)
0.80	20.37 (1.00)	20.23 (0.96)	20.29 (0.95)	21.14 (1.02)	20.04 (0.95)	19.65(0.97)
0.85	21.21 (1.06)	20.13 (0.99)	20.23 (0.99)	21.91 (1.07)	20.54 (1.01)	20.46(1.03)
0.90	20.34 (0.93)	19.44 (0.88)	19.39 (0.87)	20.88 (0.92)	19.73 (0.88)	20.22(0.91)
0.95	21.95 (0.90)	19.93 (0.78)	19.75 (0.76)	19.93 (0.78)	19.80 (0.76)	21.00(0.78)
time	205.14 (8.14)	630.14 (24.40)				380.76 (12.68)

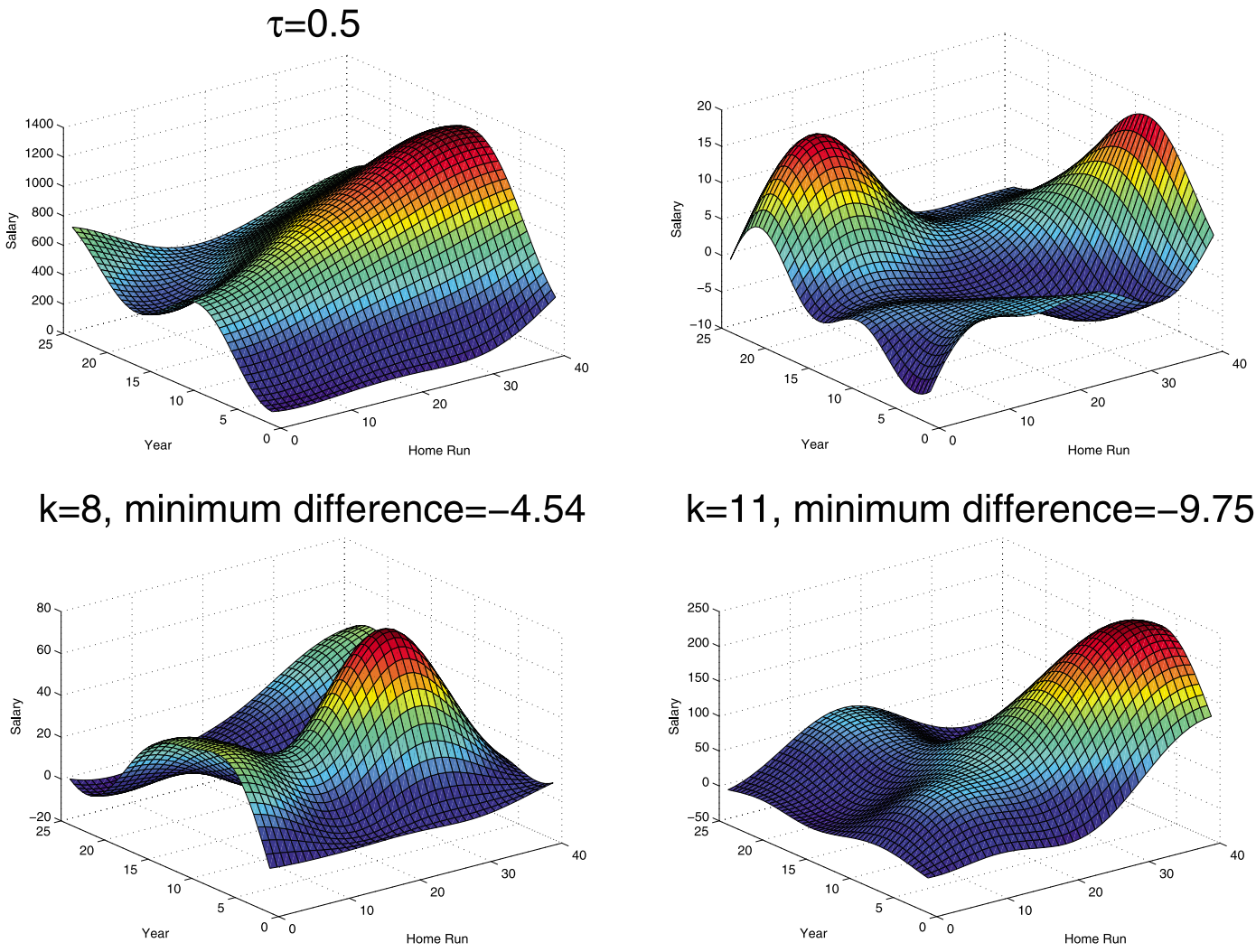

Figure 8. Plots for the Baseball data example. Top-left panel: estimated quantile function with $\tau = 0.5$ using non-crossing constraints; top-right panel: difference function between estimated quantile functions ($\tau = 0.5$) and the estimate without using non-crossing constraints; bottom-left panel: the difference between the quantile functions of τ_9 and τ_8; bottom-right panel: the difference between the quantile functions of τ_{12} and τ_{11}.

only helps to provide more meaningful results, it also improves the estimation accuracy of the resulting regression functions.

As in other regularization problems, the choice of the regularization parameter λ is very important for the performance of quantile regression. It is often for one to select a finite set of representative values for λ and then use a separate validation data set or certain model selection criterion to select a value for λ. In this article, we have used separate validation sets for simulation and cross validation for the real data analysis. As an alternative, one can use certain model selection criterion to choose λ. Two commonly used criteria are the Schwarz information criterion (Schwarz 1978, Koenker et al. 1994) (SIC) and the generalized approximate cross-validation criterion (Yuan 2006) (GACV). These criteria are well studied for unconstrained quantile regression and require further developments for our constrained methods.

Received 31 March 2009

REFERENCES

Brown, M. P. S., Grundy, W. N., Lin, D., Cristianini, N., Sugnet, C. W., Furey, T. S., Ares, M., and Haussler, D. (2000). Knowledge-Based Analysis of Microarray Gene Expression Data by Using Support Vector Machines. *Proceedings of the National Academy of Science* **97** 262–267.

Cole, T. and Green, P. (1992). Smoothing reference centile curves: the LMS method and penalized likelihood. *Statistics in Medicine* **11** 1305–1319.

He, X. (1997). Quantile curves without crossing. *American Statistician* **51**(2) 186–192.

He, X., Ng, P., and Portnoy, S. (1998). Bivariate Quantile Smoothing Splines. *Journal of the Royal Statistical Society, B* **60** 537–550. MR1625950

Heagerty, P. and Pepe, M. (1999). Semiparametric estimation of regression quantiles with application to standardizing weight for height and age in US children. *Journal of the Royal Statistical Society: Series C* **48** 533–551.

Hendricks, W. and Koenker, R. (1992). Hierarchical spline models for conditional quantiles and the demand for electricity. *Journal of the American Statistical Association* **93** 58–68.

Koenker, R. (2005). *Quantile Regression*. Cambridge University Press, New York. MR2268657

Koenker, R. and Bassett, G. (1978). Regression quantiles. *Econometrica* (1) 33–50. MR0474644

Koenker, R. and Geling, R. (2001). Reappraising medfly longevity: a quantile regression survival analysis. *Journal of the American Statistical Association* **96** 458–468. MR1939348

Koenker, R. and Hallock, K. (2001). Quantile regression. *Journal of Economic Perspectives* **15**(4) 143–156.

Koenker, R., Ng, P., and Portnoy, S. (1994). Quantile smoothing splines. *Biometrika* **81**(4) 673–680. MR1326417

Li, Y., Liu, Y., and Zhu, J. (2007). Quantile regression in reproducing kernel Hilbert spaces. *Journal of the American Statistical Association* **102** 255–268. MR2293307

Li, Y. and Zhu, J. (2008). L1-norm quantile regression. *Journal of Computational and Graphical Statistics* **17** 163–185. MR2424800

Neocleousa, T. and Portnoy, S. (2007). On monotonicity of regression quantile functions. *Statistics and Probability Letters* **78**(10) 1226–1229. MR2441467

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics* **6** 461–464. MR0468014

Takeuchi, I., Le, Q. V., Sears, T. D., and Smola, A. J. (2006). Nonparametric Quantile Estimation. *Journal of Machine Learning Research* **7** 1231–1264. MR2274404

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of Royal Statistical Society, B* **58** 267–288. MR1379242

Wang, H., Li, G., and Jiang, G. (2007). Robust regression shrinkage and consistent variable selection through the lad-lasso. *Journal of Business & Economic Statistics* **25** 347–355. MR2380753

Wu, Y. and Liu, Y. (2007). Robust Truncated Hinge Loss Support Vector Machines. *Journal of the American Statistical Association* **102** 974–983. MR2411659

Yang, S. (1999). Censored median regression using weighted empirical survival and hazard functions. *Journal of the American Statistical Association* **94** 137–145. MR1689219

Yuan, M. (2006). GACV for quantile smoothing splines. *Computational Statistics and Data Analysis* **5**(3) 813–829. MR2207010

Yu, K., Lu, Z., and Stander, J. (2003). Quantile regression: applications and current research areas. The Statistician **52** 331–350. MR2011179

Yichao Wu
Department of Statistics
North Carolina State University
Raleigh, NC 27695
USA
E-mail address: wu@stat.ncsu.edu

Yufeng Liu
Department of Statistics and Operations Research
Carolina Center for Genome Sciences
University of North Carolina
Chapel Hill, NC 27599
USA
E-mail address: yfliu@email.unc.edu