Quantile Regression in Reproducing Kernel Hilbert Spaces

Youjuan LI, Yufeng LIU, and Ji ZHU

In this article we consider quantile regression in reproducing kernel Hilbert spaces, which we call kernel quantile regression (KQR). We make three contributions: (1) we propose an efficient algorithm that computes the entire solution path of the KQR, with essentially the same computational cost as fitting one KQR model; (2) we derive a simple formula for the effective dimension of the KQR model, which allows convenient selection of the regularization parameter; and (3) we develop an asymptotic theory for the KQR model.

KEY WORDS: Degrees of freedom; Metric entropy; Model selection; Quadratic programming; Quantile regression; Reproducing kernel Hilbert space.

1. INTRODUCTION

Classical regression methods have focused mainly on estimating conditional mean functions; however, estimation of conditional quantile functions is also often of substantial practical interest. For example, it is well known that the median estimate is more robust to outliers than the traditional mean estimate. In recent years, quantile regression has emerged as a comprehensive approach to the statistical analysis of response models, and it has been widely used in many real applications, including reference charts in medicine (Cole and Green 1992; Heagerty and Pepe 1999), survival analysis (Yang 1999; Koenker and Geling 2001), and economics (Hendricks and Koenker 1992; Koenker and Hallock 2001). For comprehensive reviews of quantile regression, see the articles by Koenker and Hallock (2001) and Yu, Lu, and Stander (2003), as well as the exceptionally well-written book by Koenker (2005).

Suppose that we have a set of training data, $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)$, with input $\mathbf{x}_i \in \mathbb{R}^p$ and output $y_i \in \mathbb{R}$, and we would like to recover the $100\tau\%$ quantile of the conditional distribution of y given $\mathbf{x}$. In the case where $p = 1$, Koenker, Ng, and Portnoy (1994) suggested

$$\min_{f\in\mathcal{F}} \sum_{i=1}^{n} \rho_\tau(y_i - f(x_i)) + \frac{\lambda}{2}\left(\int_0^1 (f''(x))^q\right)^{1/q}, \tag{1}$$

where q is a positive integer and $\rho_\tau(r)$ is the so-called "check function" of Koenker and Bassett (1978) (Fig. 1),

$$\rho_\tau(r) = \begin{cases} \tau r & \text{if } r > 0 \\ -(1-\tau)r & \text{otherwise.} \end{cases} \tag{2}$$

Here $\tau \in (0,1)$ indicates the quantile of interest, and $\lambda > 0$ controls the balance between the smoothness of the fit and its fidelity to the data. For $q = 1$, with an appropriately chosen model space, Koenker et al. (1994) showed that the solution is a linear spline with knots at the data points, which leads essentially to an L_1 loss $+$ L_1 penalty problem.

For $q = 2$ (Bloomfield and Steiger 1983; Nychka, Gray, Haaland, Martin, and O'Connell 1995), (1) reduces to

$$\min_{f\in\mathcal{F}} \sum_{i=1}^{n} \rho_\tau(y_i - f(x_i)) + \frac{\lambda}{2}\int_0^1 f''(x)^2\,dx. \tag{3}$$

We can view (3) as an analogy to the more extensively studied classical least squares smoothing spline model pioneered by Wahba (1990) and her collaborators (Gu 2002). The solution to (3) over the second-order Sobolev space is a natural cubic spline with knots at the data points.

In this article we consider the more general setup of (3),

$$\min_{f\in\mathcal{H}_K} \sum_{i=1}^{n} \rho_\tau(y_i - f(\mathbf{x}_i)) + \frac{\lambda}{2}\|f\|^2_{\mathcal{H}_K}, \tag{4}$$

where $\mathbf{x}_i \in \mathbb{R}^p$ and $\mathcal{H}_K$ is a structured reproducing kernel Hilbert space (RKHS) generated by a positive definite kernel $K(\mathbf{x}, \mathbf{x}')$. This includes the entire family of smoothing splines and additive and interaction spline models (Wahba 1990). Some other popular choices of $K(\cdot,\cdot)$ in practice are

$$d\text{th-degree polynomial:} \quad K(\mathbf{x},\mathbf{x}') = (1 + \langle \mathbf{x},\mathbf{x}'\rangle)^d$$

and

$$\text{radial basis:} \quad K(\mathbf{x},\mathbf{x}') = \exp(-\|\mathbf{x}-\mathbf{x}'\|^2/2\sigma^2),$$

where d and σ are prespecified parameters.

Using the representer theorem (Kimeldorf and Wahba 1971), the solution to (4) has a finite form,

$$\hat{f}(\mathbf{x}) = \beta_0 + \frac{1}{\lambda}\sum_{i=1}^{n} \theta_i K(\mathbf{x},\mathbf{x}_i). \tag{5}$$

Note that we write $\hat{f}(\mathbf{x})$ in a way that involves λ explicitly; later, we show that $\theta_i \in [-(1-\tau), \tau]$. Given the format of the solution (5), we can in turn rewrite (4) in finite form as

$$\min_{\beta_0,\boldsymbol{\theta}} \sum_{i=1}^{n} \rho_\tau\left(y_i - \beta_0 - \frac{1}{\lambda}\sum_{i'=1}^{n}\theta_{i'}K(\mathbf{x}_i,\mathbf{x}_{i'})\right) + \frac{1}{2\lambda}\sum_{i=1}^{n}\sum_{i'=1}^{n}\theta_i\theta_{i'}K(\mathbf{x}_i,\mathbf{x}_{i'}), \tag{6}$$

which we call *kernel quantile regression* (KQR).

Youjuan Li is PhD Student, Department of Statistics, University of Michigan, Ann Arbor, MI 48109. Yufeng Liu is Assistant Professor, Department of Statistics and Operations Research, Carolina Center for Genome Sciences, University of North Carolina, Chapel Hill, NC 27599. Ji Zhu is Assistant Professor, Department of Statistics, University of Michigan, Ann Arbor, MI 48109 (E-mail: *jizhu@umich.edu*). The authors thank the joint editor, the associate editor, and three referees for their thoughtful and useful comments. They also thank Trevor Hastie, Saharon Rosset, Rob Tibshirani, and Hui Zou for helpful discussions. Partial support has been provided by the National Science Foundation (grants DMS-06-06577 and DMS-05-05432).

Journal of the American Statistical Association
March 2007, Vol. 102, No. 477, Theory and Methods
DOI 10.1198/016214506000000979

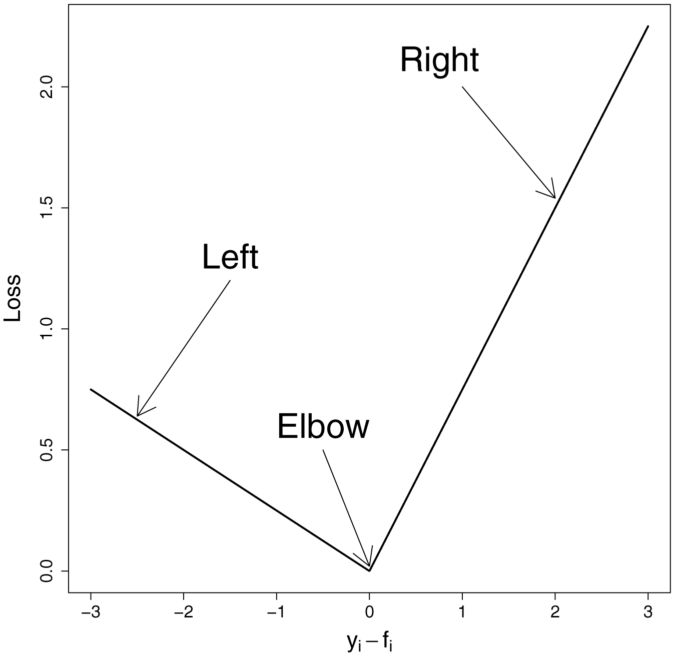

Figure 1. The Check Function, With $\tau = .75$.

The KQR model (6) can be transformed into a quadratic programming problem; thus most commercially available packages can be used to solve the KQR. Many specific algorithms for the KQR have been developed, including the interior point algorithm (Bosch, Ye, and Woodworth 1995) and the pseudodata algorithm (Nychka et al. 1995). All of these algorithms solve the KQR for a prefixed regularization parameter λ. As in any smoothing problem, then choice of regularization parameter λ is critical. In practice, people usually prespecify a finite set of values for λ that covers a wide range, then either use a separate validation dataset or certain model selection criterion to choose a value for λ that gives the best performance in the prespecified set. Two commonly used criteria for KQR are the Schwarz information criterion (SIC) (Schwarz 1978; Koenker et al. 1994) and the generalized approximate cross-validation criterion (GACV) (Yuan 2006),

$$\mathrm{SIC}(\lambda) = \ln\left(\frac{1}{n}\sum_{i=1}^{n}\rho_\tau(y_i - \hat{f}(\mathbf{x}_i))\right) + \frac{\ln n}{2n} df \tag{7}$$

and

$$\mathrm{GACV}(\lambda) = \frac{\sum_{i=1}^{n}\rho_\tau(y_i - \hat{f}(\mathbf{x}_i))}{n - df}, \tag{8}$$

where *df* is a measure of the effective dimensionality of the fitted model. Koenker et al. (1994) heuristically argued that in the case of one-dimensional quantile smoothing spline, the number of interpolated y_i's is a plausible measure for the effective dimension of the fitted model. In the case of GACV and its earlier cousin ACV, Yuan (2006) and Nychka et al. (1995) argued that the divergence,

$$\mathrm{div}(\hat{f}) = \sum_{i=1}^{n}\frac{\partial \hat{f}(\mathbf{x}_i)}{\partial y_i}, \tag{9}$$

can be used for *df*. They used a smooth approximation of the check function to compute $\mathrm{div}(\hat{f})$.

This article makes three main contributions:

- We show that the solution $\boldsymbol{\theta}(\lambda)$ is *piecewise linear* as a function of λ and derive an efficient algorithm that computes the *exact entire solution path*, $\{\boldsymbol{\theta}(\lambda), 0 \le \lambda \le \infty\}$, ranging from the least regularized model to the most regularized model.
- We prove that in the case of KQR, the divergence (9) is exactly equal to the number of interpolated y_i's, which justifies its use in selecting the regularization parameter λ.
- We develop an asymptotic theory for the KQR. In particular, using metric entropy and large deviation theories, we obtain the convergence rate of the difference between the KQR solution and the true quantile function in terms of their mean check deviations.

We acknowledge that the first result was inspired by one of the authors' earlier work in the support vector machine setting (Hastie, Rosset, Tibshirani, and Zhu 2004).

Before delving into the technical details, we illustrate the concept of piecewise linearity of the solution path with a simple example. We generate six training observations using the famous *sinc*$(\cdot)$ function,

$$y = \frac{\sin(\pi x)}{\pi x} + \epsilon,$$

where x is distributed as uniform$(-2, 2)$ and ϵ is distributed as normal$(0, .2^2)$. We use the KQR with a one-dimensional spline kernel (Wahba 1990),

$$K(x, x') = 1 + k_1(x)k_1(x') + k_2(x)k_2(x') - k_4(|x - x'|),$$

where $k_1(\cdot) = \cdot - 1/2$, $k_2 = (k_1^2 - 1/12)/2$, $k_4 = (k_1^4 - k_1^2/2 + 7/240)/24$. Figure 2 shows a subset of the piecewise linear solution path $\boldsymbol{\theta}(\lambda)$ as a function of λ, and also how the number of interpolated y_i's changes with λ.

The rest of the article is organized as follows. In Section 2 we derive the algorithm that computes the entire solution path of the KQR. In Section 3 we prove that the divergence (9) is equal to the number of interpolated y_i's for the KQR, and in Section 4 we develop an asymptotic theory for the KQR. In Section 5 we present numerical results on both simulation and real-world data. We end with some conclusions in Section 6.

2. ALGORITHM

2.1 Problem Setup

Criterion (6) can be rewritten in an equivalent way as

$$\min_{\beta_0,\boldsymbol{\theta}} \tau\sum_{i=1}^{n}\xi_i + (1-\tau)\sum_{i=1}^{n}\zeta_i + \frac{1}{2\lambda}\boldsymbol{\theta}^\top\mathbf{K}\boldsymbol{\theta}, \tag{10}$$

subject to

$$-\zeta_i \le y_i - f(\mathbf{x}_i) \le \xi_i \tag{11}$$

and

$$\zeta_i, \xi_i \ge 0, \qquad i = 1, \dots, n, \tag{12}$$

where

$$f(\mathbf{x}_i) = \beta_0 + \frac{1}{\lambda}\sum_{i'=1}^{n}\theta_{i'}K(\mathbf{x}_i, \mathbf{x}_{i'}), \qquad i = 1, \dots, n,$$

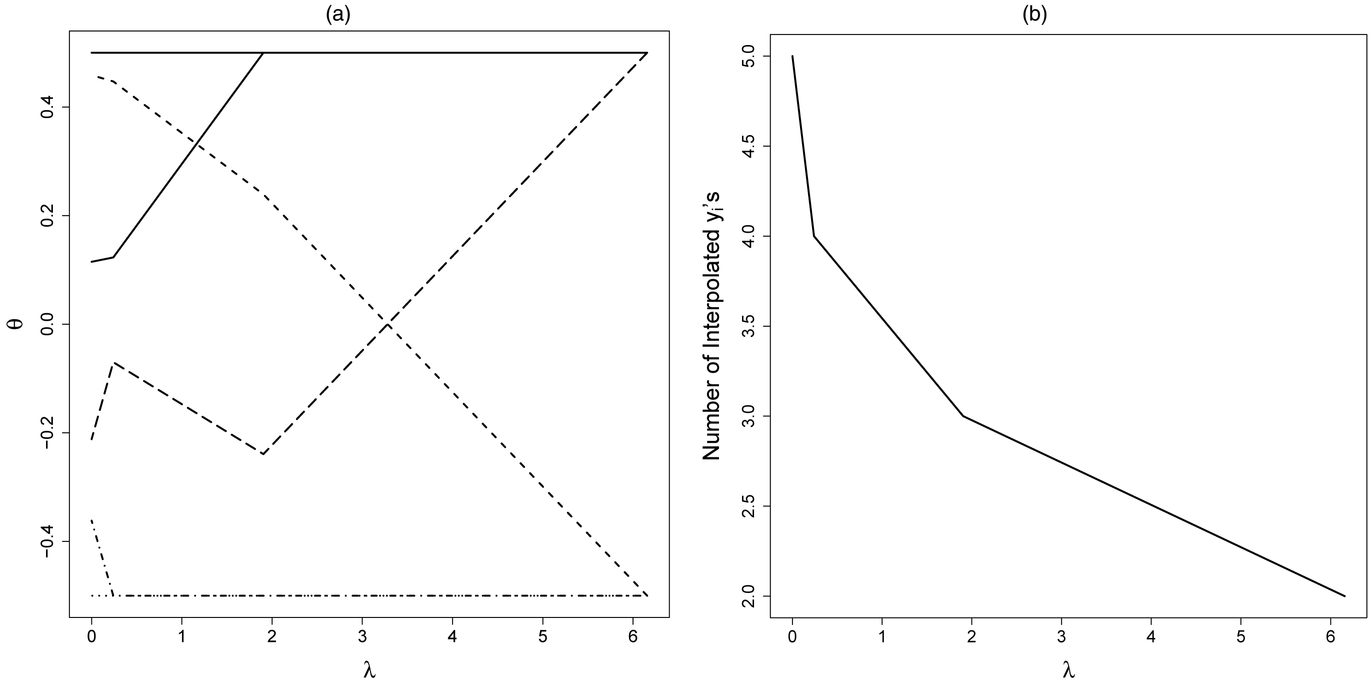

Figure 2. Illustrative Example. (a) A subset of the solution path $\boldsymbol{\theta}(\lambda)$ as a function of λ. (b) How the number of interpolated y_i's changes with λ.

and

$$\mathbf{K} = \begin{pmatrix} K(\mathbf{x}_1,\mathbf{x}_1) & \cdots & K(\mathbf{x}_1,\mathbf{x}_n) \\ \vdots & \ddots & \vdots \\ K(\mathbf{x}_n,\mathbf{x}_1) & \cdots & K(\mathbf{x}_n,\mathbf{x}_n) \end{pmatrix}_{n\times n}.$$

For the rest of the article, we assume that the data points $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)$ are in general positions and that the kernel matrix $\mathbf{K}$ is positive definite. Then the foregoing setting gives the Lagrangian primal function,

$$\begin{aligned} L_p\colon\quad & \tau\sum_{i=1}^{n}\xi_i + (1-\tau)\sum_{i=1}^{n}\zeta_i + \frac{1}{2\lambda}\boldsymbol{\theta}^\top\mathbf{K}\boldsymbol{\theta} \\ & + \sum_{i=1}^{n}\alpha_i(y_i - f(\mathbf{x}_i) - \xi_i) - \sum_{i=1}^{n}\gamma_i(y_i - f(\mathbf{x}_i) + \zeta_i) \\ & - \sum_{i=1}^{n}\kappa_i\xi_i - \sum_{i=1}^{n}\rho_i\zeta_i, \end{aligned} \tag{13}$$

where $\alpha_i, \gamma_i, \kappa_i$, and ρ_i are nonnegative Lagrange multipliers. Setting the derivatives of L_p to 0, we arrive at

$$\frac{\partial}{\partial\boldsymbol{\theta}}\colon\quad \theta_i = \alpha_i - \gamma_i, \tag{14}$$

$$\frac{\partial}{\partial\beta_0}\colon\quad \sum_{i=1}^{n}\alpha_i = \sum_{i=1}^{n}\gamma_i, \tag{15}$$

$$\frac{\partial}{\partial\xi_i}\colon\quad \alpha_i = \tau - \kappa_i, \tag{16}$$

and

$$\frac{\partial}{\partial\zeta_i}\colon\quad \gamma_i = 1 - \tau - \rho_i, \tag{17}$$

and the Karush–Kuhn–Tucker conditions are

$$\alpha_i(y_i - f(\mathbf{x}_i) - \xi_i) = 0, \tag{18}$$

$$\gamma_i(y_i - f(\mathbf{x}_i) + \zeta_i) = 0, \tag{19}$$

$$\kappa_i\xi_i = 0, \tag{20}$$

and

$$\rho_i\zeta_i = 0. \tag{21}$$

Because the Lagrange multipliers must be nonnegative, we can conclude from (16) and (17) that both $0 \le \alpha_i \le \tau$ and $0 \le \gamma_i \le 1-\tau$. We also see from (18) and (19) that if α_i is positive, then γ_i must be 0, and vice versa. These lead to the following relationships:

$$y_i - f(\mathbf{x}_i) > 0 \quad\Rightarrow\quad \alpha_i = \tau, \xi_i > 0, \gamma_i = 0, \zeta_i = 0;$$

$$y_i - f(\mathbf{x}_i) < 0 \quad\Rightarrow\quad \alpha_i = 0, \xi_i = 0, \gamma_i = 1-\tau, \zeta_i > 0;$$

and

$$y_i - f(\mathbf{x}_i) = 0 \quad\Rightarrow\quad \alpha_i \in [0,\tau], \xi_i = 0, \gamma_i \in [0, 1-\tau], \zeta_i = 0.$$

Note from (14) that for every λ, θ_i is equal to $(\alpha_i - \gamma_i)$. Hence, using these relationships, we can define the following three sets, which we use later when calculating the regularization path of the KQR:

- $\mathcal{E} = \{i : y_i - f(\mathbf{x}_i) = 0, -(1-\tau) \le \theta_i \le \tau\}$ (elbow)
- $\mathcal{L} = \{i : y_i - f(\mathbf{x}_i) < 0, \theta_i = -(1-\tau)\}$ (left of the elbow)
- $\mathcal{R} = \{i : y_i - f(\mathbf{x}_i) > 0, \theta_i = \tau\}$ (right of the elbow).

For points in $\mathcal{L}$ and $\mathcal{R}$, the values of θ_i are known; therefore, the algorithm focuses on points resting at the elbow $\mathcal{E}$.

The basic idea of our algorithm is as follows. We start with $\lambda = \infty$ and decrease it toward 0, keeping track of the locations of all data points relative to the elbow along the way. As λ decreases, points move from the left of the elbow to the right of the elbow (or vice versa). Their corresponding θ_i's change from $-(1-\tau)$ when they are on the left of the elbow to τ when they are on the right of the elbow. By continuity, points must linger on the elbow while their θ_i's change from $-(1-\tau)$ to τ. Because all points at the elbow have $y_i - f(\mathbf{x}_i) = 0$, we can establish a path for their θ_i's, and this set will remain stable until either some other point comes to the elbow or one point at the elbow has departed from the elbow.

2.2 Initialization

Initially, when $\lambda = \infty$, we can see from (5) that $\hat{f}(\mathbf{x}) = \beta_0$. We can determine the value of β_0 through a simple one-dimensional optimization. For simplicity of exposition, we focus on the case where all the values of y_i are distinct and ordered as $y_1 < y_2 < \cdots < y_n$. We distinguish between two cases: the initial β_0 is unique, and the initial β_0 is nonunique.

Case 1: The Initial β_0 Is Unique. This occurs when $n\tau$ is a noninteger, for example, when $\tau = .5$ and the number of data points n is odd. In this case, it is easy to show that β_0 must be equal to one of the observed y_i's and $\beta_0 = y_{\lfloor n\tau \rfloor + 1}$, say y_{i^*}. Therefore, all data points are initially divided into three sets:

- $\mathcal{E} = \{i^* : \text{point } (\mathbf{x}_{i^*}, y_{i^*})\}$
- $\mathcal{L} = \{i : y_i < y_{i^*}\}$
- $\mathcal{R} = \{i : y_i > y_{i^*}\}$.

From (15), we have that

$$\theta_{i^*} = n_{\mathcal{L}}(1-\tau) - n_{\mathcal{R}}\tau,$$

where $n_{\mathcal{L}} = |\mathcal{L}|$ and $n_{\mathcal{R}} = |\mathcal{R}|$. When λ decreases, due to the constraint (15), $(\mathbf{x}_{i^*}, y_{i^*})$ will stay at the elbow before another data point enters the elbow. Therefore, for sufficiently large values of λ, we have

$$\begin{aligned}\hat{f}(\mathbf{x}) &= \beta_0 + \frac{1}{\lambda}\Bigg[-(1-\tau)\sum_{i\in\mathcal{L}} K(\mathbf{x},\mathbf{x}_i) \\ &\qquad + \tau\sum_{i\in\mathcal{R}} K(\mathbf{x},\mathbf{x}_i) + \theta_{i^*}K(\mathbf{x},\mathbf{x}_{i^*})\Bigg] \\ &= \beta_0 + \frac{1}{\lambda}g(\mathbf{x}),\end{aligned}$$

where $g(\mathbf{x}) = -(1-\tau)\sum_{i\in\mathcal{L}} K(\mathbf{x},\mathbf{x}_i) + \tau\sum_{i\in\mathcal{R}} K(\mathbf{x},\mathbf{x}_i) + \theta_{i^*}K(\mathbf{x},\mathbf{x}_{i^*})$. Again, because $(\mathbf{x}_{i^*}, y_{i^*})$ stays at the elbow, the intercept β_0 is determined through $\beta_0 = y_{i^*} - \frac{1}{\lambda}g(\mathbf{x}_{i^*})$.

When a data point enters the elbow, it satisfies

$$y_i = \beta_0 + \frac{1}{\lambda}g(\mathbf{x}_i).$$

Hence the entry value of λ (i.e., the largest value of $\lambda < \infty$ that starts to change $\boldsymbol{\theta}$) is given by

$$\lambda_1 = \max_{i\neq i^*} \frac{g(\mathbf{x}_i) - g(\mathbf{x}_{i^*})}{y_i - y_{i^*}}.$$

Case 2: The Initial β_0 Is Nonunique. This occurs when $n\tau$ is an integer, for example, when $\tau = .5$ and the number of data points n is even. In this case it is easy to show that β_0 can take any value between two adjacent y_i's and $\beta_0 \in [y_{n\tau}, y_{n\tau+1}]$, say $[y_{i^*}, y_{j^*}]$.

Although β_0 is not unique, all of the θ_i's are fully determined, that is:

- $\theta_i = -(1-\tau), y_i \le y_{i^*}$
- $\theta_i = \tau, y_i \ge y_{j^*}$.

Hence again, we can divide all data points into three sets:

- $\mathcal{E} = \emptyset$
- $\mathcal{L} = \{i : y_i \le y_{i^*}\}$
- $\mathcal{R} = \{i : y_i \ge y_{j^*}\}$.

For sufficiently large values of λ, we have

$$\begin{aligned}\hat{f}(\mathbf{x}) &= \beta_0 + \frac{1}{\lambda}\Bigg[-(1-\tau)\sum_{i\in\mathcal{L}} K(\mathbf{x},\mathbf{x}_i) + \tau\sum_{i\in\mathcal{R}} K(\mathbf{x},\mathbf{x}_i)\Bigg] \\ &= \beta_0 + \frac{1}{\lambda}g(\mathbf{x}),\end{aligned}$$

where $g(\mathbf{x}) = -(1-\tau)\sum_{i\in\mathcal{L}} K(\mathbf{x},\mathbf{x}_i) + \tau\sum_{i\in\mathcal{R}} K(\mathbf{x},\mathbf{x}_i)$.

When λ decreases, by continuity and the balance between the θ_i's, $\mathcal{L}$ and $\mathcal{R}$ will stay the same; therefore,

$$y_i - \beta_0 - \frac{1}{\lambda}g(\mathbf{x}_i) \le 0, \qquad i \in \mathcal{L},$$

and

$$y_i - \beta_0 - \frac{1}{\lambda}g(\mathbf{x}_i) \ge 0, \qquad i \in \mathcal{R}.$$

These inequalities imply that the solution for β_0 is not unique and that β_0 can be any value in the interval

$$\left[\max_{i\in\mathcal{L}}\left(y_i - \frac{1}{\lambda}g(\mathbf{x}_i)\right),\ \min_{i\in\mathcal{R}}\left(y_i - \frac{1}{\lambda}g(\mathbf{x}_i)\right)\right].$$

When λ decreases, the length of this interval changes, and when two data points (from different sets) hit the elbow simultaneously, the length of the interval shrinks to 0.

2.3 The Regularization Path

The algorithm focuses on the set of points $\mathcal{E}$. These points have $\hat{f}(\mathbf{x}_i) = y_i$ with $\theta_i \in [-(1-\tau), \tau]$. As we follow the path, we examine this set until it changes, at which point we say that an *event* has occurred. Thus events can be categorized as follows:

1. A point from $\mathcal{L}$ has just entered $\mathcal{E}$, with θ_i initially $-(1-\tau)$.
2. A point from $\mathcal{R}$ has just entered $\mathcal{E}$, with θ_i initially τ.
3. Point(s) from $\mathcal{E}$ has just left the elbow to join either $\mathcal{L}$ or $\mathcal{R}$.

Until another event occurs, all sets will remain the same. As a point passes through $\mathcal{E}$, its respective θ_i must change from $-(1-\tau) \to \tau$ or $\tau \to -(1-\tau)$. Relying on the fact that $\hat{f}(\mathbf{x}_i) = y_i$ for all points in $\mathcal{E}$, we can calculate θ_i for these points.

From event 3, we may reach the situation where $\mathcal{E}$ becomes empty. When this occurs, as with initialization, the solution for

β_0 is not unique. However, we can again resort to case 2 of initialization until the length of the interval for β_0 reaches 0.

We use the subscript ℓ to index the foregoing sets immediately after the ℓth event has occurred, and let θ_i^ℓ, β_0^ℓ, and λ^ℓ be the parameter values immediately after the ℓth event; we also let f^ℓ be the function at this point. For convenience, we define $\beta_{0,\lambda} = \lambda \cdot \beta_0$ and hence $\beta_{0,\lambda}^\ell = \lambda^\ell \cdot \beta_0^\ell$. Then, because

$$\hat{f}(\mathbf{x}) = \frac{1}{\lambda}\left(\beta_{0,\lambda} + \sum_{i=1}^{n} \theta_i K(\mathbf{x}, \mathbf{x}_i)\right)$$

for $\lambda^{\ell+1} < \lambda < \lambda^\ell$, we can write

$$\begin{aligned}\hat{f}(\mathbf{x}) &= \left[f(\mathbf{x}) - \frac{\lambda^\ell}{\lambda} f^\ell(\mathbf{x})\right] + \frac{\lambda^\ell}{\lambda} f^\ell(\mathbf{x}) \\ &= \frac{1}{\lambda}\left[(\beta_{0,\lambda} - \beta_{0,\lambda}^\ell) + \sum_{i\in\mathcal{E}^\ell} (\theta_i - \theta_i^\ell) K(\mathbf{x}, \mathbf{x}_i) + \lambda^\ell f^\ell(\mathbf{x})\right],\end{aligned}$$

where the reduction occurs in the second line because the θ_i's are fixed for all points in $\mathcal{R}^\ell$ and $\mathcal{L}^\ell$ and all points remain in their respective sets. Suppose that $|\mathcal{E}^\ell| = n_{\mathcal{E}}^\ell$; then, for the $n_{\mathcal{E}}^\ell$ points staying at the elbow, we have that

$$y_k = \frac{1}{\lambda}\left[(\beta_{0,\lambda} - \beta_{0,\lambda}^\ell) + \sum_{i\in\mathcal{E}^\ell} (\theta_i - \theta_i^\ell) K(\mathbf{x}_k, \mathbf{x}_i) + \lambda^\ell f^\ell(\mathbf{x})\right], \qquad \forall k \in \mathcal{E}^\ell.$$

To simplify, let $\nu_i = \theta_i - \theta_i^\ell$ and $\nu_0 = \beta_{0,\lambda} - \beta_{0,\lambda}^\ell$. Then

$$\nu_0 + \sum_{i\in\mathcal{E}^\ell} \nu_i K(\mathbf{x}_k, \mathbf{x}_i) = (\lambda - \lambda^\ell) y_k, \qquad \forall k \in \mathcal{E}^\ell.$$

In addition, by condition (15), we have that

$$\sum_{i\in\mathcal{E}^\ell} \nu_i = 0.$$

This gives us $n_{\mathcal{E}}^\ell + 1$ linear equations that we can use to solve for each of the $n_{\mathcal{E}}^\ell + 1$ unknown variables ν_i and ν_0.

Now define $\mathbf{K}^\ell$ to be a $n_{\mathcal{E}}^\ell \times n_{\mathcal{E}}^\ell$ matrix with the entries equal to $K(\mathbf{x}_i, \mathbf{x}_k)$, where $i, k \in \mathcal{E}^\ell$, and let ν denote the vector with the components equal to ν_i, $i \in \mathcal{E}^\ell$. Finally, let $\mathbf{y}_{\mathcal{E}}^\ell$ be a vector with the components equal to y_k, $k \in \mathcal{E}^\ell$. Using these notations, we have the following two equations:

$$\nu_0 \mathbf{1} + \mathbf{K}^\ell \nu = (\lambda - \lambda^\ell)\mathbf{y}_{\mathcal{E}}^\ell \tag{22}$$

and

$$\nu^\top \mathbf{1} = 0. \tag{23}$$

Simplifying further, if we let

$$\mathbf{A}^\ell = \begin{pmatrix} 0 & \mathbf{1}^\top \\ \mathbf{1} & \mathbf{K}^\ell \end{pmatrix}, \qquad \nu_0 = \begin{pmatrix} \nu_0 \\ \nu \end{pmatrix},$$

and

$$\mathbf{y}_0 = \begin{pmatrix} 0 \\ \mathbf{y}_{\mathcal{E}}^\ell \end{pmatrix},$$

then (22) and (23) can be combined to give

$$\mathbf{A}^\ell \nu_0 = (\lambda - \lambda^\ell)\mathbf{y}_0.$$

Then if $\mathbf{A}^\ell$ has full rank, we can define

$$\mathbf{b}_0 = (\mathbf{A}^\ell)^{-1}\mathbf{y}_0$$

to give

$$\theta_i = \theta_i^\ell + (\lambda - \lambda^\ell) b_i, \qquad \forall i \in \mathcal{E}^\ell, \tag{24}$$

and

$$\beta_{0,\lambda} = \beta_{0,\lambda}^\ell + (\lambda - \lambda^\ell) b_0. \tag{25}$$

Thus for $\lambda^{\ell+1} < \lambda < \lambda^\ell$, the θ_i and $\beta_{0,\lambda}$ proceed linearly in λ. In addition,

$$\hat{f}(\mathbf{x}) = \frac{\lambda^\ell}{\lambda}[f^\ell(\mathbf{x}) - h^\ell(\mathbf{x})] + h^\ell(\mathbf{x}), \tag{26}$$

where

$$h^\ell(\mathbf{x}) = b_0 + \sum_{i\in\mathcal{E}^\ell} b_i K(\mathbf{x}, \mathbf{x}_i).$$

Given λ_ℓ, (24) and (26) allow us to compute $\lambda_{\ell+1}$, the λ at which the next event will occur. This will be the largest λ less than λ_ℓ, such that either θ_i for $i \in \mathcal{E}^\ell$ reaches τ or $-(1-\tau)$, or one of the points in $\mathcal{R}$ or $\mathcal{L}$ reaches the elbow. The latter event will occur for a point k when

$$\lambda = \lambda_\ell \left(\frac{f^\ell(\mathbf{x}_k) - h^\ell(\mathbf{x}_k)}{y_k - h^\ell(\mathbf{x}_k)}\right), \qquad \forall k \in \mathcal{R}^\ell \cup \mathcal{L}^\ell.$$

We terminate the algorithm either when the sets $\mathcal{R}$ and $\mathcal{L}$ become empty or when λ becomes sufficiently close to 0. (We set the threshold to 10^{-8} in our implementation.)

2.4 Computational Cost

The major computational cost of updating the solutions at any event ℓ involves two aspects: solving the system of $n_{\mathcal{E}}^\ell$ linear equations and computing $h^\ell(\mathbf{x})$. The former takes $O(n_{\mathcal{E}}^{\ell 2})$ calculations by using inverse updating and downdating, because the elbow set usually differs by only one point between consecutive events, and the latter requires $O(nn_{\mathcal{E}}^\ell)$ computations.

In to our experience, the total number of steps taken by the algorithm is on average some small multiple of n. Letting m be the average size of $\mathcal{E}^\ell$, the approximate computational cost of the algorithm is $O(cn^2m + nm^2)$.

3. THE EFFECTIVE DIMENSION OF KERNEL QUANTILE REGRESSION

It is well known that an appropriate value of λ is crucial to the performance of the fitted model in any smoothing problems. One advantage of computing the entire solution path is that this facilitates selection of the regularization parameter. In practice, one can first use the efficient algorithm in Section 2 to compute the entire solution path, then identify the appropriate value of λ that minimizes certain model selection criterion. This avoids the computationally more intensive cross-validation method.

3.1 The Schwarz Information Criterion and Generalized Approximate Cross-Validation

Two commonly used criteria for KQR are the SIC [see (7)] and the GACV [see (8)]. The SIC has been successfully used in quantile regression by several authors (Koenker et al. 1994; He, Ng, and Portnoy 1998). In the parametric setting, Machado (1993) proved that it is a consistent model selection procedure if one of the candidate models is actually correct. However, we acknowledge that using the SIC in the nonparametric setting is ad hoc and requires considerable further investigation. GACV (Yuan 2006) and its earlier cousin ACV (Nychka et al. 1995) are approximations to leave-one-out quantile cross-validation (RCV) (Oh, Nychka, Brown, and Charbonneau 2004),

$$\mathrm{RCV} = \frac{1}{n}\sum_{i=1}^{n} \rho_\tau\big(y_i - \hat{f}^{[-i]}(\mathbf{x}_i)\big),$$

where $\hat{f}^{[-i]}(\cdot)$ denotes the KQR model with the ith data point omitted and RCV is an approximately unbiased estimate of the generalized comparative Kullback–Leibler distance (GCKL),

$$\mathrm{GCKL} = \frac{1}{n}\sum_{i=1}^{n} \mathrm{E}_Y \rho_\tau(Y_i - \hat{f}(\mathbf{x}_i)).$$

Oh et al. (2004) justified using the RCV by arguing that the difference between E(RCV) and the mean squared error (MSE) is approximately a constant not depending on the value of λ, where MSE is defined as

$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n} \mathrm{E}(f(\mathbf{x}_i) - \hat{f}(\mathbf{x}_i))^2.$$

Here we briefly compare the SIC and the GACV. We take the logarithm of the GACV [see (8)], and it becomes

$$\ln\left(\frac{1}{n}\sum_{i=1}^{n} \rho_\tau(y_i - \hat{f}(\mathbf{x}_i))\right) - \ln\left(1 - \frac{df}{n}\right). \qquad (27)$$

Comparing it with the SIC [see (7)], we note that the two differ only in the second term (the penalty term). Figure 3 plots the second term of the SIC (i.e., $\frac{\ln n}{n} df$) and that of (27) [i.e., $-\ln(1 - \frac{df}{n})$] as functions of df/n. As we can see, the SIC penalizes more than the GACV. This explains why the SIC tends to select smoother (i.e., smaller df/n) models than the GACV. As we can also see, when df/n is small, $-\ln(1 - df/n)$ can be approximated by df/n, which leads to the Akaike information criterion for quantile regression (Koenker 2005).

3.2 The Divergence Formula

As we have seen, both the SIC and the GACV depend on a quantity *df*, which is an informative measure of the complexity of a fitted model. For the SIC, Koenker et al. (1994) argued heuristically that in the case of one-dimensional quantile smoothing spline, *df* can be estimated by the number of interpolated y_i's (i.e., $|\mathcal{E}|$), whereas for the GACV and ACV, Nychka et al. (1995) and Yuan (2006) proposed using the divergence formula (9) in Section 1 for *df*. To compute $\sum_{i=1}^n \frac{\partial \hat{f}(\mathbf{x}_i)}{\partial y_i}$, Nychka et al. (1995) and Yuan (2006) approximated the check function with a differentiable function $\rho_{\tau,\delta}(\cdot)$, which differs from $\rho_\tau(\cdot)$ within the interval $(-\delta, \delta)$,

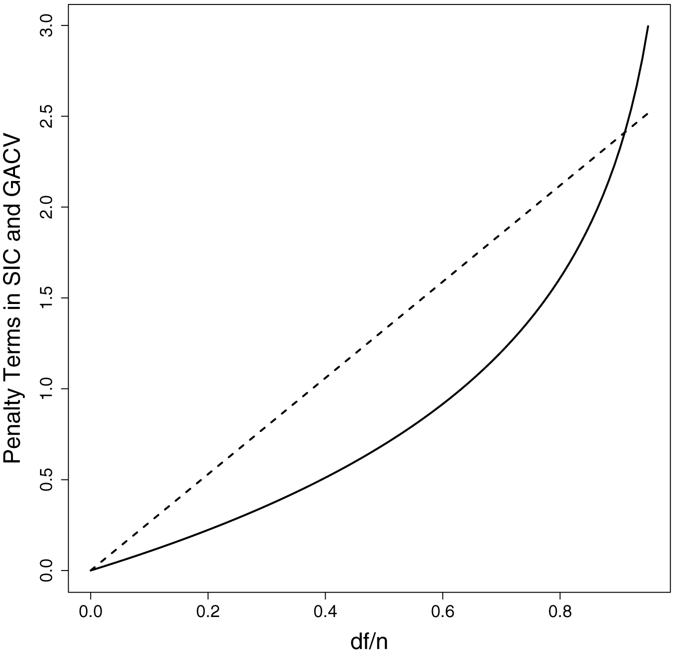

Figure 3. Comparison of the Penalty Term in the SIC (i.e., (ln n/n)df) and That in the GACV [i.e., −ln(1 − df/n)], With n = 200 (—— GACV; - - - SIC).

$$\rho_{\tau,\delta}(r) = \begin{cases} \tau r, & r \ge \delta \\ \tau r^2/\delta, & 0 \le r < \delta \\ (1-\tau) r^2/\delta, & -\delta \le r < 0 \\ -(1-\tau) r, & r < -\delta, \end{cases}$$

where δ is a small positive number.

Notice that the divergence formula (9) measures the sum of the sensitivity of each fitted value with respect to the corresponding observed value. This quantity first appeared under the framework of Stein's unbiased risk estimation (SURE) theory (Stein 1981). Given $\mathbf{x}$, and assuming that y is generated according to a homoscedastic model, we have

$$y \sim (\mu(\mathbf{x}), \sigma^2),$$

where μ is the true mean and σ^2 is the common variance. Then the degrees of freedom of a fitted model $\hat{f}(\mathbf{x})$ can be defined as

$$\sum_{i=1}^{n} \mathrm{cov}(\hat{f}(\mathbf{x}_i), y_i)/\sigma^2. \qquad (28)$$

Stein showed that under mild conditions, $\sum_{i=1}^n \partial\hat{f}(\mathbf{x}_i)/\partial y_i$ is an unbiased estimate of (28). Since then, many authors have argued that $\sum_{i=1}^n \partial\hat{f}(\mathbf{x}_i)/\partial y_i$ can be considered an estimate of the effective dimension for a general modeling procedure (e.g., Efron 1986; Ye 1998; Meyer and Woodroofe 2000; Koenker 2005). For a detailed discussion and complete references, see the article by Efron (2004).

It turns out that in the case of KQR, for every fixed λ, $\sum_{i=1}^n \partial\hat{f}(\mathbf{x}_i)/\partial y_i$ has an extremely simple formula,

$$\sum_{i=1}^{n} \frac{\partial \hat{f}(\mathbf{x}_i)}{\partial y_i} = |\mathcal{E}|. \qquad (29)$$

Therefore, $|\mathcal{E}|$ is a convenient estimate for the effective dimension of $\hat{f}(\mathbf{x})$, which agrees with the heuristic conjecture of Koenker et al. (1994). Plugging (29) into (7) and (8), we arrive at new formulas for the SIC and GACV,

$$\mathrm{SIC}(\lambda) = \ln\left(\frac{1}{n}\sum_{i=1}^{n}\rho_\tau(y_i - \hat{f}(\mathbf{x}_i))\right) + \frac{\ln n}{2n}|\mathcal{E}| \tag{30}$$

and

$$\mathrm{GACV}(\lambda) = \frac{\sum_{i=1}^{n}\rho_\tau(y_i - \hat{f}(\mathbf{x}_i))}{n - |\mathcal{E}|}. \tag{31}$$

We outline the proof of (29) in this section, and defer all of the details to the Appendix. Note that the proof relies closely on our algorithm in Section 2 and follows the spirit of Zou, Hastie, and Tibshirani (2005).

As discussed in Section 2, for a fixed response vector $\mathbf{y} = (y_1, \dots, y_n)^\top$, there is a sequence of λ's, $\infty = \lambda_0 > \lambda_1 > \lambda_2 > \cdots > \lambda_L = 0$, such that in the interior of any interval $(\lambda_{\ell+1}, \lambda_\ell)$, the sets $\mathcal{R}$, $\mathcal{L}$, and $\mathcal{E}$ are constant with respect to λ. These sets change only at each λ_ℓ. We thus define these λ_ℓ's as *event points*.

The essence of Lemmas 1–3 is to show that when we make a small enough perturbation to the dataset, the sets $\mathcal{R}$, $\mathcal{L}$, and $\mathcal{E}$ stay the same.

Lemma 1. For any fixed $\lambda > 0$, the set of $\mathbf{y} = (y_1, \dots, y_n)^\top$ such that λ is an event point is a finite collection of hyperplanes in $\mathbb{R}^n$.

Denote this set as $\mathcal{N}_\lambda$. Then for any $\mathbf{y} \in \mathbb{R}^n \backslash \mathcal{N}_\lambda$, λ is not an event point. Note that $\mathcal{N}_\lambda$ is a null set and $\mathbb{R}^n \backslash \mathcal{N}_\lambda$ is of full measure.

Lemma 2. For any fixed $\lambda > 0$, $\boldsymbol{\theta}_\lambda(\mathbf{y})$ is a continuous function of $\mathbf{y}$.

Lemma 3. For any fixed $\lambda > 0$ and any $\mathbf{y} \in \mathbb{R}^n \backslash \mathcal{N}_\lambda$, the sets $\mathcal{R}$, $\mathcal{L}$, and $\mathcal{E}$ are locally constant with respect to $\mathbf{y}$.

Theorem 1. For any fixed $\lambda > 0$ and any $\mathbf{y} \in \mathbb{R}^n \backslash \mathcal{N}_\lambda$, we have the divergence formula

$$\sum_{i=1}^{n}\frac{\partial \hat{f}(\mathbf{x}_i)}{\partial y_i} = |\mathcal{E}|.$$

4. ASYMPTOTIC THEORY

In this section we develop an asymptotic theory for the KQR. We consider the general problem

$$\min_{f \in \mathcal{F}} \frac{1}{n}\sum_{i=1}^{n}\rho_\tau(y_i - f(\mathbf{x}_i)) + \lambda J(f), \tag{32}$$

where $\mathcal{F}$ is a function class and $J(f)$ is a regularization term that measures the complexity of f. Here $\mathcal{F}$ may depend on n; for example, given a positive definite kernel function $K(\cdot,\cdot)$, $\mathcal{F} = \mathcal{H}_K = \{f : f(\mathbf{x}) = \beta_0 + 1/\lambda \sum_{i=1}^{n}\theta_i K(\mathbf{x}, \mathbf{x}_i)\}$ and $J(f) = \|f\|^2_{\mathcal{H}_K}$.

We outline the main results here and defer the details to the Appendix. Denote

$$\mathrm{R}_\tau(f) = \mathrm{E}\big[\rho_\tau(Y - f(\mathbf{X}))\big] \tag{33}$$

and

$$e_\tau(f, f^*) = R_\tau(f) - R_\tau(f^*), \tag{34}$$

where $f^*(\mathbf{x}) = 100\tau\%$ quantile of Y given $\mathbf{x}$. Hence $f^*(\mathbf{x})$ satisfies $\Pr(Y \le f^*(\mathbf{x})) = \tau$ and $\Pr(Y \ge f^*(\mathbf{x})) = 1 - \tau$ for $\forall\mathbf{x}$. We focus on the asymptotic performance of $\hat{f}(\mathbf{x})$, which is the solution of (32), using $e_\tau(\hat{f}, f^*)$.

The following lemma presents the difference between f and f^* in terms of the check function.

Lemma 4. For any $f \in \mathcal{F}$,

$$\rho_\tau(y - f(\mathbf{x})) - \rho_\tau(y - f^*(\mathbf{x})) = g_\tau(\mathbf{x}, y) + h_\tau(\mathbf{x}, y), \tag{35}$$

where

$$g_\tau(\mathbf{x}, y) = (\tau - 1)\mathbb{I}(y \le f^*(\mathbf{x}))(f^*(\mathbf{x}) - f(\mathbf{x})) \\ + \tau\mathbb{I}(y \ge f^*(\mathbf{x}))(f^*(\mathbf{x}) - f(\mathbf{x}))$$

and

$$h_\tau(\mathbf{x}, y) = \mathbb{I}\big(f^*(\mathbf{x}) \le y \le f(\mathbf{x})\big)(f(\mathbf{x}) - y) \\ + \mathbb{I}\big(f(\mathbf{x}) \le y \le f^*(\mathbf{x})\big)(y - f(\mathbf{x})).$$

We note that $h_\tau(\mathbf{x}, y) \ge 0$ for $\forall\mathbf{x}$ and $\mathrm{E}[g_\tau(\mathbf{X}, Y)] = 0$. Thus $e_\tau(f, f^*) = \mathrm{E}[h_\tau(\mathbf{X}, Y)] \ge 0$ and $f^* = \arg\min_f \mathrm{R}_\tau(f)$. Without loss of generality, we assume that $e_\tau(f, f^*) \le 1$.

Before proceeding further, we define the L_2-metric entropy with bracketing that measures the size of the function class $\mathcal{F}$. Given any $\epsilon > 0$, the set $\{(f_j^\ell, f_j^u), j = 1, \dots, m\}$ is called an "ϵ-bracketing function" of $\mathcal{F}$ if $\|f_j^u - f_j^\ell\|_2 \le \epsilon$ for all $j = 1, \dots, m$, where $\|\cdot\|_2$ is the L_2-norm, and for any $f \in \mathcal{F}$, there exists j such that $f_j^\ell \le f \le f_j^u$. The L_2-metric entropy $\mathrm{H}_B(\epsilon, \mathcal{F})$ of $\mathcal{F}$ with bracketing is then defined as logarithm of the cardinality of ϵ-bracketing function of $\mathcal{F}$ of the smallest size.

We also make two technical assumptions as follows:

- Assumption A. There exist constants $c_1 > 0$ and $0 < \alpha \le 1$ such that

$$\big(\mathrm{E}[h_\tau(\mathbf{X}, Y)]\big)^\alpha \ge c_1 \mathrm{E}|f(\mathbf{X}) - f^*(\mathbf{X})| \tag{36}$$

for any $f \in \mathcal{F}$.

- Assumption B. Denote

$$\mathcal{F}(k) = \{f \in \mathcal{F} : J(f) \le k\},$$
$$\mathcal{F} = \{f \in \mathcal{F} : J(f) < \infty\},$$

and

$$J_0 = \max\{J(f^*), 1\}.$$

We assume that for some positive constants c_2, c_3, and c_4 there exists some $\delta_n > 0$ such that

$$\sup_{k \ge 1}\phi(\delta_n, k) \le c_2 n^{1/2}, \tag{37}$$

where

$$\phi(\delta_n, k) = \frac{1}{D}\int_{c_4 D}^{c_3^{1/2}D^{\alpha/2}} \mathrm{H}_B^{1/2}(u, \mathcal{F}(k))\, du$$

and

$$D = D(\delta_n, \lambda, k) = \min\{\delta_n^2 + \lambda J_0(k - 1), 1\}.$$

Theorem 2. Suppose that Assumptions A and B are satisfied and that $f^* \in \mathcal{F}$ and $\rho_\tau(y - f(\mathbf{x})) \le T$ for some $T > 0$ and for $\forall f \in \mathcal{F}$. Then for any solution $\hat{f}$ of (32) with $\tau(1-\tau) > \eta$, $\eta > 0$, there exists a constant $c_5 > 0$ such that

$$\Pr(e_\tau(\hat{f}, f^*) \ge \delta_n^2) \le 3.5 \exp(-c_5 n(\lambda J_0)^{2-\alpha}), \qquad (38)$$

provided that $\lambda J_0 \le \delta_n^2/2$.

Corollary 1. Under the assumptions of Theorem 2,

$$e_\tau(\hat{f}, f^*) = O_p(\delta_n^2) \qquad \text{and}$$
$$\mathrm{E}|e_\tau(\hat{f}, f^*)| = O(\delta_n^2), \qquad (39)$$

provided that $n(\lambda J_0)^{2-\alpha}$ is bounded away from 0.

Theorem 2 and Corollary 1 provide probability and risk bounds for $e_\tau(\hat{f}, f^*)$. As these bounds indicate, there is a correspondence between the value of λ and the performance. To achieve the best performance, we need to choose the value of λ that provides the best balance between the size of $\mathcal{F}$ and n.

To illustrate the asymptotic theory, we consider the following simple example. Let $y = x + \epsilon$, where $x \in [-1, 1]$ and ϵ is uniformly distributed on $[-1, 1]$. It is easy to verify that $f^*(x) = x + 2\tau - 1$ and $x - 1 \le y \le x + 1$. Let $\mathcal{F}_1 = \{f : f(x) = \beta_0 + 1/\lambda \sum_{i=1}^n \theta_i K(x, x_i), f(x) \in [x-1, x+1], J(f) \le b\}$, where K is the radial basis kernel and $b > 0$ is a constant.

To verify assumption A, we note that

$$\mathrm{E}_{Y|X=x}(f(x) - Y)\mathbb{I}\big(f^*(x) \le Y \le f(x)\big) = \int_{f^*(x)}^{f(x)} \frac{1}{2}(f(x) - u)\,du$$
$$= \frac{1}{4}|f(x) - f^*(x)|^2.$$

Then

$$\mathrm{E}[h_\tau(X, Y)] = \frac{1}{2}\mathrm{E}|f(X) - f^*(X)|^2$$
$$\ge \frac{1}{2}\big(\mathrm{E}|f(X) - f^*(X)|\big)^2.$$

Thus assumption A is satisfied with $\alpha = 1/2$ and $c_1 = 2^{-1/2}$. Using the property of RKHS with a radial basis kernel (Zhou 2002), we have $\mathrm{H}_B(u, \mathcal{F}(k)) = O(\log^2(k/u))$ for any given k. Because $\mathcal{F}_1 \subset \mathcal{F}(b)$, we have $\mathrm{H}_B(u, \mathcal{F}_1) = O(\log^2(1/u))$. Let

$$\phi_1(\delta_n, k) = c_3 \log(1/D)/D^{1-\alpha/2}.$$

Then for some $c > 0$, we have

$$\sup_{k \ge 1} \phi(\delta_n, k) \le \phi_1(\delta_n, 1)$$
$$= c \log(1/\delta_n)/\delta_n^{2-\alpha}.$$

Solving (37), we get the rate

$$\delta_n^2 = \left(\frac{\log^2 n}{n}\right)^{1/(2-\alpha)}, \qquad \text{when } \lambda J_0 \sim \delta_n^2.$$

Using Corollary 1 and $\alpha = 1/2$, we can conclude that $e_\tau(\hat{f}, f^*) = O((\frac{\log^2 n}{n})^{2/3})$ except for a set with probability tending to 0.

5. NUMERICAL RESULTS

In this section we use both simulation and real-world data to demonstrate our algorithm and the selection of λ through the new SIC [see (30)] and the new GACV criterion [see (31)].

5.1 Computational Cost

We first compare the computational cost of the path algorithm with that of the interior point algorithm. Both algorithms have been implemented in the R programming language, and the comparison was done using an IBM notebook PC with an Intel Pentium CPU running at 1.7 GHz, with 256 MB of memory.

Simulations were based on the function given by Yuan (2006),

$$y = 40 \exp\big[8\big((x_1 - .5)^2 + (x_2 - .5)^2\big)\big]$$
$$\times \big(\exp\big[8\big((x_1 - .2)^2 + (x_2 - .7)^2\big)\big]$$
$$+ \exp\big[8\big((x_1 - .7)^2 + (x_2 - .2)^2\big)\big]\big)^{-1} + \epsilon, \quad (40)$$

where x_1 and x_2 are distributed as uniform$(0, 1)$. We used four different error distributions: standard normal, t-distribution with 3 degrees of freedom, double exponential, and a mixture distribution,

$$.1 \cdot \mathrm{N}(0, 5^2) + .9 \cdot \mathrm{N}(0, 1).$$

We used the radial basis kernel with $\sigma = .2$ and generated $n = 50, 100, 200$, and 400 training observations from (40), associated with each of the four error distributions. We considered three different values of τ: 10%, 30%, and 50%. Because the error distributions are all symmetric, these τ's are also representative of the upper quantiles 70% and 90%.

For each simulation dataset, we first ran our path algorithm to compute the entire solution path and retrieved the sequence of event points, $\lambda_0 > \lambda_1 > \lambda_2 > \cdots > \lambda_L$; then for each λ_ℓ, we ran the interior point algorithm to get the corresponding solution. Elapsed CPU times (in seconds) were carefully recorded using the `system.time()` function in R. We repeated this 100 times and computed the average elapsed CPU times and their corresponding standard errors; the results are summarized in Table 1. Because the results for the four error distributions are similar, for simplicity of exposition, we show results only for the normal error distribution. To see the picture more clearly, we also recorded and summarized the number of steps along the path (or the number of event points L) and the average size of the elbow $|\mathcal{E}|$ in Table 2. As we can see, in terms of the elapsed CPU time in computing the entire solution path, our path algorithm dominates the interior point algorithm by several orders. We can also see that the number of steps increases linearly with the size of the training data, and, interestingly, the average elbow size does not seem to increase much as the size of the training dataset increases.

We note that these timings should be viewed with some caution, because they can be sensitive to the details of implementation. We also note that the interior point algorithm was implemented using the *cold start* scheme; that is, for every value of λ_ℓ, the training was done from scratch. To get a more fair

Table 1. Elapsed CPU Times (in seconds)

	Path			Interior point		
n	$\tau = .1$	$\tau = .3$	$\tau = .5$	$\tau = .1$	$\tau = .3$	$\tau = .5$
50	$.09_{(.03)}$	$.13_{(.01)}$	$.14_{(.02)}$	$2.70_{(.32)}$	$4.14_{(.43)}$	$4.74_{(.52)}$
100	$.20_{(.02)}$	$.33_{(.03)}$	$.38_{(.04)}$	$12.19_{(1.58)}$	$21.12_{(1.91)}$	$23.67_{(1.90)}$
200	$.52_{(.09)}$	$.89_{(.11)}$	$.97_{(.10)}$	$104.00_{(10.03)}$	$179.85_{(17.54)}$	$196.29_{(16.17)}$
400	$3.26_{(.21)}$	$4.18_{(.35)}$	$4.58_{(.38)}$	$1{,}986.23_{(56.26)}$	$2{,}561.75_{(68.31)}$	$2{,}706.64_{(75.07)}$

NOTE: The first column n is the size of the training data. `# Steps` is the number of event points; $|\mathcal{E}|$ is the average elbow size at the event points. All results are averages of 100 independent simulations. The numbers in the parentheses are the corresponding standard errors.

comparison of our path algorithm and the interior point algorithm, we also computed the average elapsed CPU time of the interior point algorithm for a *single* value of λ_ℓ, defined as

$$\text{average elapsed CPU time for a single value of } \lambda = \frac{\text{total elapsed CPU time}}{\text{number of steps}}.$$

The results are summarized in Table 3. Comparing Tables 1, 2, and 3, we can see that it takes our algorithm about 1–3 times as long to compute the entire solution path as it takes the interior point algorithm to compute a single solution. We also note that the path algorithm gives a full presentation of the solution path without knowing the locations of the event points a priori, whereas the interior point algorithm requires a sequence of prespecified λ_ℓ's.

5.2 Simulation Data

The setups of the function and the error distributions are similar to those in Section 5.1. We generated 200 training observations from (40), associated with each of the four error distributions, along with 10,000 validation observations and 10,000 test observations. We used the radial basis kernel with $\sigma = .2$. Again, we considered three different values of τ: $10\%, 30\%$, and 50%. We then found the λ's that minimized the SIC and the GACV criterion. We used the validation set to select the gold standard λ's that minimized the prediction error. Using these λ's, we calculated the prediction error and the mean absolute deviation with the test data for each criterion. Suppose that the fitted quantile function is $\hat{f}(\mathbf{x})$ and the true quantile function is $f(\mathbf{x})$; then the prediction error and the mean absolute deviation are defined as

$$\text{prediction error} = \frac{1}{10{,}000} \sum_{i=1}^{10{,}000} \rho_\tau(y_i - \hat{f}(\mathbf{x}_i))$$

and

$$\text{mean absolute deviation} = \frac{1}{10{,}000} \sum_{i=1}^{10{,}000} |f(\mathbf{x}_i) - \hat{f}(\mathbf{x}_i)|.$$

We repeated this 100 times and computed the average prediction error, the average mean absolute deviation, and their corresponding standard errors. We also compared the degrees of freedom selected by the three different methods. The results are summarized in Tables 4–6.

We can see, several trends in the results when using (30) and (31) to select the regularization parameter λ:

1. In terms of the prediction error and the mean absolute deviation, both the SIC and the GACV perform close to the gold standard, and they get closer to the gold standard as τ gets closer to .5.
2. As τ gets closer to .5, the performance of the SIC and the GACV also get closer.
3. The SIC always performs slightly better than the GACV.
4. The variance of the GACV is always slightly greater than that of the SIC.
5. The SIC tends to select a simpler model than the gold standard, whereas the GACV tends to select a more complicated model than the gold standard.

5.3 Real Data

In this section we consider applications to a real dataset: the annual salary of baseball players.

Annual Salary of Baseball Players. This is a widely analyzed dataset, provided by He et al. (1998), comprising records of 263 North American major league baseball players for the 1986 season. We reanalyzed the data to further demonstrate our algorithm and compare the SIC and GACV criterion. Following He et al. (1998) and Yuan (2006), we used 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 (in thousands of dollars). The multiplicative spline kernel was used. The results are given in Figure 4. As shown, for the 50% quantile surface, the SIC and the GACV give similar results, but for the 25% and the 75% quantile surfaces, the SIC fits are again smoother than the GACV fits.

Table 2. Characteristics of the Path

	Average $\lvert\mathcal{E}\rvert$			Average number of steps		
n	$\tau = .1$	$\tau = .3$	$\tau = .5$	$\tau = .1$	$\tau = .3$	$\tau = .5$
50	$19.70_{(2.16)}$	$18.24_{(1.35)}$	$19.09_{(1.38)}$	$49.70_{(5.18)}$	$78.30_{(7.62)}$	$91.30_{(9.45)}$
100	$24.46_{(1.78)}$	$24.18_{(1.33)}$	$24.63_{(1.68)}$	$94.64_{(11.90)}$	$172.20_{(14.91)}$	$199.62_{(14.59)}$
200	$27.01_{(1.91)}$	$27.35_{(1.97)}$	$27.73_{(1.55)}$	$179.50_{(16.87)}$	$342.8_{(18.23)}$	$381.6_{(26.66)}$
400	$43.70_{(3.58)}$	$42.33_{(2.24)}$	$42.39_{(2.65)}$	$539.25_{(47.71)}$	$822.75_{(53.87)}$	$892.75_{(62.68)}$

NOTE: The first column n is the size of the training data. `# Steps` is the number of event points; $|\mathcal{E}|$ is the average elbow size at the event points. All results are averages of 100 independent simulations. The numbers in the parentheses are the corresponding standard errors.

Table 3. Average Elapsed CPU Time (in seconds) of the Interior Point Algorithm for a Single Value of λ_ℓ

n	$\tau = .1$	$\tau = .3$	$\tau = .5$
50	.054$_{(.004)}$	.053$_{(.002)}$	.052$_{(.001)}$
100	.129$_{(.003)}$	.123$_{(.003)}$	.119$_{(.003)}$
200	.579$_{(.017)}$	.524$_{(.076)}$	.514$_{(.020)}$
400	3.692$_{(.378)}$	3.167$_{(.223)}$	3.073$_{(.274)}$

Table 4. Simulation Example, $\tau = 10\%$

	SIC	GACV	Gold standard
Prediction error			
Normal	.216$_{(.016)}$	.355$_{(.165)}$	.208$_{(.011)}$
DE	.324$_{(.065)}$	.532$_{(.212)}$	.296$_{(.010)}$
T3	.352$_{(.073)}$	.567$_{(.226)}$	.323$_{(.011)}$
Mixture	.382$_{(.059)}$	.590$_{(.220)}$	.350$_{(.012)}$
Mean absolute deviation			
Normal	.505$_{(.075)}$	.857$_{(.379)}$	.478$_{(.072)}$
DE	.761$_{(.133)}$	1.235$_{(.437)}$	.672$_{(.114)}$
T3	.842$_{(.212)}$	1.494$_{(.496)}$	.724$_{(.133)}$
Mixture	.904$_{(.337)}$	1.491$_{(.610)}$	.697$_{(.165)}$
Degrees of freedom			
Normal	23.2$_{(4.8)}$	52.5$_{(28.1)}$	22.0$_{(3.5)}$
DE	22.1$_{(6.8)}$	53.1$_{(28.4)}$	18.4$_{(3.8)}$
T3	21.9$_{(9.2)}$	60.2$_{(27.1)}$	18.0$_{(4.2)}$
Mixture	22.7$_{(12.4)}$	56.1$_{(26.6)}$	18.1$_{(4.2)}$

NOTE: Prediction errors of the true conditional quantile functions are .173 (normal), .258 (DE), .285 (T3), and .311 (mixture).

Table 5. Simulation Example, $\tau = 30\%$

	SIC	GACV	Gold standard
Prediction error			
Normal	.393$_{(.016)}$	.470$_{(.111)}$	.384$_{(.011)}$
DE	.503$_{(.019)}$	.603$_{(.167)}$	.490$_{(.012)}$
T3	.551$_{(.018)}$	.644$_{(.195)}$	.538$_{(.011)}$
Mixture	.570$_{(.022)}$	.606$_{(.093)}$	.557$_{(.013)}$
Mean absolute deviation			
Normal	.398$_{(.058)}$	.683$_{(.347)}$	.368$_{(.052)}$
DE	.462$_{(.071)}$	.737$_{(.361)}$	.423$_{(.061)}$
T3	.493$_{(.088)}$	.779$_{(.453)}$	.440$_{(.070)}$
Mixture	.478$_{(.082)}$	.602$_{(.414)}$	.405$_{(.051)}$
Degrees of freedom			
Normal	18.9$_{(4.5)}$	52.3$_{(23.1)}$	25.0$_{(4.2)}$
DE	17.2$_{(4.1)}$	44.1$_{(21.1)}$	23.6$_{(3.5)}$
T3	15.8$_{(4.5)}$	43.9$_{(25.3)}$	23.3$_{(3.8)}$
Mixture	15.5$_{(4.2)}$	34.2$_{(17.9)}$	24.3$_{(4.2)}$

NOTE: Prediction errors of the true conditional quantile functions are .346 (normal), .448 (DE), .493 (T3), and .515 (mixture).

Table 6. Simulation Example, $\tau = 50\%$

	SIC	GACV	Gold standard
Prediction error			
Normal	.448$_{(.017)}$	.481$_{(.077)}$	.438$_{(.010)}$
DE	.561$_{(.021)}$	.591$_{(.111)}$	.545$_{(.012)}$
T3	.615$_{(.023)}$	.655$_{(.136)}$	.600$_{(.013)}$
Mixture	.619$_{(.022)}$	.626$_{(.050)}$	.606$_{(.010)}$
Mean absolute deviation			
Normal	.375$_{(.062)}$	.538$_{(.237)}$	.348$_{(.048)}$
DE	.397$_{(.073)}$	.463$_{(.238)}$	.355$_{(.060)}$
T3	.429$_{(.077)}$	.482$_{(.214)}$	.384$_{(.054)}$
Mixture	.425$_{(.075)}$	.490$_{(.247)}$	.384$_{(.051)}$
Degrees of freedom			
Normal	18.3$_{(4.0)}$	43.1$_{(19.6)}$	26.0$_{(4.2)}$
DE	16.2$_{(3.3)}$	30.9$_{(13.3)}$	26.6$_{(4.8)}$
T3	15.6$_{(3.7)}$	28.6$_{(12.9)}$	24.6$_{(5.0)}$
Mixture	15.5$_{(3.5)}$	28.8$_{(15.5)}$	24.7$_{(4.8)}$

NOTE: Prediction errors of the true conditional quantile functions are .398 (normal), .503 (DE), .553 (T3), and .565 (mixture).

6. CONCLUSION

In this article we have proposed an efficient algorithm that computes the entire regularization path of the KQR; we have derived a simple formula for the effective dimension of the fitted KQR model, which can be used to select the regularization parameter λ; we also have developed an asymptotic theory for the KQR.

We acknowledge that in this article we have taken the *loss* + *penalty* approach to the problem of nonparametric estimation of conditional quantile functions. There is also an extensive literature covering the same problem using the local polynomial approach (e.g., Stone 1977; Chaudhuri 1991; Yu and Jones 1998).

Finally, we would like to point out an interesting direction where our work can be extended. As Koenker et al. (1994) pointed out, in the case of L_1 loss + L_1 penalty, the solution path is piecewise constant in τ (for fixed λ). We plan to investigate whether a similar result holds for the KQR, that is, whether the solution path is also piecewise linear in τ. When changing τ, one disturbing problem in quantile regression is that the fitted quantile curves (or surfaces) can cross each other (He 1997). For example, in the right column of Figure 4, the fitted median surface is higher than the 75% quantile surface in the region of `Year` > 20 and `Home Run` > 30. Although this is due to lack of data in that region, avoiding such confusion is of practical importance.

APPENDIX: PROOFS

Proof of Lemma 1

For any fixed $\lambda > 0$, suppose that $\mathcal{R}$, $\mathcal{L}$, and $\mathcal{E}$ are given; then we have

$$\frac{1}{\lambda}\left(\beta_{0,\lambda} + \sum_{i\in\mathcal{E}} \theta_i K(\mathbf{x}_k, \mathbf{x}_i) - (1-\tau)\sum_{i\in\mathcal{L}} K(\mathbf{x}_k, \mathbf{x}_i) + \tau \sum_{i\in\mathcal{R}} K(\mathbf{x}_k, \mathbf{x}_i)\right) = y_k, \qquad \forall k \in \mathcal{E}, \tag{A.1}$$

and

$$\sum_{i\in\mathcal{E}} \theta_i - (1-\tau) n_{\mathcal{L}} + \tau n_{\mathcal{R}} = 0. \tag{A.2}$$

These can be reexpressed as

$$\begin{pmatrix} 0 & \mathbf{1}^\top \\ \mathbf{1} & \mathbf{K}_\mathcal{E} \end{pmatrix} \begin{pmatrix} \beta_{0,\lambda} \\ \boldsymbol{\theta}_\mathcal{E} \end{pmatrix} = \begin{pmatrix} b \\ \lambda \mathbf{y}_\mathcal{E} - \mathbf{a} \end{pmatrix},$$

where $\mathbf{K}_\mathcal{E}$ is a $n_\mathcal{E} \times n_\mathcal{E}$ matrix, with entries equal to $K(\mathbf{x}_k, \mathbf{x}_{k'})$, $k, k' \in \mathcal{E}$, and $\boldsymbol{\theta}_\mathcal{E}$ and $\mathbf{y}_\mathcal{E}$ are vectors of length $n_\mathcal{E}$, with elements equal to θ_k and y_k, $k \in \mathcal{E}$. Here $\mathbf{a}$ is also a vector of length $n_\mathcal{E}$, with elements equal to $-(1-\tau)\sum_{i\in\mathcal{L}} K(\mathbf{x}_k, \mathbf{x}_i) + \tau \sum_{i\in\mathcal{R}} K(\mathbf{x}_k, \mathbf{x}_i)$, $k \in \mathcal{E}$, and b is a scalar $b = -(1-\tau)n_\mathcal{L} + \tau n_\mathcal{R}$. Note that λ, $\mathcal{R}$, $\mathcal{L}$, and $\mathcal{E}$ are fixed, $\mathbf{K}_\mathcal{E}$, $\mathbf{a}$, and b are also fixed.

Then $\beta_{0,\lambda}$ and $\boldsymbol{\theta}_\mathcal{E}$ can be expressed as

$$\begin{pmatrix} \beta_{0,\lambda} \\ \boldsymbol{\theta}_\mathcal{E} \end{pmatrix} = \tilde{\mathbf{K}} \begin{pmatrix} b \\ \lambda \mathbf{y}_\mathcal{E} - \mathbf{a} \end{pmatrix},$$

where

$$\tilde{\mathbf{K}} = \begin{pmatrix} 0 & \mathbf{1}^\top \\ \mathbf{1} & \mathbf{K}_\mathcal{E} \end{pmatrix}^{-1}.$$

Note that $\beta_{0,\lambda}$ and $\boldsymbol{\theta}_\mathcal{E}$ are linear in $\mathbf{y}_\mathcal{E}$.

Now, corresponding to the three events listed at the beginning of Section 2.3, if λ is an event point, then one of the following conditions must be satisfied:

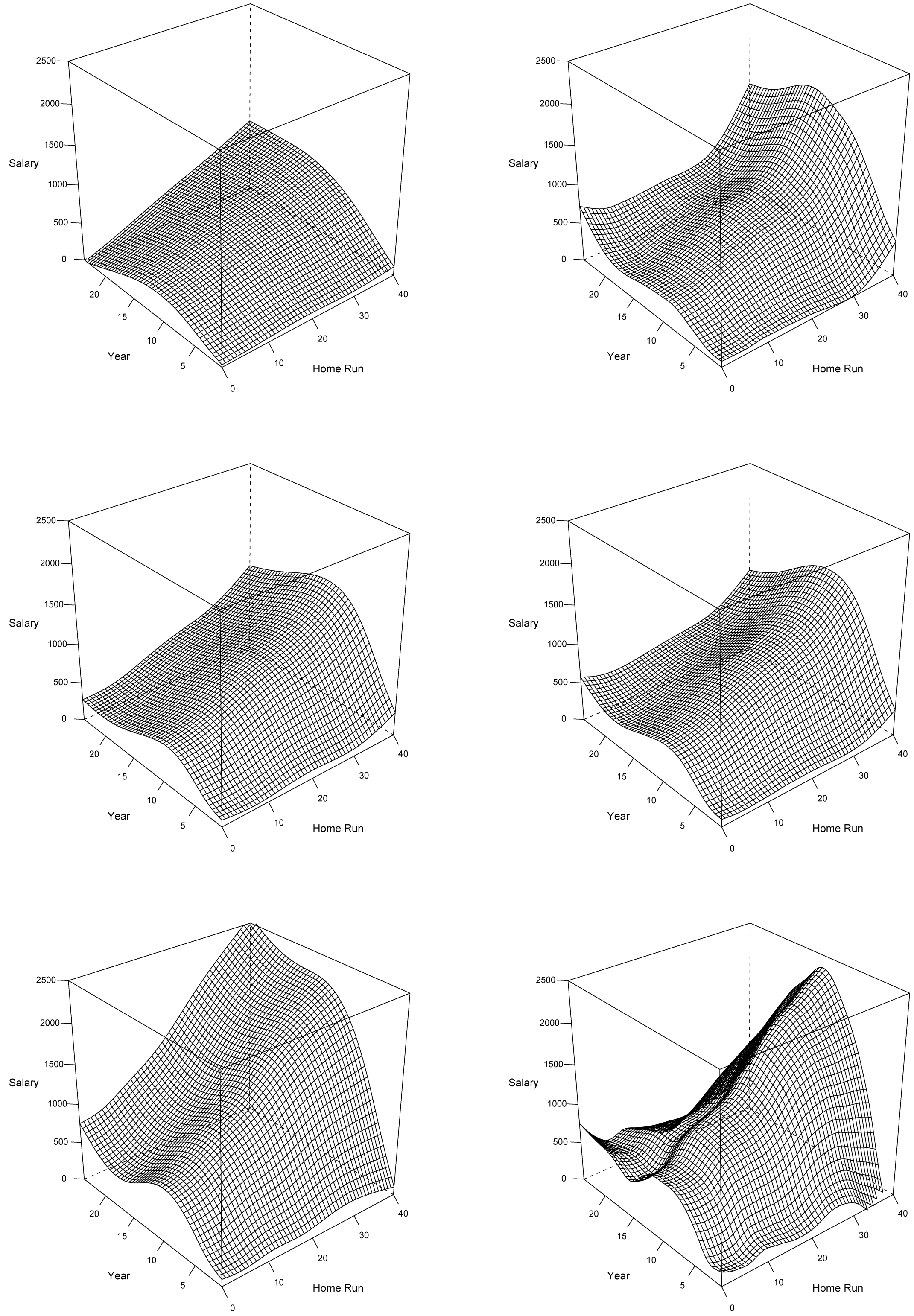

Figure 4. Baseball Data. The left column contains the SIC results; the right column, the GACV results. The first row corresponds to the 25% quantile surfaces, the middle row corresponds to the 50% quantile surfaces, and the third row corresponds to the 75% quantile surfaces.

1. $\theta_k = -(1-\tau)$, $\exists k \in \mathcal{E}$
2. $\theta_k = \tau$, $\exists k \in \mathcal{E}$
3. $y_k = \frac{1}{\lambda}(\beta_{0,\lambda} + \sum_{i\in\mathcal{E}} \theta_i K(\mathbf{x}_k, \mathbf{x}_i) - (1-\tau)\sum_{i\in\mathcal{L}} K(\mathbf{x}_k, \mathbf{x}_i) + \tau \sum_{i\in\mathcal{R}} K(\mathbf{x}_k, \mathbf{x}_i))$, $\exists k \in \mathcal{L} \cup \mathcal{R}$.

For any fixed λ, $\mathcal{R}$, $\mathcal{L}$, and $\mathcal{E}$, each of the foregoing conditions defines a hyperplane of $\mathbf{y}$ in $\mathbb{R}^n$. Taking into account all possible combinations of $\mathcal{R}$, $\mathcal{L}$, and $\mathcal{E}$, the set of $\mathbf{y}$ such that λ is an event point is a collection of finite number of hyperplanes.

Proof of Lemma 2

For any fixed $\lambda > 0$ and any fixed $\mathbf{y}_0 \in \mathbb{R}^n$, we wish to show that if a sequence $\mathbf{y}_m$ converges to y_0, then $\boldsymbol{\theta}(\mathbf{y}_m)$ converges to $\boldsymbol{\theta}(\mathbf{y}_0)$. Because $\boldsymbol{\theta}(\mathbf{y}_m)$ is bounded, it is equivalent to show that for every converging subsequence, say $\boldsymbol{\theta}(\mathbf{y}_{m_k})$, the subsequence converges to $\boldsymbol{\theta}(\mathbf{y}_0)$. Suppose that $\boldsymbol{\theta}(\mathbf{y}_{m_k})$ converges to $\boldsymbol{\theta}_\infty$; we show that $\boldsymbol{\theta}_\infty = \boldsymbol{\theta}(\mathbf{y}_0)$. Denote (6) as $g(\boldsymbol{\theta}(\mathbf{y}), \mathbf{y})$, and let

$$\Delta g(\boldsymbol{\theta}(\mathbf{y}), \mathbf{y}, \mathbf{y}') = g(\boldsymbol{\theta}(\mathbf{y}), \mathbf{y}) - g(\boldsymbol{\theta}(\mathbf{y}), \mathbf{y}').$$

Then we have

$$\begin{aligned} g(\boldsymbol{\theta}(\mathbf{y}_0), \mathbf{y}_0) &= g\big(\boldsymbol{\theta}(\mathbf{y}_0), \mathbf{y}_{m_k}\big) + \Delta g\big(\boldsymbol{\theta}(\mathbf{y}_0), \mathbf{y}_0, \mathbf{y}_{m_k}\big) \\ &\geq g\big(\boldsymbol{\theta}(\mathbf{y}_{m_k}), \mathbf{y}_{m_k}\big) + \Delta g\big(\boldsymbol{\theta}(\mathbf{y}_0), \mathbf{y}_0, \mathbf{y}_{m_k}\big) \\ &= g\big(\boldsymbol{\theta}(\mathbf{y}_{m_k}), \mathbf{y}_0\big) + \Delta g\big(\boldsymbol{\theta}(\mathbf{y}_{m_k}), \mathbf{y}_{m_k}, \mathbf{y}_0\big) \\ &\quad + \Delta g\big(\boldsymbol{\theta}(\mathbf{y}_0), \mathbf{y}_0, \mathbf{y}_{m_k}\big). \end{aligned} \tag{A.3}$$

Using the fact that $|a| - |b| \leq |a - b|$ and $\mathbf{y}_{m_k} \to \mathbf{y}_0$, it is easy to show that for sufficiently large m_k, we have

$$\Delta g\big(\boldsymbol{\theta}(\mathbf{y}_{m_k}), \mathbf{y}_{m_k}, \mathbf{y}_0\big) + \Delta g\big(\boldsymbol{\theta}(\mathbf{y}_0), \mathbf{y}_0, \mathbf{y}_{m_k}\big) \leq c\big\|\mathbf{y}_0 - \mathbf{y}_{m_k}\big\|_1,$$

where $c > 0$ is a constant. Furthermore, using $\mathbf{y}_{m_k} \to \mathbf{y}_0$ and $\boldsymbol{\theta}(\mathbf{y}_{m_k}) \to \boldsymbol{\theta}_\infty$, we reduce (A.3) to

$$g(\boldsymbol{\theta}(\mathbf{y}_0), \mathbf{y}_0) \geq g(\boldsymbol{\theta}_\infty, \mathbf{y}_0).$$

Because $\boldsymbol{\theta}(\mathbf{y}_0)$ is the unique minimizer of $g(\boldsymbol{\theta}, \mathbf{y}_0)$, we have that $\boldsymbol{\theta}_\infty = \boldsymbol{\theta}(\mathbf{y}_0)$.

Proof of Lemma 3

For any fixed $\lambda > 0$ and any fixed $\mathbf{y}_0 \in \mathbb{R}^n \backslash \mathcal{N}_\lambda$, because $\mathbb{R}^n \backslash \mathcal{N}_\lambda$ is an open set, we can always find a small enough $\epsilon > 0$ such that $\mathrm{Ball}(\mathbf{y}_0, \epsilon) \subset \mathbb{R}^n \backslash \mathcal{N}_\lambda$. Thus λ is not an event point for any $\mathbf{y} \in \mathrm{Ball}(\mathbf{y}_0, \epsilon)$. We claim that if ϵ is small enough, then the sets $\mathcal{R}$, $\mathcal{L}$, and $\mathcal{E}$ stay the same for all $\mathbf{y} \in \mathrm{Ball}(\mathbf{y}_0, \epsilon)$. Consider $\mathbf{y}$ and $\mathbf{y}_0$. Let $\mathcal{R}_\mathbf{y}, \mathcal{L}_\mathbf{y}, \mathcal{E}_\mathbf{y}, \mathcal{R}_0, \mathcal{L}_0$, and $\mathcal{E}_0$ denote the corresponding sets and $\boldsymbol{\theta}^\mathbf{y}, f^\mathbf{y}$, $\boldsymbol{\theta}^0$, and f^0 denote the corresponding fits. For any $i \in \mathcal{E}_0$, because λ is not an event point, we have $-(1-\tau) < \theta_i^0 < \tau$. Therefore, by continuity, we also have $-(1-\tau) < \theta_i^\mathbf{y} < \tau$, $i \in \mathcal{E}_0$ for $\mathbf{y}$ close enough to $\mathbf{y}_0$ or, equivalently, $\mathcal{E}_0 \subseteq \mathcal{E}_\mathbf{y}$, $\forall \mathbf{y} \in \mathrm{Ball}(\mathbf{y}_0, \epsilon)$ for small enough ϵ. Similarly, for any $i \in \mathcal{R}_0$, because $y_i^0 - f^0(\mathbf{x}_i) > 0$, again, by continuity, we have $y_i - f^\mathbf{y}(\mathbf{x}_i) > 0$ for $\mathbf{y}$ close enough to $\mathbf{y}_0$ or, equivalently, $\mathcal{R}_0 \subseteq \mathcal{R}_\mathbf{y}$, $\forall \mathbf{y} \in \mathrm{Ball}(\mathbf{y}_0, \epsilon)$ for small enough ϵ. The same applies to $\mathcal{L}_0$ and $\mathcal{L}_\mathbf{y}$ as well. Overall, we then must have $\mathcal{E}_0 = \mathcal{E}_\mathbf{y}$, $\mathcal{R}_0 = \mathcal{R}_\mathbf{y}$, and $\mathcal{L}_0 = \mathcal{L}_\mathbf{y}$ for all $\mathbf{y} \in \mathrm{Ball}(\mathbf{y}_0, \epsilon)$ when ϵ is small enough.

Proof of Theorem 1

Using Lemma 3, we know that there exists $\epsilon > 0$ such that for all $\mathbf{y} \in \mathrm{Ball}(\mathbf{y}, \epsilon)$, the sets $\mathcal{R}$, $\mathcal{L}$, and $\mathcal{E}$ stay the same. This implies that for points in $\mathcal{E}$, we have

$$\frac{\partial \hat{f}(\mathbf{x}_i)}{\partial y_i} = 1, \qquad i \in \mathcal{E}.$$

Furthermore, from (A.1) and (A.2), we can see that for points in $\mathcal{R}$ and $\mathcal{L}$, their θ_i's are fixed at either τ or $\tau - 1$, and the other θ_i's are determined by $\mathbf{y}_\mathcal{E}$. Hence

$$\frac{\partial \hat{f}(\mathbf{x}_i)}{\partial y_i} = 0, \qquad i \in \mathcal{R} \cup \mathcal{L}.$$

Overall, we have

$$\sum_{i=1}^n \frac{\partial \hat{f}(\mathbf{x}_i)}{\partial y_i} = |\mathcal{E}|.$$

Proof of Lemma 4

The desired result can be proved directly by expressing $\rho_\tau(y - f(\mathbf{x})) - \rho_\tau(y - f^*(\mathbf{x}))$ using the definition of the check function with all possible orderings among f, f^*, and y.

Proof of Theorem 2

We first introduce some notations. Denote $z_i = (\mathbf{x}_i, y_i)$, $Z_i = (\mathbf{X}_i, Y_i)$, $\ell_\tau(f, z_i) = \rho_\tau(y_i - f(\mathbf{x}_i))$, and $\tilde{\ell}_\tau(f, z_i) = \ell_\tau(f, z_i) + \lambda J(f)$. We let

$$A_{ij} = \big\{f \in \mathcal{F} : 2^{i-1}\delta_n^2 \leq e_\tau(f, f^*) < 2^i\delta_n^2, 2^{j-1}J_0 \leq J(f) < 2^j J_0\big\}, \qquad i, j = 1, 2, \ldots,$$

and

$$A_{i0} = \big\{f \in \mathcal{F} : 2^{i-1}\delta_n^2 \leq e_\tau(f, f^*) < 2^i\delta_n^2, J(f) < J_0\big\}, \qquad i = 1, 2, \ldots.$$

We then define the scaled empirical process, $\mathrm{E}_n(\tilde{\ell}_\tau(f, Z) - \tilde{\ell}_\tau(f^*, Z))$, as

$$\begin{aligned} &\mathrm{E}_n\big(\tilde{\ell}_\tau(f, Z) - \tilde{\ell}_\tau(f^*, Z)\big) \\ &\quad = n^{-1}\sum_{i=1}^n \big(\tilde{\ell}_\tau(f, Z) - \tilde{\ell}_\tau(f^*, Z) - \mathrm{E}[\tilde{\ell}_\tau(f, Z) - \tilde{\ell}_\tau(f^*, Z)]\big) \\ &\quad = \mathrm{E}_n[\ell_\tau(f, Z) - \ell_\tau(f^*, Z)]. \end{aligned}$$

To bound $\Pr(e_\tau(\hat{f}, f^*) \geq \delta_n^2)$, we use theorem 3 of Shen and Wong (1994), a large deviation inequality for empirical processes by controlling the corresponding mean and variance. First, we note that

$$\{e_\tau(\hat{f}, f^*) \geq \delta_n^2\} \subset \left\{\sup_{\{f \in \mathcal{F} : e_\tau(f, f^*) \geq \delta_n^2\}} n^{-1}\sum_{i=1}^n \big(\tilde{\ell}_\tau(f^*, Z_i) - \tilde{\ell}_\tau(f, Z_i)\big) \geq 0\right\}.$$

Hence

$$\Pr(e_\tau(\hat{f}, f^*) \geq \delta_n^2) \leq \overset{*}{\Pr}\left(\sup_{\{f \in \mathcal{F} : e_\tau(f, f^*) \geq \delta_n^2\}} n^{-1}\sum_{i=1}^n \big(\tilde{\ell}_\tau(f^*, Z_i) - \tilde{\ell}_\tau(f, Z_i)\big) \geq 0\right), \tag{A.4}$$

where $\Pr^*$ denotes the outer probability measure. We denote the probability on the right side of (A.4) as I. To bound I, it is sufficient to bound $\Pr^*(\sup_{\{f \in A_{ij}\}} n^{-1}\sum_{i=1}^n (\tilde{\ell}_\tau(f^*, Z_i) - \tilde{\ell}_\tau(f, Z_i)) \geq 0)$ for each i, j. Toward this end, we need some inequalities regarding the first and second moments of $\tilde{\ell}_\tau(f^*, Z_i) - \tilde{\ell}_\tau(f, Z_i)$ for $f \in A_{ij}$.

For the first moment, using the definition of A_{ij}, we have

$$\inf_{A_{ij}} \mathrm{E}[\tilde{\ell}_\tau(f, Z_i) - \tilde{\ell}_\tau(f^*, Z_i)] \geq 2^{i-1}\delta_n^2 + \lambda(2^{j-1} - 1)J_0 \overset{\text{def}}{=} M(i, j) \tag{A.5}$$

and

$$\inf_{A_{i0}} \mathrm{E}[\tilde{\ell}_\tau(f, Z_i) - \tilde{\ell}_\tau(f^*, Z_i)] \geq 2^{i-2}\delta_n^2 \overset{\text{def}}{=} M(i, 0), \tag{A.6}$$

where $i, j \geq 1$. Note that (A.6) follows from the assumption that $\lambda J_0 \leq \delta_n^2/2$ and the fact that $2^i - 1 \leq 2^{i-1}$.

For the second moment, because ℓ_τ is bounded by T, we have

$$\mathrm{E}[\ell_\tau(f, Z) - \ell_\tau(f^*, Z)]^2 \leq T \cdot \mathrm{E}|\ell_\tau(f, Z) - \ell_\tau(f^*, Z)|.$$

Furthermore, using Lemma 4, we get

$$\mathrm{E}|\ell_\tau(f, Z) - \ell_\tau(f^*, Z)| \leq \mathrm{E}\big[|g_\tau(\mathbf{X}, Y)| + |h_\tau(\mathbf{X}, Y)|\big]. \tag{A.7}$$

Note that $\mathrm{E}|h_\tau(\mathbf{X}, Y)| = \mathrm{E}[h_\tau(\mathbf{X}, Y)] = e_\tau(f, f^*)$, and using Assumption A,

$$\begin{aligned} \mathrm{E}|g_\tau(\mathbf{X}, Y)| &= \tau(1-\tau)\mathrm{E}|f(\mathbf{X}) - f^*(\mathbf{X})| \\ &\leq c_1^{-1}\tau(1-\tau)\big(\mathrm{E}[h_\tau(\mathbf{X}, Y)]\big)^\alpha \\ &= c_1^{-1}\tau(1-\tau)(e_\tau(f, f^*))^\alpha. \end{aligned}$$

Thus we have

$$\mathrm{E}[\ell_\tau(f,Z)-\ell_\tau(f^*,Z)]^2 \le T(c_1^{-1}\tau(1-\tau)+1)(e_\tau(f,f^*))^\alpha. \tag{A.8}$$

The first and second moments can then be connected as follows. For any $f\in A_{ij}$, we have

$$\begin{aligned}\sup_{A_{ij}}\mathrm{E}[\ell_\tau(f,Z)-\ell_\tau(f^*,Z)]^2 &\le T(c_1^{-1}\tau(1-\tau)+1)(2^i\delta_n^2)^\alpha\\ &\le c_3(M(i,j))^\alpha \overset{\text{def}}{=} v(i,j)^2,\end{aligned}$$

where $c_3=4^\alpha T(c_1^{-1}\tau(1-\tau)+1)$.

Now we are ready to bound I. Using (A.5) and (A.6), we get

$$\begin{aligned}I &\le \sum_{i\ge1,j\ge0}\Pr{}^*\Big(\sup_{A_{ij}}\mathrm{E}_n\big(\tilde\ell_\tau(f^*,Z_i)-\tilde\ell_\tau(f,Z_i)\big)\ge 0\Big)\\ &\le \sum_{i\ge1,j\ge0}\Pr{}^*\Big(\sup_{A_{ij}}\mathrm{E}_n\big(\ell_\tau(f^*,Z_i)-\ell_\tau(f,Z_i)\big)\ge M(i,j)\Big)\\ &= I_1+I_2,\end{aligned}$$

where

$$I_1=\sum_{i,j\ge1}\Pr{}^*\Big(\sup_{A_{ij}}\mathrm{E}_n\big(\ell_\tau(f^*,Z_i)-\ell_\tau(f,Z_i)\big)\ge M(i,j)\Big)$$

and

$$I_2=\sum_{i\ge1}\Pr{}^*\Big(\sup_{A_{i0}}\mathrm{E}_n\big(\ell_\tau(f^*,Z_i)-\ell_\tau(f,Z_i)\big)\ge M(i,0)\Big).$$

Next, we use the large deviation inequality of Shen and Wong (1994) to bound I_1 and I_2. We first verify the required conditions (4.5)–(4.7) in theorem 3 of Shen and Wong (1994).

To compute the metric entropy in (4.7) of Shen and Wong (1994), we need to construct a bracketing function of $\ell_\tau(f^*,Z)-\ell_\tau(f,Z)$. Denote $\mathcal{F}_{\ell_\tau}(k)=\{\ell_\tau(f,z)-\ell_\tau(f^*,z):f\in\mathcal{F}(k)\}$. We let

$$\ell_j^\ell=\min\{\ell_\tau(f_j^\ell,z),\ell_\tau(f_j^u,z),0\}-\ell_\tau(f^*,z)$$

and

$$\ell_j^u=\max\{\ell_\tau(f_j^\ell,z),\ell_\tau(f_j^u,z)\}-\ell_\tau(f^*,z).$$

Then for any $f\in\mathcal{F}$ with $J(f)\le 2^j$, there exists $j\in\{1,\dots,m\}$ such that $f_j^\ell\le f\le f_j^u$, which implies that $\ell_j^\ell\le\ell_\tau(f,z)-\ell_\tau(f^*,z)\le\ell_j^u$. Hence $\{(\ell_k^\ell,\ell_k^u),k=1,\dots,m\}$ is a bracket function set of $\ell_\tau(f,z)-\ell_\tau(f^*,z)$. Furthermore, using the property of ℓ_τ, we have

$$\|\ell_k^u-\ell_k^\ell\|_2\le\max(\tau,1-\tau)\|f_j^u-f_j^\ell\|_2\le\|f_j^u-f_j^\ell\|_2.$$

Hence $H_B(u,\mathcal{F}_{\ell_\tau}(2^j))\le H_B(u,\mathcal{F}(2^j))$.

Using the fact that $\int_{aM(i,j)}^{v(i,j)}H_B^{1/2}(u,\mathcal{F}(2^j))\,du/M(i,j)$ is nonincreasing in i and $M(i,j)$, we have

$$\begin{aligned}&\int_{aM(i,j)}^{v(i,j)}H_B^{1/2}(u,\mathcal{F}(2^j))\,du/M(i,j)\\ &\qquad\le\int_{aM(1,j)}^{\sqrt{c_3}M(1,j)^{\alpha/2}}H_B^{1/2}(u,\mathcal{F}(2^j))\,du/M(1,j)\\ &\qquad\le\phi(\delta_n,2^j),\end{aligned}$$

where $a=\epsilon/32$ with ϵ as defined later. Thus (4.7) of Shen and Wong (1994) holds with $M=n^{1/2}M(i,j)$ and $v=v(i,j)^2$, as does (4.5). Without loss of generality, we assume that $M(i,j)\le1$ and $v(i,j)^2\le1$. Then, $M(i,j)/v(i,j)^2\le c_3^{-1}=\epsilon/(4T)$ implies (4.6) of Shen and Wong (1994) with $\epsilon=4Tc_3^{-1}=2^{2-2\alpha}(c_1^{-1}\tau(1-\tau)+1)^{-1}$. Note that $0<\delta_n\le1$ and $\lambda J_0\le\delta_n^2/2$. Thus, an application of theorem 3 of Shen and Wong (1994) yields

$$\begin{aligned}I_1&\le\sum_{j=1}^\infty\sum_{i=1}^\infty 3\exp\Big(-\frac{(1-\epsilon)nM(i,j)^2}{2(4v(i,j)^2+M(i,j)T/3)}\Big)\\ &\le\sum_{j=1}^\infty\sum_{i=1}^\infty 3\exp\big(-c_5nM(i,j)^{2-\alpha}\big)\\ &\le\sum_{j=1}^\infty\sum_{i=1}^\infty 3\exp\big(-c_5n\big[(2^{i-1}\delta_n^2)^{2-\alpha}+((2^{j-1}-1)\lambda J_0)^{2-\alpha}\big]\big)\\ &\le\sum_{j=1}^\infty\sum_{i=1}^\infty 3\exp\big(-c_5n[i(\lambda J_0)^{2-\alpha}+(j-1)(\lambda J_0)^{2-\alpha}]\big)\\ &\le 3\exp(-c_5n(\lambda J_0)^{2-\alpha})/\big[1-\exp(-c_5n(\lambda J_0)^{2-\alpha})\big]^2,\end{aligned}$$

where $c_5>0$ is a generic constant. I_2 can be bounded in a similar way. Finally, we have

$$I\le 6\exp(-c_5n(\lambda J_0)^{2-\alpha})/\big[1-\exp(-c_5n(\lambda J_0)^{2-\alpha})\big]^2,$$

which implies that $I^{1/2}\le(5/2+I^{1/2})\exp(-c_5n(\lambda J_0)^{2-\alpha})$. Because $I\le I^{1/2}\le1$, we finally have

$$I\le 3.5\exp(-c_5n(\lambda J_0)^{2-\alpha}),$$

which is the desired result.

Proof of Corollary 1

To show that $e_\tau(\hat f,f^*)=O_p(\delta_n^2)$, it is sufficient to show that

$$\Pr(e(\hat f,f^*)\ge G\delta_n^2)\le 3.5\exp(-c_5nG(\lambda J_0)^{2-\alpha})$$

for any $G\ge1$.

Toward this end, we need to modify the proof of Theorem 2 only slightly. We redefine

$$\begin{aligned}A_{ij}=\big\{f\in\mathcal{F}:2^{i-1}\delta_n^2G\le e_\tau(f,f^*)<2^i\delta_n^2G,&\\ 2^{j-1}J_0G\le J(f)<2^jJ_0G\big\},&\qquad i,j=1,2,\dots,\end{aligned}$$

and

$$A_{i0}=\big\{f\in\mathcal{F}:2^{i-1}\delta_n^2G\le e_\tau(f,f^*)<2^i\delta_n^2G,J(f)<J_0G\big\},\qquad i=1,2,\dots.$$

Using these new definitions, an analogous proof to that of Theorem 2 can be obtained with $M(i,j)=2^{i-1}\delta_n^2G+\lambda(2^{j-1}-1)J_0G$, and the desired result then follows.

[Received September 2005. Revised July 2006.]

REFERENCES

Bloomfield, P., and Steiger, W. (1983), *Least Absolute Deviations: Theory, Applications and Algorithms*, Boston: Birkhäuser-Verlag.

Bosch, R., Ye, Y., and Woodworth, G. (1995), "A Convergent Algorithm for Quantile Regression With Smoothing Splines," *Computational Statistics & Data Analysis*, 19, 613–630.

Chaudhuri, P. (1991), "Nonparametric Estimates of Regression Quantiles and Their Local Bahadur Representation," *The Annals of Statistics*, 2, 760–777.

Cole, T., and Green, P. (1992), "Smoothing Reference Centile Curves: The LMS Method and Penalized Likelihood," *Statistics in Medicine*, 11, 1305–1319.

Efron, B. (1986), "How Biased Is the Apparent Error Rate of a Prediction Rule?" *Journal of the American Statistical Association*, 81, 461–470.

——— (2004), "The Estimation of Prediction Error: Covariance Penalties and Cross-Validation," *Journal of the American Statistical Association*, 99, 619–632.

Gu, C. (2002), *Smoothing Spline ANOVA Models*, New York: Springer-Verlag.

Hastie, T., Rosset, S., Tibshirani, R., and Zhu, J. (2004), "The Entire Regularization Path for the Support Vector Machine," *Journal of Machine Learning Research*, 5, 1391–1415.

He, X. (1997), “Quantile Curves Without Crossing,” *The American Statistician*, 51, 186–192.

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

Heagerty, P., and Pepe, M. (1999), “Semiparametric Estimation of Regression Quantiles With Application to Standardizing Weight for Height and Age in U.S. Children,” *Journal of the Royal Statistical Society*, Ser. 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.

Kimeldorf, G., and Wahba, G. (1971), “Some Results on Tchebycheffian Spline Functions,” *Journal of Mathematical Analysis and Applications*, 33, 82–95.

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

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

Koenker, R., and Geling, R. (2001), “Reappraising Medfly Longevity: A Quantile Regression Survival Analysis,” *Journal of the American Statistical Association*, 96, 458–468.

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

Koenker, R., Ng, P., and Portnoy, S. (1994), “Quantile Smoothing Splines,” *Biometrika*, 81, 673–680.

Machado, J. (1993), “Robust Model Selection and M-Estimation,” *Econometric Theory*, 9, 478–493.

Meyer, M., and Woodroofe, M. (2000), “On the Degrees of Freedom in Shape-Restricted Regression,” *The Annals of Statistics*, 28, 1083–1104.

Nychka, D., Gray, G., Haaland, P., Martin, D., and O’Connell, M. (1995), “A Nonparametric Regression Approach to Syringe Grading for Quality Improvement,” *Journal of the American Statistical Association*, 90, 1171–1178.

Oh, H., Nychka, D., Brown, T., and Charbonneau, P. (2004), “Period Analysis of Variable Stars by Robust Smoothing,” *Journal of the Royal Statistical Society*, Ser. A, 53, 15–30.

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

Shen, X., and Wong, W. (1994), “Convergence Rate of Sieve Estimates,” *The Annals of Statistics*, 22, 580–615.

Stein, C. (1981), “Estimation of the Mean of a Multivariate Normal Distribution,” *The Annals of Statistics*, 9, 1135–1151.

Stone, C. (1977), “Consistent Nonparametric Regression” (with discussion), *The Annals of Statistics*, 5, 595–645.

Wahba, G. (1990), *Spline Models for Observational Data*, Philadelphia: SIAM.

Yang, S. (1999), “Censored Median Regression Using Weighted Empirical Survival and Hazard Functions,” *Journal of the American Statistical Association*, 94, 137–145.

Ye, J. (1998), “On Measuring and Correcting the Effects of Data Mining and Model Selection,” *Journal of the American Statistical Association*, 93, 120–131.

Yu, K., and Jones, M. (1998), “Local Linear Regression Quantile Estimation,” *Journal of the American Statistical Association*, 93, 228–238.

Yu, K., Lu, Z., and Stander, J. (2003), “Quantile Regression: Applications and Current Research Areas,” *The Statistician*, 52, 331–350.

Yuan, M. (2006), “GACV for Quantile Smoothing Splines,” *Computational Statistics and Data Analysis*, 5, 813–829.

Zhou, D.-X. (2002), “The Covering Number in Learning Theory,” *Journal of Complexity*, 18, 739–767.

Zou, H., Hastie, T., and Tibshirani, R. (2005), “On the Degrees of Freedom of the Lasso,” technical report, Stanford University, Dept. of Statistics.