Quantile Regression in Reproducing Kernel Hilbert Spaces

Youjuan Li

Department of Statistics

University of Michigan

Ann Arbor, MI 48109

Yufeng Liu

Department of Statistics and Operations Research

Carolina Center for Genome Sciences

University of North Carolina

Chapel Hill, NC 27599

Ji Zhu*

Department of Statistics

University of Michigan

Ann Arbor, MI 48109

June 22, 2006

*Address for correspondence: Ji Zhu, 439 West Hall, 1085 South University, Ann Arbor, MI 48109-1107. E-mail: jizhu@umich.edu.

Abstract

In this paper we consider quantile regression in reproducing kernel Hilbert spaces, which we refer to as 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.

Keywords: Degrees of freedom; Metric entropy; Model selection; Quadratic programming; Quantile regression; RKHS

1 Introduction

Classical regression methods have mainly focused 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, such as, reference charts in medicine (Cole & Green 1992, Heagerty & Pepe 1999), survival analysis (Yang 1999, Koenker & Geling 2001) and economics (Hendricks & Koenker 1992, Koenker & Hallock 2001). For comprehensive reviews on the quantile regression, we refer the authors to Koenker & Hallock (2001), Yu, Lu & Stander (2003) and the exceptionally well-written book Koenker (2005).

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

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

where q is a positive integer and $\rho_\tau(r)$ is the so called check function of Koenker & Bassett (1978) (Figure 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) show 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 & Steiger 1983, Nychka, Gray, Haaland, Martin & 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.

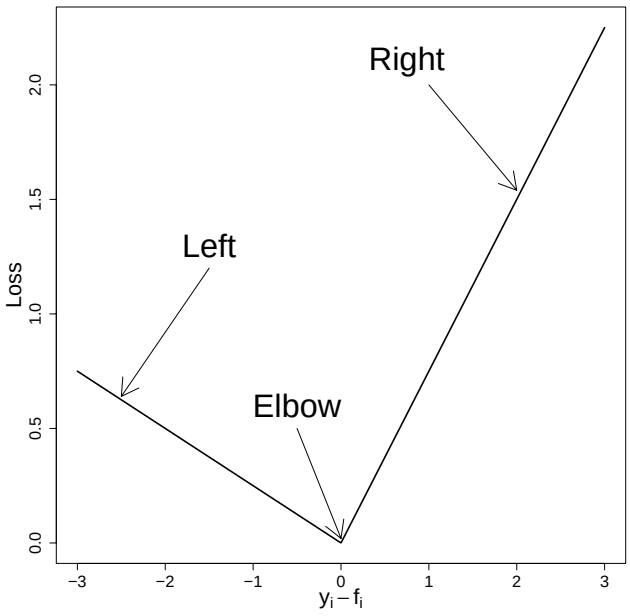

Figure 1: The check function, with $\tau = 0.75$.

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

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

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

$$\begin{aligned} d\text{th Degree polynomial} \quad &: \quad K(\boldsymbol{x}, \boldsymbol{x}') = (1 + \langle \boldsymbol{x}, \boldsymbol{x}' \rangle)^d \\ \text{Radial basis} \quad &: \quad K(\boldsymbol{x}, \boldsymbol{x}') = \exp(-\|\boldsymbol{x} - \boldsymbol{x}'\|^2 / 2\sigma^2) \end{aligned}$$

where d and σ are pre-specified parameters.

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

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

Notice that we write $\hat{f}(\boldsymbol{x})$ in a way that involves λ explicitly, and we will see later that $\theta_i \in [-(1-\tau), \tau]$. Given the format of the solution (5), we can in turn re-write (4) in a finite form:

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

which we will refer to as the kernel quantile regression (KQR).

The KQR model (6) can be transformed into a quadratic programming problem, hence most commercially available packages can be used to solve the KQR. In the past years, many specific algorithms for the KQR have also been developed, for example, the interior point algorithm (Bosch, Ye & Woodworth 1995), and the pseudo-data algorithm (Nychka et al. 1995). All these algorithms solve the KQR for a pre-fixed regularization parameter λ. As in any smoothing problem, choice of the regularization parameter λ is critical. In practice, people usually pre-specify a finite set of values for λ that covers a wide range, then either use a separate validation data set or certain model selection criterion to pick a value for λ that gives the best performance among the pre-specified set. Two commonly used criteria for KQR are the Schwarz information criterion (Schwarz 1978, Koenker et al. 1994) (SIC) and the generalized approximate cross-validation criterion (Yuan 2006) (GACV):

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

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

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

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

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

In this paper, we make three main contributions:

- We show that the solution $\boldsymbol{\theta}(\lambda)$ is *piecewise linear* as a function of λ, and we derive an efficient algorithm that computes the *exact entire solution path* $\{\boldsymbol{\theta}(\lambda), 0 \leq \lambda \leq \infty\}$, 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 usage in selecting the regularization parameter λ.

- We also 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 & 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$(0, 0.2^2)$. 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 paper 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 the divergence (9) is equal to the number of interpolated y_i's for the KQR. In Section 4, we develop an asymptotic theory for the KQR. In Section 5, we present numerical results on both simulation data and real world data. We conclude the paper with Section 6.

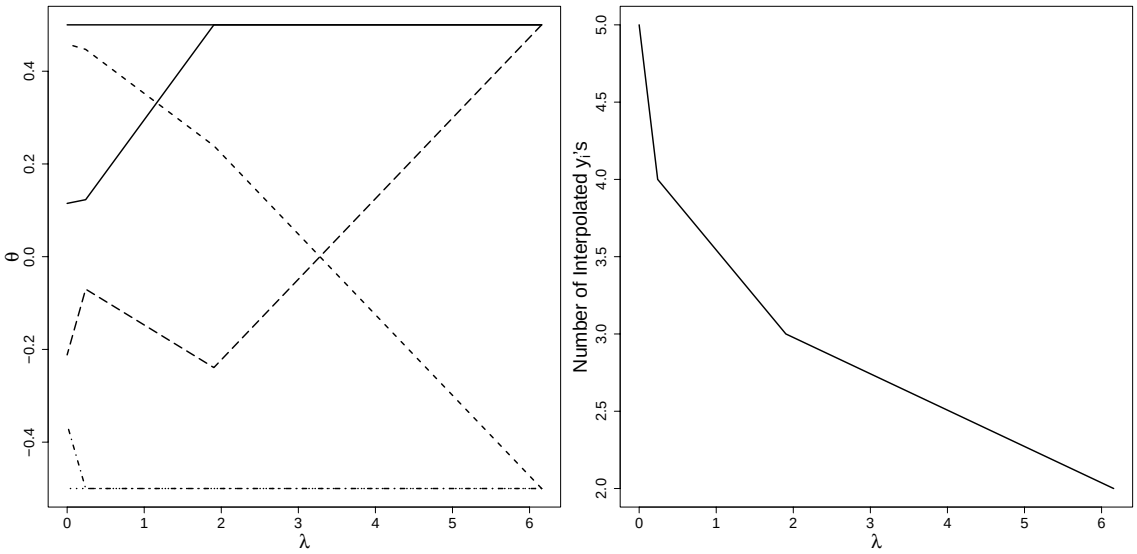

Figure 2: Illustrative example. Left panel: A subset of the solution path $\boldsymbol{\theta}(\lambda)$ as a function of λ. Right panel: How the number of interpolated y_i's changes with λ.

2 Algorithm

2.1 Problem Setup

Criterion (6) can be re-written in an equivalent way:

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

$$\text{subject to} \quad -\zeta_i \leq y_i - f(\boldsymbol{x}_i) \leq \xi_i \tag{11}$$

$$\zeta_i, \xi_i \geq 0, \quad i = 1, \ldots, n \tag{12}$$

where

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

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

For the rest of the paper, we assume the data points $(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_n, y_n)$ are in general positions and the kernel matrix $\boldsymbol{K}$ is positive definite. Then the above setting gives the

Lagrangian primal function

$$L_p : \quad \tau \sum_{i=1}^{n} \xi_i + (1-\tau) \sum_{i=1}^{n} \zeta_i + \frac{1}{2\lambda} \boldsymbol{\theta}^{\mathsf{T}} \boldsymbol{K} \boldsymbol{\theta} + \sum_{i=1}^{n} \alpha_i (y_i - f(\boldsymbol{x}_i) - \xi_i) - \sum_{i=1}^{n} \gamma_i (y_i - f(\boldsymbol{x}_i) + \zeta_i) - \sum_{i=1}^{n} \kappa_i \xi_i - \sum_{i=1}^{n} \rho_i \zeta_i, \tag{13}$$

where $\alpha_i, \gamma_i, \kappa_i$ and ρ_i are non-negative Lagrange multipliers. Setting the derivatives of L_p to zero we arrive at

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

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

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

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

and the Karush-Kuhn-Tucker conditions are

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

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

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

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

Since the Lagrange multipliers must be non-negative, 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 zero, and vice versa. These lead to the following relationships:

$$\begin{aligned}
&y_i - f(\boldsymbol{x}_i) > 0 \quad \Rightarrow \quad \alpha_i = \tau, \qquad \xi_i > 0, \quad \gamma_i = 0, \qquad \zeta_i = 0; \\
&y_i - f(\boldsymbol{x}_i) < 0 \quad \Rightarrow \quad \alpha_i = 0, \qquad \xi_i = 0, \quad \gamma_i = 1 - \tau, \qquad \zeta_i > 0; \\
&y_i - f(\boldsymbol{x}_i) = 0 \quad \Rightarrow \quad \alpha_i \in [0, \tau], \quad \xi_i = 0, \quad \gamma_i \in [0, 1 - \tau], \quad \zeta_i = 0.
\end{aligned}$$

Notice 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 that will be used later when we calculate the regularization path of the KQR:

- $\mathcal{E} = \{i : y_i - f(\boldsymbol{x}_i) = 0, -(1-\tau) \leq \theta_i \leq \tau\}$ (Elbow)
- $\mathcal{L} = \{i : y_i - f(\boldsymbol{x}_i) < 0, \theta_i = -(1-\tau)\}$ (Left of the elbow)
- $\mathcal{R} = \{i : y_i - f(\boldsymbol{x}_i) > 0, \theta_i = \tau\}$ (Right of the elbow)

For points in $\mathcal{L}$ and $\mathcal{R}$, the values of θ_i are known; therefore, the algorithm will focus 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 zero, keeping track of the locations of all data points relative to the elbow along the way. As λ decreases, points move from 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. Since all points at the elbow have $y_i - f(\boldsymbol{x}_i) = 0$, we can establish a path for their θ_i, and this set will stay 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 $\hat{f}(\boldsymbol{x}) = \beta_0$. We can determine the value of β_0 via a simple one-dimensional optimization. For exposition simplicity, we focus on the case that all the values of y_i are distinct and ordered $y_1 < y_2 < \cdots < y_n$. We distinguish between two cases: the initial β_0 is unique, and the initial β_0 is non-unique.

Case 1: the initial β_0 is unique

This happens when $n\tau$ is a non-integer, for example, when $\tau = 0.5$ 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 it is y_{i^*}. All data points are therefore initially divided into the three sets:

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

- $\mathcal{R} = \{i : y_i > y_{i^*}\}$

From (15), we have

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

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

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

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

When a data point enters the elbow, it satisfies

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

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

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

Case 2: the initial β_0 is non-unique

This happens when $n\tau$ is an integer, for example, when $\tau = 0.5$ 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 they are $[y_{i^*}, y_{j^*}]$.

Although β_0 is not unique, all the θ_i's are fully determined, i.e.,

- $\theta_i = -(1-\tau), y_i \leq y_{i^*}$
- $\theta_i = \tau, y_i \geq y_{j^*}$

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

- $\mathcal{E} = \emptyset$

- $\mathcal{L} = \{i : y_i \leq y_{i^*}\}$
- $\mathcal{R} = \{i : y_i \geq y_{j^*}\}$

For sufficiently large values of λ, we have

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

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

When λ decreases, by continuity and the balance between θ_i, $\mathcal{L}$ and $\mathcal{R}$ will stay the same. Therefore

$$\begin{aligned} y_i - \beta_0 - \frac{1}{\lambda}g(\boldsymbol{x}_i) &\leq 0, \quad i \in \mathcal{L} \\ y_i - \beta_0 - \frac{1}{\lambda}g(\boldsymbol{x}_i) &\geq 0, \quad i \in \mathcal{R}\end{aligned}$$

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

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

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 zero,

2.3 The Regularization Path

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

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 has occurred, 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 $\hat{f}(\boldsymbol{x}_i) = y_i$ for all points in $\mathcal{E}$, we can calculate θ_i for these points.

From event 3, we may reach the situation that $\mathcal{E}$ 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 zero.

We use the subscript ℓ to index the sets above 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. Also let f^ℓ be the function at this point. We define for convenience $\beta_{0,\lambda} = \lambda \cdot \beta_0$ and hence $\beta_{0,\lambda}^\ell = \lambda^\ell \cdot \beta_0^\ell$. Then since

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

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

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

where the reduction occurs in the second line, since 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 $|\mathcal{E}^\ell| = n_{\mathcal{E}}^\ell$, so for the $n_{\mathcal{E}}^\ell$ points staying at the elbow, we have that

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

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

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

Also, 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 we can use to solve for each of the $n_{\mathcal{E}}^{\ell}+1$ unknown variables ν_i and ν_0.

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

$$\begin{aligned} \nu_0 \mathbf{1} + \boldsymbol{K}^{\ell} \boldsymbol{\nu} &= (\lambda - \lambda^{\ell}) \boldsymbol{y}_{\mathcal{E}}^{\ell} && (22) \\ \boldsymbol{\nu}^{\mathsf{T}} \mathbf{1} &= 0 && (23) \end{aligned}$$

Simplifying further, if we let

$$\boldsymbol{A}^{\ell} = \begin{pmatrix} 0 & \mathbf{1}^{\mathsf{T}} \\ \mathbf{1} & \boldsymbol{K}^{\ell} \end{pmatrix}, \quad \boldsymbol{\nu}_0 = \begin{pmatrix} \nu_0 \\ \boldsymbol{\nu} \end{pmatrix}, \quad \text{and} \quad \boldsymbol{y}_0 = \begin{pmatrix} 0 \\ \boldsymbol{y}_{\mathcal{E}}^{\ell} \end{pmatrix},$$

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

$$\boldsymbol{A}^{\ell} \boldsymbol{\nu}_0 = (\lambda - \lambda^{\ell}) \boldsymbol{y}_0.$$

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

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

to give us

$$\begin{aligned} \theta_i &= \theta_i^{\ell} + (\lambda - \lambda^{\ell}) b_i, \quad \forall i \in \mathcal{E}^{\ell}, && (24) \\ \beta_{0,\lambda} &= \beta_{0,\lambda}^{\ell} + (\lambda - \lambda^{\ell}) b_0. && (25) \end{aligned}$$

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

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

where

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

Given λ_{ℓ}, equations (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

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

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

2.4 Computational Cost

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

According 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 $\mathcal{E}^\ell$, then the approximate computational cost of the algorithm is $O\left(cn^2m + nm^2\right)$.

3 The Effective Dimension of KQR

It is well known that an appropriate value of λ is crucial for the performance of the fitted model in any smoothing problems. One advantage of computing the entire solution path is to facilitate the 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 SIC and the GACV

Two commonly used criteria for KQR are the SIC (7) and the GACV (8). The SIC criterion has been successfully used in the quantile regression literature by several authors (Koenker et al. 1994, He, Ng & 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 the use of the SIC in the nonparametric setting is ad hoc and needs considerable further investigation. The GACV (Yuan 2006) and its earlier cousin ACV (Nychka et al. 1995) are approximations to the leave-one-out quantile cross-validation (RCV) (Oh, Nychka, Brown & Charbonneau 2004)

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

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

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

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

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

Here we make a brief comparison between the SIC and the GACV. We take the logarithm of the GACV (8), and it becomes

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

Comparing it with the SIC (7), we notice that they 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 (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 AIC 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, where df is an informative measure of the complexity of a fitted model. For the SIC, Koenker et al. (1994)

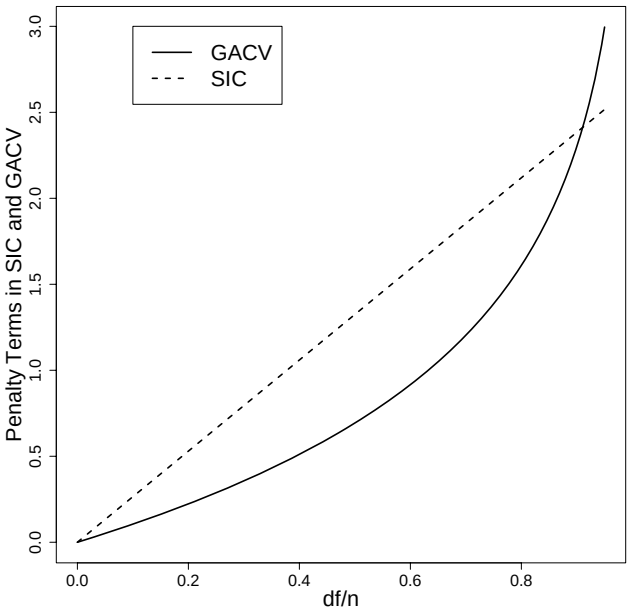

Figure 3: Comparison of the penalty term in the SIC, i.e., $\frac{\ln n}{n} df$ and that in the GACV, i.e., $-\ln(1-\frac{df}{n})$, with $n = 200$.

argue 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., $|\mathcal{E}|$; while for the GACV and ACV, Nychka et al. (1995) and Yuan (2006) propose to use the divergence formula (9) in Section 1 for *df*. To compute $\sum_{i=1}^{n} \frac{\partial \hat{f}(\boldsymbol{x}_i)}{\partial y_i}$, Nychka et al. (1995) and Yuan (2006) approximate the check function with a differentiable function $\rho_{\tau,\delta}(\cdot)$, which differs from $\rho_\tau(\cdot)$ within the interval $(-\delta, \delta)$:

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

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 $\boldsymbol{x}$, assuming y is generated according to a homoskedastic model:

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

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

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

Stein shows that under mild conditions, $\sum_{i=1}^{n} \partial \hat{f}(\boldsymbol{x}_i)/\partial y_i$ is an unbiased estimate of (28). Ever since then, many authors have argued that $\sum_{i=1}^{n} \partial \hat{f}(\boldsymbol{x}_i)/\partial y_i$ can be considered as an estimate of the effective dimension for a general modeling procedure, for example, Efron (1986), Ye (1998), Meyer & Woodroofe (2000) and Koenker (2005). For detailed discussion and complete references, we refer the readers to Efron (2004).

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

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

Therefore, $|\mathcal{E}|$ is a convenient estimate for the effective dimension of $\hat{f}(\boldsymbol{x})$, 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 the GACV:

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

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

We outline the proof of (29) in this section, and leave all the details in the appendix. We make a note that the proof relies closely on our algorithm in Section 2, and follows the spirit of Zou, Hastie & Tibshirani (2005).

As we have seen in Section 2, for a fixed response vector $\boldsymbol{y} = (y_1, \ldots, y_n)^\mathsf{T}$, 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 only change 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 $\boldsymbol{y} = (y_1, \ldots, y_n)^\top$ 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 $\boldsymbol{y} \in \mathbb{R}^n \backslash \mathcal{N}_\lambda$, λ is not an event point. Notice $\mathcal{N}_\lambda$ is a null set, and $\mathbb{R}^n \backslash \mathcal{N}_\lambda$ is of full measure.

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

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

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

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

4 Asymptotic Theory

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

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

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

We outline the main results here and leave all the details in the appendix.

Denote

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

and

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

where $f^*(\boldsymbol{x}) = 100\tau\%$ quantile of Y given $\boldsymbol{x}$. Hence $f^*(\boldsymbol{x})$ satisfies $\Pr(Y \leq f^*(\boldsymbol{x})) = \tau$ and $\Pr(Y \geq f^*(\boldsymbol{x})) = 1 - \tau$ for $\forall \boldsymbol{x}$. We will focus on the asymptotic performance of $\hat{f}(\boldsymbol{x})$, 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 $f \in \mathcal{F}$,*

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

where

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

and

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

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

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 $\left\{(f_j^\ell, f_j^u), j = 1, \ldots, m\right\}$ is called an ϵ-bracketing function of $\mathcal{F}$ if $\|f_j^u - f_j^\ell\|_2 \le \epsilon$ for all $j = 1, \cdots, m$, where $\|\cdot\|_2$ is the L_2-norm, and for any $f \in \mathcal{F}$, there exists a 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

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

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

- **Assumption B:** Denote

$$\begin{aligned} \mathcal{F}(k) &= \{f \in \mathcal{F} : J(f) \le k\} \\ \mathcal{F} &= \{f \in \mathcal{F} : J(f) < \infty\} \\ J_0 &= \max\{J(f^*), 1\} \end{aligned}$$

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

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

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 assumptions A and B are satisfied, and suppose* $f^* \in \mathcal{F}$, *and* $\rho_\tau(y - f(\boldsymbol{x})) \leq T$ *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*

$$\Pr\left(e_\tau(\hat{f}, f^*) \geq \delta_n^2\right) \leq 3.5 \exp\left(-c_5 n (\lambda J_0)^{2-\alpha}\right) \tag{38}$$

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

Corollary 1 *Under the assumptions of Theorem 2,*

$$e_\tau(\hat{f}, f^*) = O_p(\delta_n^2) \quad \textit{and} \quad \mathrm{E}\left|e_\tau(\hat{f}, f^*)\right| = O(\delta_n^2) \tag{39}$$

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

Theorem 2 and Corollary 1 provide probability and risk bounds for $e_\tau(\hat{f}, f^*)$. As shown in the bounds, there is a correspondence between the value of λ and the performance. In order to achieve the best performance, we need to choose λ that gives 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 $x - 1 \leq y \leq x + 1$. Let $\mathcal{F}_1 = \{f : f(x) = \beta_0 + 1/\lambda \cdot \sum_{i=1}^n \theta_i K(x, x_i), f(x) \in [x-1, x+1], J(f) \leq b\}$ where K is the radial basis kernel, and $b > 0$ is a constant.

To verify assumption A, we notice

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

Then

$$\begin{aligned} \mathrm{E}\left[h_\tau(X,Y)\right] &= \frac{1}{2}\mathrm{E}|f(X)-f^*(X)|^2 \\ &\ge \frac{1}{2}\left(\mathrm{E}|f(X)-f^*(X)|\right)^2. \end{aligned}$$

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. Since $\mathcal{F}_1 \subset \mathcal{F}(b)$, 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

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

Solving (37), we get the rate

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

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 zero.

5 Numerical Results

In this section, we use both simulation data and real world data to demonstrate our algorithm and the selection of λ via the new SIC criterion (30) and the new GACV criterion (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 on an IBM notebook that has an Intel Pentium CPU running at 1.7GHz, and the system has 256MB memory.

Simulations were based on the function found in Yuan (2006):

$$y = \frac{40 \exp\left[8\left((x_1 - 0.5)^2 + (x_2 - 0.5)^2\right)\right]}{\exp\left[8\left((x_1 - 0.2)^2 + (x_2 - 0.7)^2\right)\right] + \exp\left[8\left((x_1 - 0.7)^2 + (x_2 - 0.2)^2\right)\right]} + \epsilon, \qquad (40)$$

where x_1, x_2 are distributed as Uniform(0,1). Four different error distributions were used: standard Normal, t-distribution with degrees of freedom 3, Double Exponential, and a mixture distribution:

$$0.1 \cdot N(0, 5^2) + 0.9 \cdot N(0, 1).$$

We used the radial basis kernel with $\sigma = 0.2$, and we generated $n = 50, 100, 200, 400$ training observations from (40), associated with each of the four error distributions. We considered three different values of τ: $10\%, 30\%$ and 50%. Since the error distributions are all symmetric, these τ's are also representative of the upper quantiles 70% and 90%.

For each simulation data set, 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 recorded carefully 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. Since the results for the four error distributions are similar, for exposition simplicity, we only show results 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 1. 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 doesn't seem to increase much when the size of the training data increases.

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

$$\text{Average Elapsed CPU Time for a Single Value of } \lambda = \frac{\text{Total Elapsed CPU Time}}{\#\text{ Steps}}.$$

The results are summarized in Table 2. Comparing Tables 1 and 2, 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 make a note that the path algorithm gives a full presentation of the solution path without knowing the locations of the event points a priori, while the interior point algorithm needs a sequence of pre-specified λ_ℓ's.

5.2 Simulation Data

The setup 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. The radial basis kernel with $\sigma = 0.2$ was used. Again, we considered three different values of τ: $10\%, 30\%$ and 50%. We then found the λ's that minimized the SIC criterion and the GACV criterion, respectively. The validation set was used to select the *gold standard* λ which 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 the fitted quantile function is $\hat{f}(\boldsymbol{x})$ and the true quantile function is $f(\boldsymbol{x})$, the prediction error and the mean absolute deviation are defined as the following:

$$\begin{aligned}
\text{Prediction Error} &= \frac{1}{10,000}\sum_{i=1}^{10,000}\rho_\tau(y_i - \hat{f}(\boldsymbol{x}_i)) \\
\text{Mean Absolute Deviation} &= \frac{1}{10,000}\sum_{i=1}^{10,000}\left|f(\boldsymbol{x}_i) - \hat{f}(\boldsymbol{x}_i)\right|
\end{aligned}$$

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 3 – 5.

As we can see, there are 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 closely to the gold standard, and they get closer to the gold standard as τ gets closer to 0.5.

2. As τ gets closer to 0.5, 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 bigger than that of the SIC.

5. The SIC tends to select a simpler model than the gold standard, while the GACV tends to select a more complicated model than the gold standard.

5.3 Real Data

In this section, we consider the 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), that consists of records of 263 North American Major League baseball players for the 1986 season. We re-analyzed the data to further demonstrate our algorithm and compare the SIC and GACV criteria. 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 shown in Figure 4. As we can see,

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.

6 Conclusion

In this paper, 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 have also developed an asymptotic theory for the KQR.

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

Finally, we would like to point out an interesting direction where our work can be extended. As Koenker et al. (1994) point 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, i.e., 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 with 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, it is of practical importance to avoid such confusion.

Acknowledgments

We would like to thank the Editor, the Associate Editor, and three reviewers for their thoughtful and useful comments. We would also like to thank Trevor Hastie, Saharon Rosset, Rob Tibshirani and Hui Zou for helpful discussions. Li and Zhu are partially supported by grant DMS-0505432 from the National Science Foundation. Liu is partially supported by grant DMS-0606577 from the National Science Foundation.

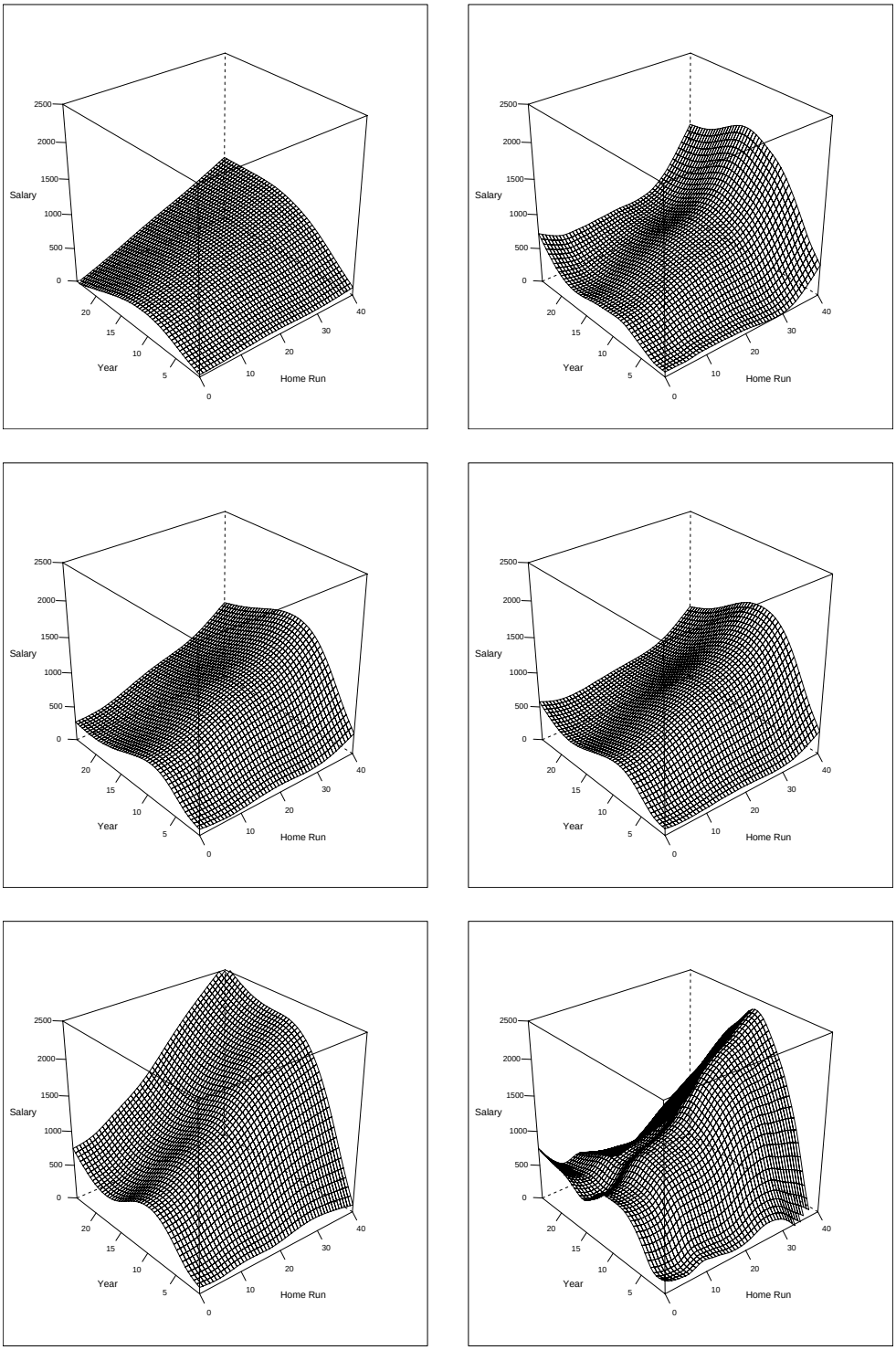

Figure 4: Baseball data. The left column contains the SIC results, and the right column contains the GACV results. The first row corresponds to the 25% quantile surfaces, the middle row corresponds to the 50% quantile surfaces, and the last row corresponds to the 75% quantile surfaces.

Appendix: Proofs

Proof of Lemma 1

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

$$\frac{1}{\lambda}\left(\beta_{0,\lambda} + \sum_{i\in\mathcal{E}} \theta_i K(\boldsymbol{x}_k, \boldsymbol{x}_i) - (1-\tau)\sum_{i\in\mathcal{L}} K(\boldsymbol{x}_k, \boldsymbol{x}_i) + \tau \sum_{i\in\mathcal{R}} K(\boldsymbol{x}_k, \boldsymbol{x}_i)\right) = y_k, \forall k \in \mathcal{E} \quad (41)$$

$$\sum_{i\in\mathcal{E}} \theta_i - (1-\tau)n_{\mathcal{L}} + \tau n_{\mathcal{R}} = 0 \quad (42)$$

These can be re-expressed as

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

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

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

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

where

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

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

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

1. $\theta_k = -(1-\tau),\ \exists k \in \mathcal{E}$
2. $\theta_k = \tau,\ \exists k \in \mathcal{E}$

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

For any fixed $\lambda, \mathcal{R}, \mathcal{L}$ and $\mathcal{E}$, each of the above conditions defines a hyperplane of $\boldsymbol{y}$ in $\mathbb{R}^n$. Taking into account all possible combinations of $\mathcal{R}, \mathcal{L}$ and $\mathcal{E}$, the set of $\boldsymbol{y}$ 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 $\boldsymbol{y}_0 \in \mathbb{R}^n$, we wish to show that if a sequence $\boldsymbol{y}_m$ converges to y_0, then $\boldsymbol{\theta}(\boldsymbol{y}_m)$ converges to $\boldsymbol{\theta}(\boldsymbol{y}_0)$.

Since $\boldsymbol{\theta}(\boldsymbol{y}_m)$ are bounded, it is equivalent to show that for every converging subsequence, say $\boldsymbol{\theta}(\boldsymbol{y}_{m_k})$, the subsequence converges to $\boldsymbol{\theta}(\boldsymbol{y}_0)$. Suppose $\boldsymbol{\theta}(\boldsymbol{y}_{m_k})$ converges to $\boldsymbol{\theta}_\infty$, we will show $\boldsymbol{\theta}_\infty = \boldsymbol{\theta}(\boldsymbol{y}_0)$.

Denote (6) as $g(\boldsymbol{\theta}(\boldsymbol{y}), \boldsymbol{y})$, and let

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

Then we have

$$\begin{aligned} g(\boldsymbol{\theta}(\boldsymbol{y}_0), \boldsymbol{y}_0) &= g(\boldsymbol{\theta}(\boldsymbol{y}_0), \boldsymbol{y}_{m_k}) + \Delta g(\boldsymbol{\theta}(\boldsymbol{y}_0), \boldsymbol{y}_0, \boldsymbol{y}_{m_k}) \\ &\geq g(\boldsymbol{\theta}(\boldsymbol{y}_{m_k}), \boldsymbol{y}_{m_k}) + \Delta g(\boldsymbol{\theta}(\boldsymbol{y}_0), \boldsymbol{y}_0, \boldsymbol{y}_{m_k}) \\ &= g(\boldsymbol{\theta}(\boldsymbol{y}_{m_k}), \boldsymbol{y}_0) + \Delta g(\boldsymbol{\theta}(\boldsymbol{y}_{m_k}), \boldsymbol{y}_{m_k}, \boldsymbol{y}_0) + \Delta g(\boldsymbol{\theta}(\boldsymbol{y}_0), \boldsymbol{y}_0, \boldsymbol{y}_{m_k}). \end{aligned} \quad (43)$$

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

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

where $c > 0$ is a constant. Furthermore, using $\boldsymbol{y}_{m_k} \to \boldsymbol{y}_0$ and $\boldsymbol{\theta}(\boldsymbol{y}_{m_k}) \to \boldsymbol{\theta}_\infty$, we reduce (43) to

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

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

Proof of Lemma 3

For any fixed $\lambda > 0$ and any fixed $\boldsymbol{y}_0 \in \mathbb{R}^n \backslash \mathcal{N}_\lambda$, since $\mathbb{R}^n \backslash \mathcal{N}_\lambda$ is an open set, we can always find a small enough $\epsilon > 0$, such that $\text{Ball}(\boldsymbol{y}_0, \epsilon) \subset \mathbb{R}^n \backslash \mathcal{N}_\lambda$. So λ is not an event point for any $\boldsymbol{y} \in \text{Ball}(\boldsymbol{y}_0, \epsilon)$.

We claim that if ϵ is small enough, the sets $\mathcal{R}, \mathcal{L}$ and $\mathcal{E}$ stay the same for all $\boldsymbol{y} \in \text{Ball}(\boldsymbol{y}_0, \epsilon)$.

Consider $\boldsymbol{y}$ and $\boldsymbol{y}_0$. Let $\mathcal{R}_{\boldsymbol{y}}, \mathcal{L}_{\boldsymbol{y}}, \mathcal{E}_{\boldsymbol{y}}, \mathcal{R}_0, \mathcal{L}_0, \mathcal{E}_0$ denote the corresponding sets, and $\boldsymbol{\theta}^{\boldsymbol{y}}, f^{\boldsymbol{y}}, \boldsymbol{\theta}^0, f^0$ denote the corresponding fits.

For any $i \in \mathcal{E}_0$, since λ is not an event point, we have $-(1-\tau) < \theta_i^0 < \tau$. Therefore, by continuity, we also have $-(1-\tau) < \theta_i^{\boldsymbol{y}} < \tau,\ i \in \mathcal{E}_0$ for $\boldsymbol{y}$ close enough to $\boldsymbol{y}_0$; or equivalently, $\mathcal{E}_0 \subseteq \mathcal{E}_{\boldsymbol{y}}, \forall \boldsymbol{y} \in \text{Ball}(\boldsymbol{y}_0, \epsilon)$ for small enough ϵ.

Similarly, for any $i \in \mathcal{R}_0$, since $y_i^0 - f^0(\boldsymbol{x}_i) > 0$, again by continuity, we have $y_i - f^{\boldsymbol{y}}(\boldsymbol{x}_i) > 0$ for $\boldsymbol{y}$ close enough to $\boldsymbol{y}_0$; or equivalently, $\mathcal{R}_0 \subseteq \mathcal{R}_{\boldsymbol{y}}, \forall \boldsymbol{y} \in \text{Ball}(\boldsymbol{y}_0, \epsilon)$ for small enough ϵ. The same applies to $\mathcal{L}_0$ and $\mathcal{L}_{\boldsymbol{y}}$ as well.

Overall, we then must have $\mathcal{E}_0 = \mathcal{E}_{\boldsymbol{y}}$, $\mathcal{R}_0 = \mathcal{R}_{\boldsymbol{y}}$ and $\mathcal{L}_0 = \mathcal{L}_{\boldsymbol{y}}$ for all $\boldsymbol{y} \in \text{Ball}(\boldsymbol{y}_0, \epsilon)$ when ϵ is small enough. □

Proof of Theorem 1

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

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

Furthermore, from (41) and (42), 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 $\boldsymbol{y}_{\mathcal{E}}$. Hence

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

Overall, we have

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

□

Proof of Lemma 4

The desired result can be proved directly by expressing $\rho_\tau(y - f(\boldsymbol{x})) - \rho_\tau(y - f^*(\boldsymbol{x}))$ 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 $z_i = (\boldsymbol{x}_i, y_i)$, $Z_i = (\boldsymbol{X}_i, Y_i)$, $\ell_\tau(f, z_i) = \rho_\tau(y_i - f(\boldsymbol{x}_i))$, and $\tilde{\ell}_\tau(f, z_i) = \ell_\tau(f, z_i) + \lambda J(f)$. We let

$$\begin{aligned} A_{ij} &= \{f \in \mathcal{F} : 2^{i-1}\delta_n^2 \le e_\tau(f, f^*) < 2^i\delta_n^2, 2^{j-1}J_0 \le J(f) < 2^j J_0\}, \quad i, j = 1, 2, \ldots \\ A_{i0} &= \{f \in \mathcal{F} : 2^{i-1}\delta_n^2 \le e_\tau(f, f^*) < 2^i\delta_n^2, J(f) < J_0\}, \quad i = 1, 2, \ldots \end{aligned}$$

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

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

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

First, we notice that

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

Hence

$$\Pr(e_\tau(\hat{f}, f^*) \ge \delta_n^2) \le \Pr^* \left(\sup_{\{f \in \mathcal{F} : e_\tau(f, f^*) \ge \delta_n^2\}} n^{-1}\sum_{i=1}^n (\tilde{\ell}_\tau(f^*, Z_i) - \tilde{\ell}_\tau(f, Z_i)) \ge 0 \right), \qquad (44)$$

where $\Pr^*$ denotes the outer probability measure.

We denote the probability on the right-hand side of (44) as I. To bound I, it is sufficient to bound $\Pr^*\left(\sup_{\{f \in A_{ij}\}} n^{-1}\sum_{i=1}^n (\tilde{\ell}_\tau(f^*, Z_i) - \tilde{\ell}_\tau(f, Z_i)) \ge 0\right)$ for each i, j. To this end, we need some inequalities regarding to 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

$$\inf_{A_{ij}} \mathrm{E}\left[\tilde{\ell}_\tau(f, Z_i) - \tilde{\ell}_\tau(f^*, Z_i)\right] \geq 2^{i-1}\delta_n^2 + \lambda(2^{j-1} - 1)J_0 \stackrel{\text{def}}{=} M(i, j), \tag{45}$$

and

$$\inf_{A_{i0}} \mathrm{E}\left[\tilde{\ell}_\tau(f, Z_i) - \tilde{\ell}_\tau(f^*, Z_i)\right] \geq 2^{i-2}\delta_n^2 \stackrel{\text{def}}{=} M(i, 0), \tag{46}$$

where $i, j \geq 1$. Note that (46) follows from the assumption $\lambda J_0 \leq \delta_n^2/2$ and the fact that $2^i - 1 \leq 2^{i-1}$.

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

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

Furthermore, using Lemma 4, we get

$$\mathrm{E}\left|\ell_\tau(f, Z) - \ell_\tau(f^*, Z)\right| \leq \mathrm{E}\left[|g_\tau(\boldsymbol{X}, Y)| + |h_\tau(\boldsymbol{X}, Y)|\right]. \tag{47}$$

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

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

Thus, we have

$$\mathrm{E}\left[\ell_\tau(f, Z) - \ell_\tau(f^*, Z)\right]^2 \leq T(c_1^{-1}\tau(1-\tau) + 1)(e_\tau(f, f^*))^\alpha. \tag{48}$$

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

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

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

Now we are ready to bound I. Using (45) and (46), we get

$$\begin{aligned} I &\leq \sum_{i\geq 1, j\geq 0} \Pr^* \left(\sup_{A_{ij}} \mathrm{E}_n(\tilde{\ell}_\tau(f^*, Z_i) - \tilde{\ell}_\tau(f, Z_i)) \geq 0 \right) \\ &\leq \sum_{i\geq 1, j\geq 0} \Pr^* \left(\sup_{A_{ij}} \mathrm{E}_n(\ell_\tau(f^*, Z_i) - \ell_\tau(f, Z_i)) \geq M(i,j) \right) \\ &= I_1 + I_2, \end{aligned}$$

where

$$\begin{aligned} I_1 &= \sum_{i,j\geq 1} \Pr^* \left(\sup_{A_{ij}} \mathrm{E}_n(\ell_\tau(f^*, Z_i) - \ell_\tau(f, Z_i)) \geq M(i,j) \right) \\ I_2 &= \sum_{i\geq 1} \Pr^* \left(\sup_{A_{i0}} \mathrm{E}_n(\ell_\tau(f^*, Z_i) - \ell_\tau(f, Z_i)) \geq M(i,0) \right) \end{aligned}$$

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

To compute the metric entropy in (4.7) of Shen & 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

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

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

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

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

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 non-increasing in i and $M(i,j)$, we have

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

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

$$\begin{aligned}
I_1 &\leq \sum_{j=1}^{\infty}\sum_{i=1}^{\infty} 3\exp\left(-\frac{(1-\epsilon)nM(i,j)^2}{2(4v(i,j)^2 + M(i,j)T/3)}\right) \\
&\leq \sum_{j=1}^{\infty}\sum_{i=1}^{\infty} 3\exp\left(-c_5 nM(i,j)^{2-\alpha}\right) \\
&\leq \sum_{j=1}^{\infty}\sum_{i=1}^{\infty} 3\exp\left(-c_5 n[(2^{i-1}\delta_n^2)^{2-\alpha} + ((2^{j-1}-1)\lambda J_0)^{2-\alpha}]\right) \\
&\leq \sum_{j=1}^{\infty}\sum_{i=1}^{\infty} 3\exp\left(-c_5 n[i(\lambda J_0)^{2-\alpha} + (j-1)(\lambda J_0)^{2-\alpha}]\right) \\
&\leq 3\exp\left(-c_5 n(\lambda J_0)^{2-\alpha}\right)/[1-\exp\left(-c_5 n(\lambda J_0)^{2-\alpha}\right)]^2,
\end{aligned}$$

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

$$I \leq 6\exp(-c_5 n(\lambda J_0)^{2-\alpha})/[1-\exp(-c_5 n(\lambda J_0)^{2-\alpha})]^2,$$

which implies that $I^{1/2} \leq (5/2 + I^{1/2})\exp(-c_5 n(\lambda J_0)^{2-\alpha})$. Since $I \leq I^{1/2} \leq 1$, we finally have

$$I \leq 3.5\exp\left(-c_5 n(\lambda J_0)^{2-\alpha}\right),$$

which is the desired result. □

Proof of Corollary 1

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

$$\Pr\left(e(\hat{f}, f^*) \geq G\delta_n^2\right) \leq 3.5\exp\left(-c_5 nG(\lambda J_0)^{2-\alpha}\right)$$

for any $G \geq 1$.

To this end, we only need to slightly modify the proof of Theorem 2. We re-define

$$\begin{aligned} A_{ij} &= \{f \in \mathcal{F} : 2^{i-1}\delta_n^2 G \leq e_\tau(f, f^*) < 2^i \delta_n^2 G, 2^{j-1} J_0 G \leq J(f) < 2^j J_0 G\}, \quad i, j = 1, 2, \ldots \\ A_{i0} &= \{f \in \mathcal{F} : 2^{i-1}\delta_n^2 G \leq e_\tau(f, f^*) < 2^i \delta_n^2 G, J(f) < J_0 G\}, \quad i = 1, 2, \ldots \end{aligned}$$

Using these new definitions, an analogous proof to that of Theorem 2 can be obtained with $M(i, j) = 2^{i-1}\delta_n^2 G + \lambda(2^{j-1} - 1)J_0 G$ and the desired result then follows. □

References

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

Bosch, R., Ye, Y. & 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', *Annals of Statistics* **2**, 760–777.

Cole, T. & 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.

Efron, B. (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*, Springer, New York.

Hastie, T., Rosset, S., Tibshirani, R. & 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', *American Statistician* **51**(2), 186–192.

He, X., Ng, P. & Portnoy, S. (1998), 'Bivariate quantile smoothing splines', *Journal of the Royal Statistical Society Series B* **60**(3), 537–550.

Heagerty, P. & 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. & 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. & Wahba, G. (1971), 'Some results on Tchebycheffian spline functions', *Journal of Mathematical Analysis and Applications* **33**, 82–95.

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

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

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

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

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

Machado, J. (1993), 'Robust model selection and M-estimation', *Econometric Theory* **9**, 478–493.

Meyer, M. & Woodroofe, M. (2000), 'On the degrees of freedom in shape-restricted regression', *Annals of Statistics* **28**, 1083–1104.

Nychka, D., Gray, G., Haaland, P., Martin, D. & O'Connell, M. (1995), 'A nonparametric regression approach to syringe grading for quality improvement', *Journal of the American Statistical Association* **90**(432), 1171–1178.

Oh, H., Nychka, D., Brown, T. & Charbonneau, P. (2004), 'Period analysis of variable stars by robust smoothing', *Journal of the Royal Statistical Society Series A* **53**, 15–30.

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

Shen, X. & Wong, W. (1994), 'Convergence rate of sieve estimates', *Annals of Statistics* **22**, 580–615.

Stein, C. (1981), 'Estimation of the mean of a multivariate normal distribution', *Annals of Statistics* **9**(6), 1135–1151.

Stone, C. (1977), 'Consistent nonparametric regression, with discussion', *Annals of Statistics* **5**, 595–645.

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

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**(441), 120–131.

Yu, K. & Jones, M. (1998), 'Local linear regression quantile estimation', *Journal of the American Statistical Association* **93**, 228–238.

Yu, K., Lu, Z. & 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**(3), 813–829.

Zhou, D.-X. (2002), 'The covering number in learning theory', *Journal of Complexity* **18**, 739–767.

Zou, H., Hastie, T. & Tibshirani, R. (2005), On the degrees of freedom of the lasso, Technical report, Department of Statistics, Stanford University.

Table 1: The upper part summarizes the elapsed CPU times (in seconds), and the lower part summarizes characteristics of the path. 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.

Elapsed CPU Times						
	Path			Interior Point		
n	$\tau = 0.1$	$\tau = 0.3$	$\tau = 0.5$	$\tau = 0.1$	$\tau = 0.3$	$\tau = 0.5$
50	0.09 (0.03)	0.13 (0.01)	0.14 (0.02)	2.70 (0.32)	4.14 (0.43)	4.74 (0.52)
100	0.20 (0.02)	0.33 (0.03)	0.38 (0.04)	12.19 (1.58)	21.12 (1.91)	23.67 (1.90)
200	0.52 (0.09)	0.89 (0.11)	0.97 (0.10)	104.00 (10.03)	179.85 (17.54)	196.29 (16.17)
400	3.26 (0.21)	4.18 (0.35)	4.58 (0.38)	1986.23 (56.26)	2561.75 (68.31)	2706.64 (75.07)
Path Characteristics						
	Average $\|\mathcal{E}\|$			Average # Steps		
n	$\tau = 0.1$	$\tau = 0.3$	$\tau = 0.5$	$\tau = 0.1$	$\tau = 0.3$	$\tau = 0.5$
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)

Table 2: Average elapsed CPU time (in seconds) of the interior point algorithm for a single value of λ_ℓ

n	$\tau = 0.1$	$\tau = 0.3$	$\tau = 0.5$
50	0.054 (0.004)	0.053 (0.002)	0.052 (0.001)
100	0.129 (0.003)	0.123 (0.003)	0.119 (0.003)
200	0.579 (0.017)	0.524 (0.076)	0.514 (0.020)
400	3.692 (0.378)	3.167 (0.223)	3.073 (0.274)

Table 3: Simulation example, $\tau = 10\%$. Prediction errors of the true conditional quantile functions are 0.173 (Normal), 0.258 (DE), 0.285 (T3) and 0.311 (Mixture), respectively.

Prediction Error			
	SIC	GACV	Gold Standard
Normal	0.216 (0.016)	0.355 (0.165)	0.208 (0.011)
DE	0.324 (0.065)	0.532 (0.212)	0.296 (0.010)
T3	0.352 (0.073)	0.567 (0.226)	0.323 (0.011)
Mixture	0.382 (0.059)	0.590 (0.220)	0.350 (0.012)
Mean Absolute Deviation			
	SIC	GACV	Gold Standard
Normal	0.505 (0.075)	0.857 (0.379)	0.478 (0.072)
DE	0.761 (0.133)	1.235 (0.437)	0.672 (0.114)
T3	0.842 (0.212)	1.494 (0.496)	0.724 (0.133)
Mixture	0.904 (0.337)	1.491 (0.610)	0.697 (0.165)
Degrees of Freedom			
	SIC	GACV	Gold Standard
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)

Table 4: Simulation Example, $\tau = 30\%$. Prediction errors of the true conditional quantile functions are 0.346 (Normal), 0.448 (DE), 0.493 (T3) and 0.515 (Mixture), respectively.

Prediction Error			
	SIC	GACV	Gold Standard
Normal	0.393 (0.016)	0.470 (0.111)	0.384 (0.011)
DE	0.503 (0.019)	0.603 (0.167)	0.490 (0.012)
T3	0.551 (0.018)	0.644 (0.195)	0.538 (0.011)
Mixture	0.570 (0.022)	0.606 (0.093)	0.557 (0.013)
Mean Absolute Deviation			
	SIC	GACV	Gold Standard
Normal	0.398 (0.058)	0.683 (0.347)	0.368 (0.052)
DE	0.462 (0.071)	0.737 (0.361)	0.423 (0.061)
T3	0.493 (0.088)	0.779 (0.453)	0.440 (0.070)
Mixture	0.478 (0.082)	0.602 (0.414)	0.405 (0.051)
Degrees of Freedom			
	SIC	GACV	Gold Standard
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)

Table 5: Simulation example, $\tau = 50\%$. Prediction errors of the true conditional quantile functions are 0.398 (Normal), 0.503 (DE), 0.553 (T3) and 0.565 (Mixture), respectively.

Prediction Error			
	SIC	GACV	Gold Standard
Normal	0.448 (0.017)	0.481 (0.077)	0.438 (0.010)
DE	0.561 (0.021)	0.591 (0.111)	0.545 (0.012)
T3	0.615 (0.023)	0.655 (0.136)	0.600 (0.013)
Mixture	0.619 (0.022)	0.626 (0.050)	0.606 (0.010)
Mean Absolute Deviation			
	SIC	GACV	Gold Standard
Normal	0.375 (0.062)	0.538 (0.237)	0.348 (0.048)
DE	0.397 (0.073)	0.463 (0.238)	0.355 (0.060)
T3	0.429 (0.077)	0.482 (0.214)	0.384 (0.054)
Mixture	0.425 (0.075)	0.490 (0.247)	0.384 (0.051)
Degrees of Freedom			
	SIC	GACV	Gold Standard
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)