ARTICLE IN PRESS

Available online at www.sciencedirect.com

SCIENCE DIRECT®

ELSEVIER

Computational Statistics & Data Analysis ▮▮▮ (▮▮▮▮) ▮▮▮–▮▮▮

COMPUTATIONAL STATISTICS & DATA ANALYSIS

www.elsevier.com/locate/csda

GACV for quantile smoothing splines

Ming Yuan*

School of Industrial and Systems Engineering Georgia Institute of Technology, 755 Ferst Drive, Atlanta, GA 30332-0205, USA

Received 12 September 2003; received in revised form 19 October 2004; accepted 21 October 2004

Abstract

Quantile smoothing splines provide nonparametric estimation of conditional quantile functions. Like other nonparametric smoothing techniques, the choice of smoothing parameters considerably affects the performance of quantile smoothing splines. The robust cross-validation (RCV) has been commonly used as a tuning criterion in practice. To explain its success, Oh et al. (J. Roy. Statist. Soc. Ser. A, in press) argued that the RCV curve, as a function of smoothing parameters in quantile smoothing splines, differs from the mean squared error (MSE) curve only by a constant. In this article, we consider an alternative loss function, the generalized comparative Kullback–Leibler distance (GCKL) for the quantile smoothing spline. We argue that RCV is an estimator of GCKL. A disadvantage of RCV is its computational intensity. To reduce the associated computational cost, Nychka et al. (J. Amer. Statist. Assoc. 90 (432) (1995) 1171) has previously proposed an approximation to RCV, namely ACV. However, we find in our simulations that the ACV-based tuning method will often fail in practice. We first reexamine the theoretical basis for ACV. This exercise enables us to explain the failure of ACV. Then we continue to propose a remedy, the generalized approximate cross-validation (GACV) as a computable proxy for the GCKL. Some preliminary simulations suggest that the GACV score is a good estimate of the GCKL score and that the GACV-based tuning technique compares favorably with both ACV and another commonly used criteria, Schwartz information criterion. A real dataset is examined to illustrate the empirical performance of the proposed method.

Keywords: Smoothing parameters; Quantile smoothing splines; Generalized approximate cross-validation; Generalized comparative Kullback–Leibler distance

* Corresponding author.
E-mail address: myuan@isye.gatech.edu (M. Yuan).

0167-9473/$ - see front matter © 2004 Published by Elsevier B.V.
doi:10.1016/j.csda.2004.10.008

1. Introduction

The focus of most nonparametric regression methodologies centers around the conditional mean. However, conditional quantile functions may provide more complete information and often uncover additional features between the responses and the corresponding covariates. For practical examples, the readers are referred to Koenker and Hallock (2001) among many other recent review articles.

In a typical regression setting, we are interested in the functional dependence between a univariate response y and its p-variate covariate $\mathbf{x} = (x_1, \ldots, x_p)$. To uncover the relationship, n independent copies of $(\mathbf{x}, y)$, $(\mathbf{x}_i, y_i)$, $i = 1, \ldots, n$ are observed. Now suppose the functional dependence which we are interested in is the αth quantile of the conditional distribution of y given $\mathbf{x}$, $f(\mathbf{x})$. A popular nonparametric estimator of f can be obtained through regularization.

Following the pioneering article of Koenker and Bassett (1978), define the check function as

$$\rho_\alpha(u) = (\alpha I(u > 0) + (1 - \alpha) I(u < 0))|u|, \tag{1.1}$$

where $I(\cdot)$ is an indicator function. This check function highlights the basic difference between the conditional mean and the conditional quantile function. The mean minimizes the expected squared loss; whereas, the αth quantile minimizes the expectation of the weighted absolute loss defined by (1.1). A quantile spline estimator of f can then be given as the minimizer of

$$\frac{1}{n} \sum_{i=1}^{n} \rho_\alpha (y_i - f(\mathbf{x}_i)) + \lambda J(f) \tag{1.2}$$

over a reproducing kernel Hilbert space $\mathscr{H}$, where J is a quadratic functional defined over $\mathscr{H}$. The first term of (1.2) measures the fidelity to the observations. The second term shrinks the solution toward the null space of J. Different variants of the quantile splines have been studied by several authors recently.

For univariate covariates, following the traditional smoothing splines, Nychka et al. (1995) chose $J(f) = \int |f''|^2$. A pseudo-data algorithm was also proposed in the same paper to carry out the minimization efficiently. Alternatively Koenker et al. (1994) set $J(f) = \left(\int |f''|^p\right)^{1/p}$. It has been shown that solutions of (1.2) under this setting are linear splines when $p = 1$. This enables us to express (1.2) as a linear programming problem. Both results have been extended to two-dimensional covariates cases, namely elastic and plastic splines, as termed by Koenker and Mizera (2002).

Like other nonparametric smoothing methods, smoothing parameter λ plays a crucial role on determining the trade-off between the fidelity to the data and the penalty. When λ is too large, there is too much penalty placed on the estimate. As a consequence, the data is oversmoothed. On the other hand, when λ is too small, we tend to interpolate the data more and this will lead to undersmoothing. The main goal here is to pick a λ such that the distance between the resulting estimate and the true function is minimized. The major difficulty is that we do not observe the true function. Therefore we cannot directly evaluate the distance. Instead, we should rely on some other proxies.

One standard proxy is the robust cross-validation (RCV). The problem with RCV is its high computational cost. To reduce the computational burden incurred by RCV, approximate cross-validation (ACV) was proposed by Nychka et al. (1995) as an approximation to RCV. However, as our simulation in this paper reveals, ACV fails frequently as a tuning method. The reason is that it seriously overestimates the loss. To explain this failure, we first reexamine the theoretical motivations of ACV based on a linearization argument first introduced by Xiang and Wahba (1996). Using the insight gained from this exercise, we propose GACV as a remedy to ACV. Simulation results suggest that GACV has a fairly high statistical efficiency and compares favorably with another commonly used tuning method, namely SIC.

In this paper, we estimate the quantile smoothing spline by the iteratively reweighting scheme introduced by Nychka et al. (1995). The approach will be formulated in the next section. In Section 3, we will introduce GCKL. We argue that RCV and ACV are approximate unbiased estimates of GCKL. This notion leads to the definition of GACV, which we suggest to be used to tune smoothing parameters. Section 4 presents simulations to evaluate the performance of GACV as a smoothing parameter tuning technique. In the final section a real dataset is analyzed to illustrate the method.

2. Quantile smoothing splines

Quantile smoothing splines defined in (1.2) can be viewed analogously to the traditional smoothing spline for estimating the conditional mean

$$\frac{1}{n}\sum_{i=1}^{n}(y_i - f(\mathbf{x}_i))^2 + \lambda J(f). \tag{2.1}$$

Denote $f(\mathbf{x}_i)$ by f_i and let K be the $n \times n$ semi-positive definite matrix associated with the penalty J such that $J(f) = \mathbf{f}'K\mathbf{f}$, where $\mathbf{f} = (f_1, \ldots, f_n)'$. Using these notation, we can re-express (2.1) as

$$\frac{1}{n}\sum_{i=1}^{n}(y_i - f_i)^2 + \lambda \mathbf{f}'K\mathbf{f}. \tag{2.2}$$

The solution of (2.2) satisfies

$$\frac{1}{n}(y_i - f_i) + \lambda(K\mathbf{f})_i = 0, \quad \forall i. \tag{2.3}$$

Now if we pretend that ρ_α is differentiable for a moment, then the solution of (1.2) will satisfy similar equations as (2.3)

$$\frac{1}{n}(y_i - f_i)\frac{\rho'_\alpha(y_i - f_i)}{2(y_i - f_i)} + \lambda(K\mathbf{f})_i = 0, \quad \forall i. \tag{2.4}$$

Comparing (2.4) and (2.3), Nychka et al. (1995) proposed to minimize (1.2) by iteratively solving a weighted smoothing spline problem with weights $\{\rho'_\alpha(y_i - f_i)/2(y_i - f_i)\}$.

To get around the nondifferentiability of ρ_α at 0, they suggested to approximate ρ_α by a differentiable function $\rho_{\alpha,\delta}$, which differs from ρ_α only in the region $(-\delta, \delta)$, where

$$\rho_{\alpha,\delta}(u) = (\alpha I(u > 0) + (1-\alpha) I(u < 0))\, u^2/\delta. \tag{2.5}$$

By setting δ small enough, we can get a good approximate solution to (1.2).

To sum up the procedure, an approximate solution to (1.2) for a fixed smoothing parameter can be computed in the following way:

(a) Set an approximation threshold δ.
(b) Initialize a solution $\mathbf{f}^{(0)}$.
(c) Given the current estimate $\mathbf{f}^{(k)}$, fit a weighted smoothing spline with response $\{y_i\}$, covariates $\{\mathbf{x}_i\}$ and weights $\left\{\rho'_{\alpha,\delta}\left(y_i - f_i^{(k)}\right) \Big/ 2\left(y_i - f_i^{(k)}\right)\right\}$. Denote the solution by $\mathbf{f}^{(k+1)}$.
(d) Iterate step (c) until sequence $\left\{\mathbf{f}^{(k)}\right\}$ meets certain convergence criteria.

Let $\widehat{f}_{\alpha,\lambda}$ be the αth quantile spline with smoothing parameter λ and $\widehat{f}_{\alpha,\lambda}^{[-i]}$ be defined similarly as $\widehat{f}_{\alpha,\lambda}$ but with the ith observation omitted. For simplicity, we will abbreviate the subscripts α and/or λ when no notational confusion occurs.

To choose an appropriate smoothing parameter λ, a commonly used technique is to minimize the cross-validation score, which is defined as

$$RCV(\lambda) = \frac{1}{n}\sum_{i=1}^{n} \rho_\alpha\left(y_i - \widehat{f}_\lambda^{[-i]}(\mathbf{x}_i)\right). \tag{2.6}$$

In principle (2.6) could be evaluated and used as a tuning criterion. But the computational cost associated with (2.6) is formidable since for each candidate smoothing parameter λ, $(n+1)$ quantile splines $\widehat{f}_{\alpha,\lambda}, \widehat{f}_{\alpha,\lambda}^{[-1]}, \ldots, \widehat{f}_{\alpha,\lambda}^{[-n]}$ should be evaluated.

To reduce the computational burden of (2.6), Nychka et al. (1995) suggested the following approximation to RCV:

$$ACV(\lambda) = \frac{1}{n}\sum_{i=1}^{n} \rho_\alpha\left(\frac{y_i - \widehat{f}_\lambda(\mathbf{x}_i)}{1 - h_{ii}}\right), \tag{2.7}$$

where $h_{ii} = \partial \widehat{f}_{\alpha,\lambda}(\mathbf{x}_i)/\partial y_i$.

The basic rationale behind the above approximation is a similar identity for the smoothing splines (Craven and Wahba, 1979). RCV enjoys a great success when applied in practice. Oh et al. (2002) argued that it is because $RCV(\lambda) - MSE(\lambda)$ is approximately a constant not depending on λ, where

$$MSE(\lambda) = \frac{1}{n}\sum_{i=1}^{n}\left(f(\mathbf{x}_i) - \widehat{f}(\mathbf{x}_i)\right)^2.$$

In the next section, we will take a different approach to explain its success. Then we will check the theoretical basis for approximating RCV by ACV. The discussion leads us to a

modification of ACV, which is shown to be preferable to ACV both from intuition and from our simulations of Section 4.

3. Generalized approximate cross-validation

Instead of minimizing $MSE(\lambda)$, we aim at a smoothing parameter λ which minimizes the risk

$$GCKL(\lambda) = \frac{1}{n}\sum_{i=1}^{n} E_z \rho_\alpha \left(z_i - \widehat{f}_\lambda(\mathbf{x}_i)\right), \tag{3.1}$$

where z_i is an independent copy of y_i and (3.1) is also known as GCKL. However, this quantity is not computable since the true distribution of y_i is unknown. To tackle this problem, we can minimize an exact or approximately unbiased estimate of GCKL instead. In general, it is not clear whether there exists an exact unbiased estimate of GCKL. Thus, approximately unbiased estimates are commonly used. There are many different options to derive approximately unbiased estimates of GCKL. The most popular choice is the cross-validation score.

Comparing (3.1) with (2.6), we can see that for large enough sample sizes,

$$E_z(RCV(\lambda)) \approx GCKL(\lambda). \tag{3.2}$$

In other words, RCV is an approximate unbiased estimate of GCKL. Although RCV is quite hard to compute, it is very intuitive and provides an accurate estimate to GCKL. To preserve these two virtues, we can start with RCV to search for a more computable estimate of GCKL.

A first-order Taylor expansion gives

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n} \rho_\alpha \left(y_i - \widehat{f}^{[-i]}(\mathbf{x}_i)\right) \\
&\quad = \frac{1}{n}\sum_{i=1}^{n} \rho_\alpha \left(y_i - \widehat{f}(\mathbf{x}_i)\right) + \frac{1}{n}\sum_{i=1}^{n} \left[\rho_\alpha \left(y_i - \widehat{f}^{[-i]}(\mathbf{x}_i)\right) - \rho_\alpha \left(y_i - \widehat{f}(\mathbf{x}_i)\right)\right] \\
&\quad \approx \frac{1}{n}\sum_{i=1}^{n} \rho_\alpha \left(y_i - \widehat{f}(\mathbf{x}_i)\right) + \frac{1}{n}\sum_{i=1}^{n} \rho'_{\alpha,\delta} \left(y_i - \widehat{f}(\mathbf{x}_i)\right) \\
&\qquad \times \left(\widehat{f}(\mathbf{x}_i) - \widehat{f}^{[-i]}(\mathbf{x}_i)\right).
\end{aligned} \tag{3.3}$$

Our strategy is to approximate the second term of right-hand side of (3.3) so that it does not rely on $\widehat{f}^{[-i]}$, $i = 1, \ldots, n$ but only on $\widehat{f}$. Before proceeding, we first need the following version of the leaving-out-one lemma.

Lemma 3.1. *Let $\tilde{f}^{[i]}_{\alpha,\lambda}$ be defined in the same way as $\widehat{f}_{\alpha,\lambda}$ except that the ith response y_i is replaced by $\widehat{f}^{[-i]}_{\alpha,\lambda}(\mathbf{x}_i)$. Then $\tilde{f}^{[i]}_{\alpha,\lambda} = \widehat{f}^{[-i]}_{\alpha,\lambda}$.*

Proof. Denote $\tilde{y}_j = y_j$ for all $j \neq i$ and $\tilde{y}_i = \widehat{f}_{\alpha,\lambda}^{[-i]}(\mathbf{x}_i)$. For any $f \in \mathscr{H}$,

$$\begin{aligned}
&\frac{1}{n}\sum_{j=1}^{n}\rho_\alpha\left(\tilde{y}_j - f\left(\mathbf{x}_j\right)\right) + \lambda J(f)\\
&\geqslant \frac{1}{n}\sum_{j=1,j\neq i}^{n}\rho_\alpha\left(y_j - f\left(\mathbf{x}_j\right)\right) + \lambda J(f)\\
&\geqslant \frac{1}{n}\sum_{j=1,j\neq i}^{n}\rho_\alpha\left(y_j - \widehat{f}^{[-i]}\left(\mathbf{x}_j\right)\right) + \lambda J\left(\widehat{f}^{[-i]}\right)\\
&= \frac{1}{n}\sum_{j=1}^{n}\rho_\alpha\left(\tilde{y}_j - \widehat{f}^{[-i]}\left(\mathbf{x}_j\right)\right) + \lambda J\left(\widehat{f}^{[-i]}\right)\\
&\geqslant \frac{1}{n}\sum_{j=1}^{n}\rho_\alpha\left(\tilde{y}_j - \widehat{f}\left(\mathbf{x}_j\right)\right) + \lambda J\left(\widehat{f}\right), \qquad (3.4)
\end{aligned}$$

where the first inequality holds by the nonnegativity of ρ_α and the last two inequalities hold according to the definitions of $\widehat{f}^{[-i]}$ and $\widehat{f}$, respectively. Thus, the proof is completed by replacing f with $\widehat{f}_\lambda$ in (3.4). □

The leaving-out-one lemma suggests that,

$$\widehat{f}(\mathbf{x}_i) - \widehat{f}^{[-i]}(\mathbf{x}_i) \approx \frac{\partial \widehat{f}(\mathbf{x}_i)}{\partial y_i}\left(y_i - \widehat{f}^{[-i]}(\mathbf{x}_i)\right). \qquad (3.5)$$

Thus,

$$\begin{aligned}
&\rho'_{\alpha,\delta}\left(y_i - \widehat{f}(\mathbf{x}_i)\right)\left(\widehat{f}(\mathbf{x}_i) - \widehat{f}^{[-i]}(\mathbf{x}_i)\right)\\
&\quad\approx \rho'_{\alpha,\delta}\left(y_i - \widehat{f}(\mathbf{x}_i)\right)\frac{\partial \widehat{f}(\mathbf{x}_i)}{\partial y_i}\left(y_i - \widehat{f}^{[-i]}(\mathbf{x}_i)\right)\\
&\quad= \rho'_{\alpha,\delta}\left(y_i - \widehat{f}(\mathbf{x}_i)\right)\frac{\partial \widehat{f}(\mathbf{x}_i)}{\partial y_i}\frac{y_i - \widehat{f}(\mathbf{x}_i)}{1 - \frac{\widehat{f}(\mathbf{x}_i) - \widehat{f}^{[-i]}(\mathbf{x}_i)}{y_i - \widehat{f}^{[-i]}(\mathbf{x}_i)}}\\
&\quad\approx \rho'_{\alpha,\delta}\left(y_i - \widehat{f}(\mathbf{x}_i)\right)\frac{\partial \widehat{f}(\mathbf{x}_i)}{\partial y_i}\frac{y_i - \widehat{f}(\mathbf{x}_i)}{1 - \partial \widehat{f}(\mathbf{x}_i)/\partial y_i}\\
&\quad\approx \rho_{\alpha,\delta}\left(y_i - \widehat{f}(\mathbf{x}_i)\right)\frac{\partial \widehat{f}(\mathbf{x}_i)/\partial y_i}{1 - \partial \widehat{f}(\mathbf{x}_i)/\partial y_i}\\
&\quad\approx \rho_{\alpha}\left(y_i - \widehat{f}(\mathbf{x}_i)\right)\frac{\partial \widehat{f}(\mathbf{x}_i)/\partial y_i}{1 - \partial \widehat{f}(\mathbf{x}_i)/\partial y_i}. \qquad (3.6)
\end{aligned}$$

Now, (3.3) can be approximated by

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n}\rho_\alpha\left(y_i-\widehat{f}^{[-i]}(\mathbf{x}_i)\right)\\
&\quad\approx\frac{1}{n}\sum_{i=1}^{n}\rho_\alpha\left(y_i-\widehat{f}(\mathbf{x}_i)\right)+\frac{1}{n}\sum_{i=1}^{n}\rho_\alpha\left(y_i-\widehat{f}(\mathbf{x}_i)\right)\frac{\partial\widehat{f}(\mathbf{x}_i)/\partial y_i}{1-\partial\widehat{f}(\mathbf{x}_i)/\partial y_i}\\
&\quad=\frac{1}{n}\sum_{i=1}^{n}\rho_\alpha\left(y_i-\widehat{f}(\mathbf{x}_i)\right)\frac{1}{1-\partial\widehat{f}(\mathbf{x}_i)/\partial y_i}.
\end{aligned}\tag{3.7}$$

The right-hand side is ACV proposed by Nychka et al. (1995). However, $ACV(\lambda)$ is not a good approximate unbiased estimate to GCKL. Although it is much more computable than RCV, it loses the accuracy of RCV. To see this, we simulated a dataset using the setting of Section 4.1 with sample size 200. We computed the median smoothing splines using the *qsreg* function of the *R* package *fields* for 80 different smoothing parameters whose logarithms are equally spaced on $[-32,-12]$. This specific interval for smoothing parameters is chosen based on empirical evidences. In general, more sophisticated optimization tools could be employed to automate this procedure instead of grid search. For the sake of brevity, we are not going to explore this possibility in this paper.

The top left panel of Fig. 1 gives the GCKL and ACV curves. We can see that the ACV seriously overestimate the loss unless λ is very small. To understand how this happens, let us go back to (3.5). The performance of ACV depends on how good this approximation is. More specifically, we would want the approximation error of (3.5) to be of a smaller order than the first-order term of Taylor expansion, i.e.

$$\widehat{f}(\mathbf{x}_i)-\widehat{f}^{[-i]}(\mathbf{x}_i)=\left(\frac{\partial\widehat{f}(\mathbf{x}_i)}{\partial y_i}\left(y_i-\widehat{f}^{[-i]}(\mathbf{x}_i)\right)\right)(1+o(1)).\tag{3.8}$$

The approximation error of (3.5) depends on the relative magnitude of the second-order term of the Taylor expansion to the first-order term. If $\partial\widehat{f}(\mathbf{x}_i)/\partial y_i$ is too close to 0, the second-order term, which has been omitted in approximation (3.5) may dominate the first-order approximation. Thus, (3.8) will be violated and the accuracy of (3.5) may not be guaranteed. To see the effect of this, we picked three different λ's. The kernel density estimates of $\left\{\partial\widehat{f}(\mathbf{x}_i)/\partial y_i\right\}$ for these three different smoothing parameters are provided in the remaining three panels of Fig. 1. From this figure, we can see that ACV approximates GCKL better if $\left\{\partial\widehat{f}(\mathbf{x}_i)/\partial y_i\right\}$ are more evenly spread between 0 and 1. Our experiences with other examples also supported this finding.

To alleviate this problem, we borrow a trick from the derivation of GCV in (Craven and Wahba, 1979). Replace $\partial\widehat{f}(\mathbf{x}_i)/\partial y_i$ by their average $tr(H)/n$ in (3.7), where H is the so-called hat matrix with the (i,j) entry $\partial\widehat{f}(\mathbf{x}_i)/\partial y_j$. This gives us

$$\frac{1}{n}\sum_{i=1}^{n}\rho_\alpha\left(y_i-\widehat{f}^{[-i]}(\mathbf{x}_i)\right)\approx\frac{\sum_{i=1}^{n}\rho_\alpha\left(y_i-\widehat{f}(\mathbf{x}_i)\right)}{n-tr(H)}.\tag{3.9}$$

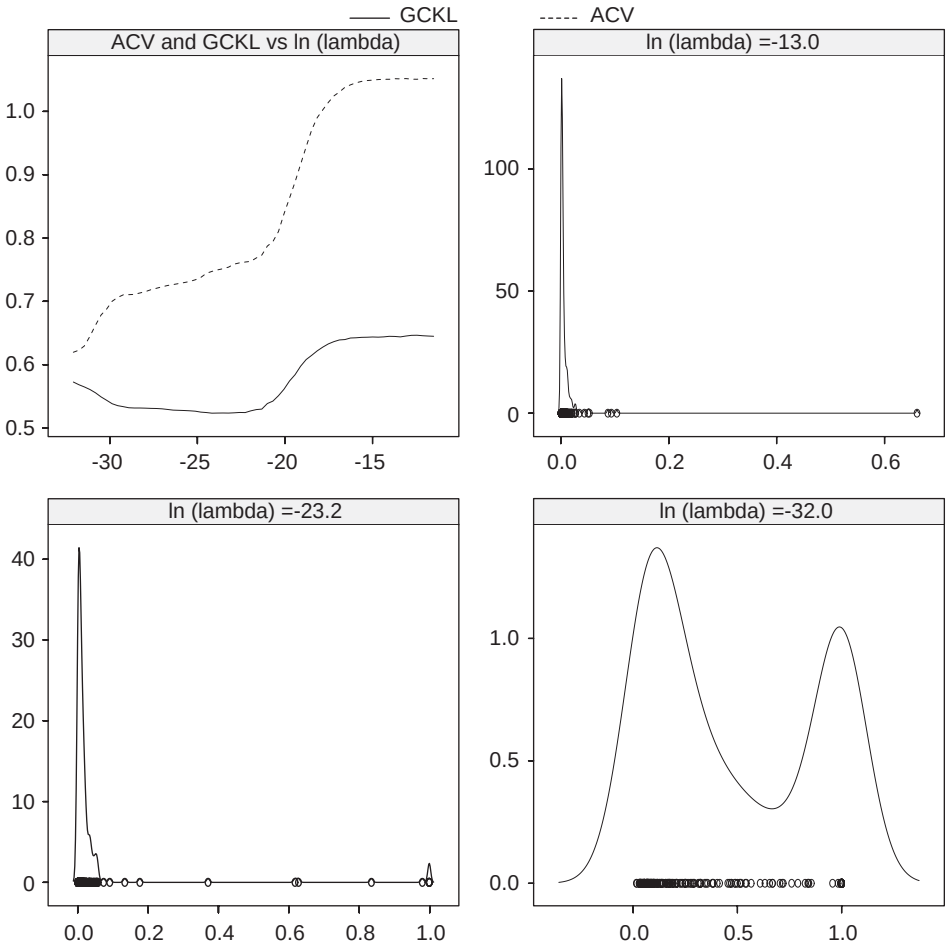

Fig. 1. *Failure of ACV*: This figure shows how ACV fails to be a good estimate of GCKL. The bottom left panel gives the GCKL and ACV curves. From this plot, we see that ACV overestimates GCKL. This problem becomes mitigated for small λ's. In the rest three panels, we give kernel density estimates of the diagonal elements of the "Hat" matrices for three different λ's. The circles of these density plots correspond to the diagonal elements. From this figure, we see that ACV approximates GCKL badly if the diagonal elements concentrate around 0.

We call the right-hand side of (3.9) GACV. It is interesting to notice that (3.9) shares a similar form with GCV for usual smoothing splines. The differences are that GACV has the sum of absolute deviations as the numerator while GCV has the sum of squared losses. Also, the denominator of GACV is the square root of the denominator of GCV.

In the next section, we will compare ACV and GACV through Monte Carlo simulations. We also include another popular method, SIC in our comparisons.

4. Monte Carlo simulations

4.1. Approximately unbiased risk estimate

In the last section, we argued that the main motivation of ACV and GACV is to estimate GCKL given by (3.1). In our first set of simulations, we will examine how good the approximations are. Consider the following model:

$$y_i = f(x_i) + \varepsilon_i, \quad i = 1, \ldots, 200, \tag{4.1}$$

where

$$f(x) = \sin(2\pi x) \tag{4.2}$$

and x_i's are independently sampled from $U(0, 1)$. The errors ε's are independent and identically distributed random variables from a double exponential distribution, whose density function is given by

$$\tfrac{1}{2}\exp(-|\varepsilon|), \quad \varepsilon \in (-\infty, +\infty).$$

The double exponential distribution allows a closed form evaluation of (3.1).

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n} E_z \rho_\alpha\left(z_i - \widehat{f}(x_i)\right) \\
&\quad \equiv \frac{1}{n}\sum_{i=1}^{n} E_e \rho_\alpha\left(e_i + \delta_i\right) \\
&\quad = \frac{1}{n}\sum_{i=1}^{n} E_e\left(\alpha - I\left(e_i + \delta_i < 0\right)\right)\left(e_i + \delta_i\right) \\
&\quad = \frac{1}{n}\sum_{i=1}^{n}\left[\alpha E_e\left(e_i + \delta_i\right) - E_e\left(e_i + \delta_i\right) I\left(e_i + \delta_i < 0\right)\right] \\
&\quad = \frac{1}{n}\sum_{i=1}^{n}\left[\alpha\delta_i - \left(-\frac{1}{2}\exp\left(-\left|\delta_i\right|\right) + \delta_i I\left(\delta_i < 0\right)\right)\right],
\end{aligned} \tag{4.3}$$

where e_i is an independent copy of ε_i and

$$\delta_i = f(\mathbf{x}_i) - \widehat{f}(\mathbf{x}_i).$$

We consider quantile regression with four different α's: 20%, 30%, 40% and 50%. Because the double exponential distribution is symmetric, the study of these lower quantiles should also be representative of the upper quantiles 60%, 70% and 80%. One hundred datasets were generated. Four quantile regressions were computed for each simulated dataset. Then for

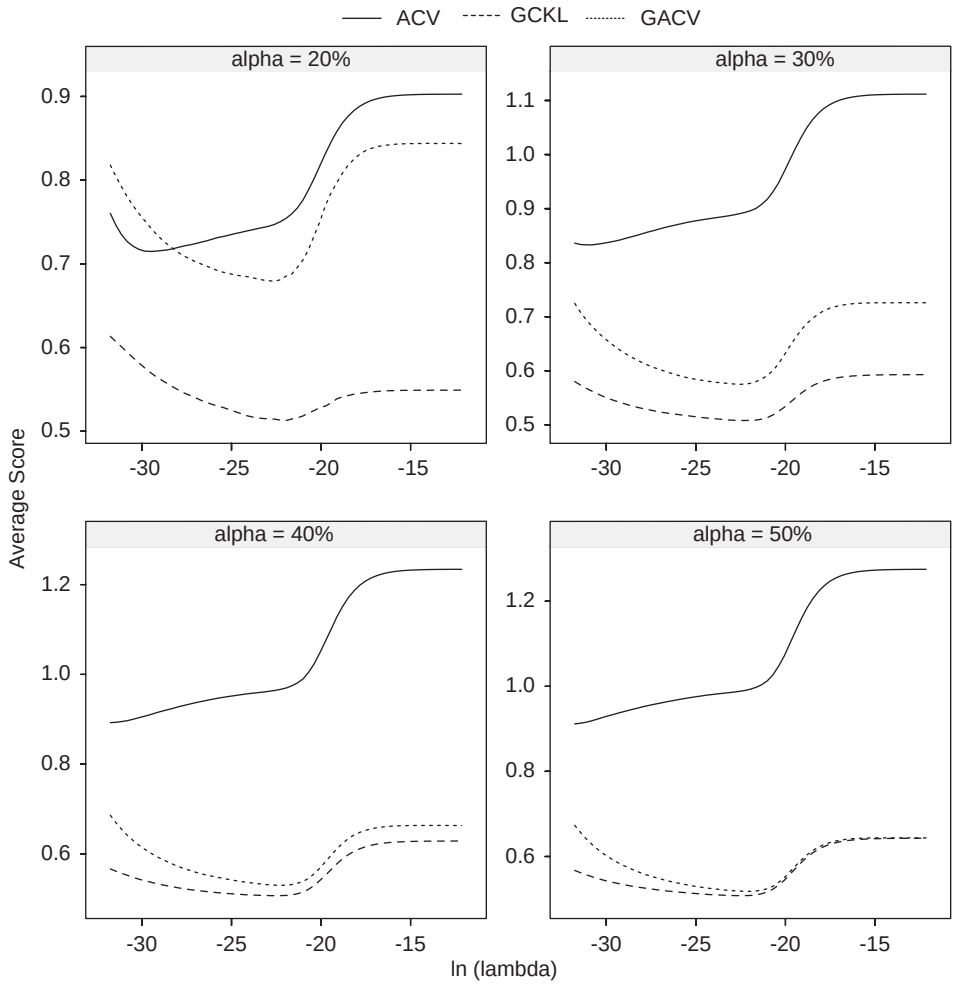

Fig. 2. *GACV and ACV*: We compare the biases of GACV and ACV as estimates of GCKL. In each panel, we plot the average GACV, ACV and GCKL curves for a given quantile from 100 simulated datasets. From these plots, we conclude that GACV is a better estimate of GCKL. It has a very small bias for percentiles close to 50%. It still preserves the shape of GCKL curve even for percentiles far from 50%.

each quantile, the GCKL scores together with GACV scores were evaluated for 80 different smoothing parameters. The natural logarithm of the 80 smoothing parameter are equally spaced in the interval $[-32, -12]$. We averaged the scores over all simulated datasets. Fig. 2 depicts the average curves of GCKL, GACV and ACV versus the natural logarithm of the smoothing parameters. From the figure, we find that the average GACV curves are very close to the average GCKL curves. The closer α is to 50%, the better GACV approximates GCKL. But for all four quantiles, the shape of the average GACV curves are very similar

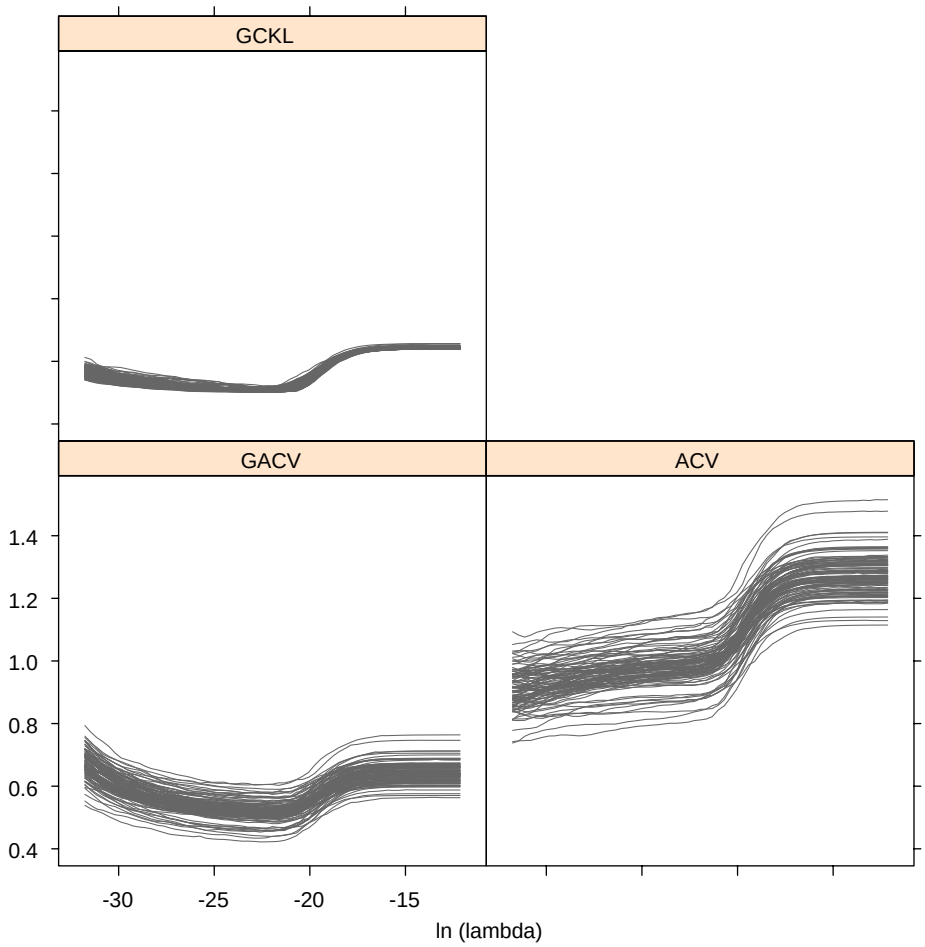

Fig. 3. *Performance of GACV*: This figure presents individual GACV, ACV and GCKL curves for the median smoothing splines from 100 simulated datasets. From this figure, we notice that GACV provides good estimates to GCKL. We also find that ACV fails to capture the minimizer of GCKL very often.

to GCKL curves. In contrast, the average ACV curves have a shape different from GCKL curves and overestimate GCKL all the time.

To examine this issue more closely. We plot individual GACV, ACV, and GCKL curves for $\alpha = 50\%$ from each simulated dataset in Fig. 3. First, we find that for most datasets, ACV fails to pick a reasonable smoothing parameter. On the other hand, GACV is quite successful for most datasets.

It is also worth noting that for extreme quantiles, although GACV and ACV curves have very different shape, they both overestimate the GCKL curve. This similarity, however, vanishes as α gets closer to 50% when the bias of GACV quickly diminishes. According

Table 1
MSE comparisons for different criteria

	MSE	GACV	SIC
DE	0.0324(0.0190)	0.0498(0.0289)	0.1701(0.3377)
Normal	0.0372(0.0223)	0.0528(0.0306)	0.3522(0.3828)
t_3	0.0481(0.0273)	0.0645(0.0345)	0.2421(0.7808)
Mixture	0.0410(0.0250)	0.0589(0.0370)	0.1619(0.2538)
Slash	0.1303(0.0722)	0.3342(0.1248)	0.4054(0.0837)

to our experience, it is generally true that GACV and ACV are relatively more alike for extreme quantiles.

4.2. MSE comparison

In this set of simulations, we focus on the MSE performance of quantile smoothing splines with the smoothing parameter automatically tuned using GACV. We compare GACV and another commonly used criterion, which is defined as

$$SIC(\lambda) = \ln\left(\frac{1}{n}\sum_{i=1}^{n}\rho_\alpha\left(y_i - \widehat{f}(\mathbf{x}_i)\right)\right) + \frac{\ln n}{2n}tr(H).$$

Datasets were stimulated from (4.1) with

$$f(x) = 2\left[\exp\left(-30(x-0.25)^2\right) + \sin\left(\pi x^2\right)\right].$$

This time, we consider five different error distributions from which ε's are sampled: double exponential, standard normal, t-distribution with degree of freedom 3, a mixture distribution

$$0.05N(0, 25) + 0.95\mathrm{N}(0, 1)$$

and a distribution known as the slash distribution, $\mathrm{N}(0, 1)/U(0, 1)$. For each of these five error distributions, 100 datasets were simulated. Given a simulated dataset, we computed the 50% quantile smoothing spline for 80 different smoothing parameters as in the last example. GACV, and SIC scores were computed together with MSE. A GACV-based tuning method will pick smoothing parameter λ_{GACV} such that the associated GACV score is minimized. Similarly, we define λ_{SIC}. We recorded the true *MSE* for $\widehat{f}_{\alpha,\lambda_{\mathrm{GACV}}}$ and $\widehat{f}_{\alpha,\lambda_{\mathrm{SIC}}}$. For the purpose of contrast, we also reported the minimum of *MSE*, which represents the optimal estimate. Table 1 summarizes the sample mean and sample standard deviation (figures in the bracket) for each combination of error distributions and tuning methods.

First, we find that GACV is superior to SIC in all cases both in terms of mean and standard deviation. We also notice that the GACV-based tuning method enjoys a performance close to the optimal.

In Fig. 4 we provide the pairwise comparison between GACV and SIC. Each point in the plot corresponds to a simulated dataset. The y-axis is $MSE(\lambda_{\mathrm{GACV}})$ and the x-axis is $MSE(\lambda_{\mathrm{SIC}})$. From this figure we see that GACV has better performance than SIC for

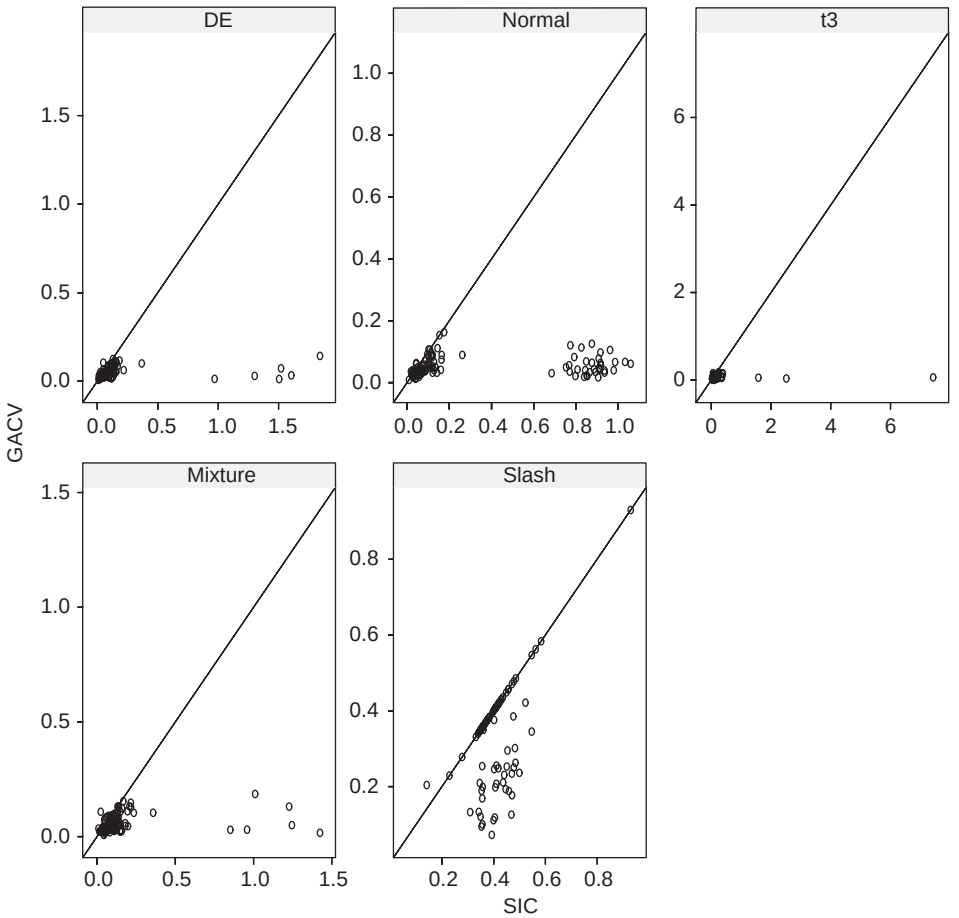

Fig. 4. *GACV versus SIC*: For five different error distributions, we simulated 100 datasets as described in Section 4.2. Each panel presents the MSEs of a GACV-based median smoothing spline versus the MSEs of a SIC-based median smoothing spline. We see a better performance of GACV in terms of MSE.

almost all simulated datasets. An interesting phenomenon is that for each error distribution, SIC broke down for some of the simulated datasets.

4.3. A two-dimensional example

In the last set of simulations, a two-dimensional example is considered. To compute the quantile smoothing splines, we used the thin-plate spline penalty in (1.1)

$$J(f) = \int\int \left(\frac{\partial^2 f}{\partial u^2} + 2\frac{\partial^2 f}{\partial u \partial v} + \frac{\partial^2 f}{\partial v^2} \right)^2 \mathrm{d}u\, \mathrm{d}v.$$

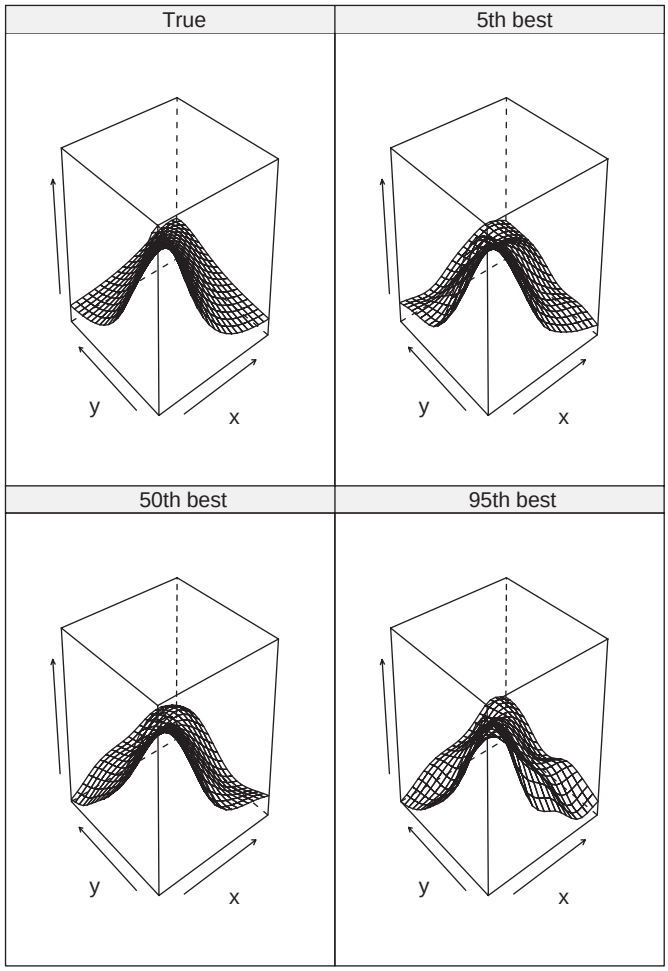

Fig. 5. *Bivariate example*: This figure features a bivariate example described in Section 4.3. The top right panel gives the true test function. Hundred datasets were simulated and GACV-based median smoothing splines were computed. The estimates were ranked by the MSE. The 5th, 50th and 95th best fits are given in the rest three panels.

We chose the following test function

$$f(x_1, x_2) = \frac{40 \exp\left(8\left((x_1 - 0.5)^2 + (x_2 - 0.5)^2\right)\right)}{\exp\left(8\left((x_1 - 0.2)^2 + (x_2 - 0.7)^2\right)\right) + \exp\left(8\left((x_1 - 0.7)^2 + (x_2 - 0.2)^2\right)\right)}.$$

One hundred datasets were generated with the following procedure: 200 independent copies of $\mathbf{x} = (x_1, x_2)$ are sampled from $U(0, 1)^2$. Then response y_i's were sampled according to (4.1) with the double exponential error distribution. For each dataset the 50% quantile smoothing spline was computed with λ chosen to minimize the GACV score. For each

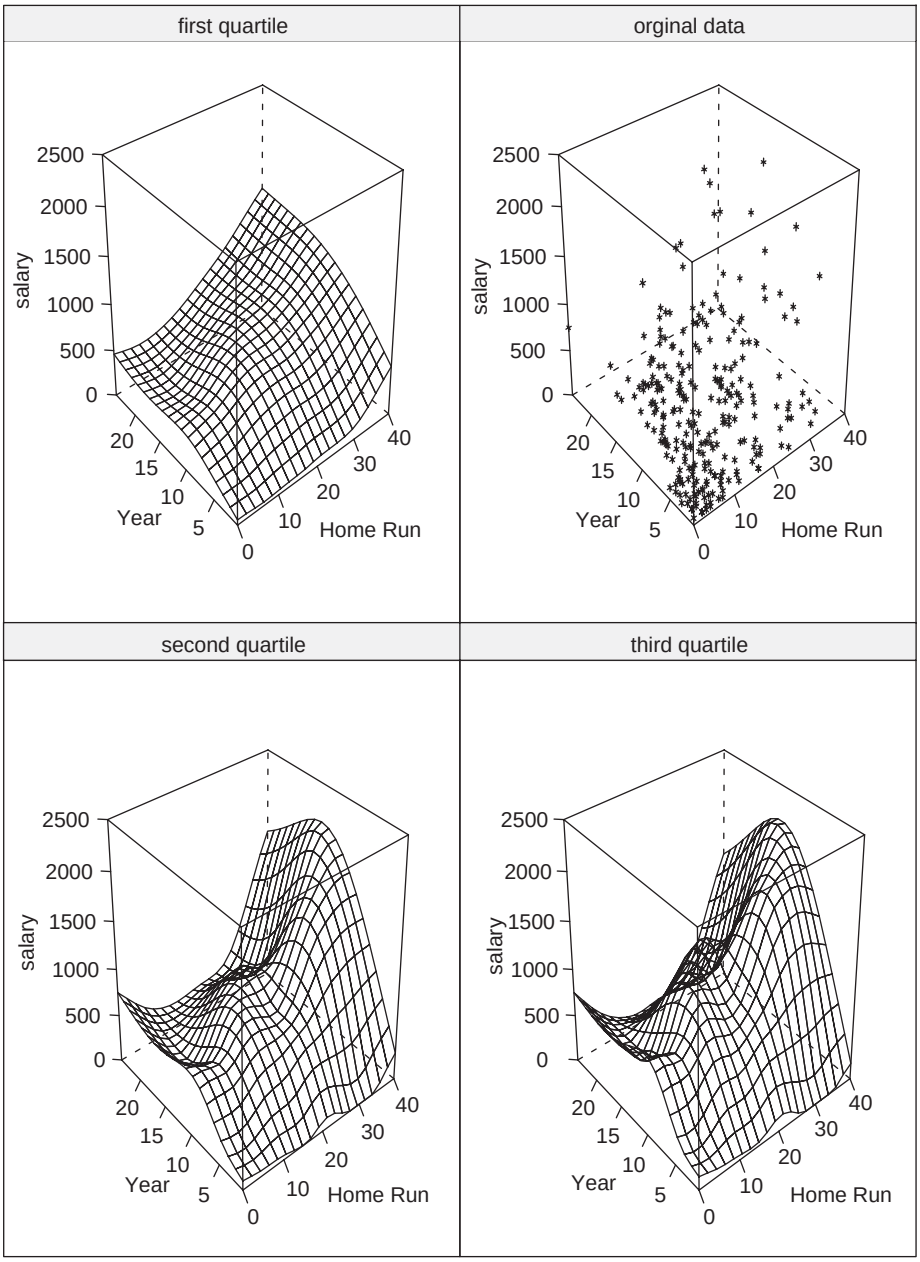

Fig. 6. *Quantile fits for baseball data*: The bottom right panel presents the scatterplot of the Baseball data. The rest three panels give the quantile smoothing splines with smoothing parameters chosen by minimizing GACV scores.

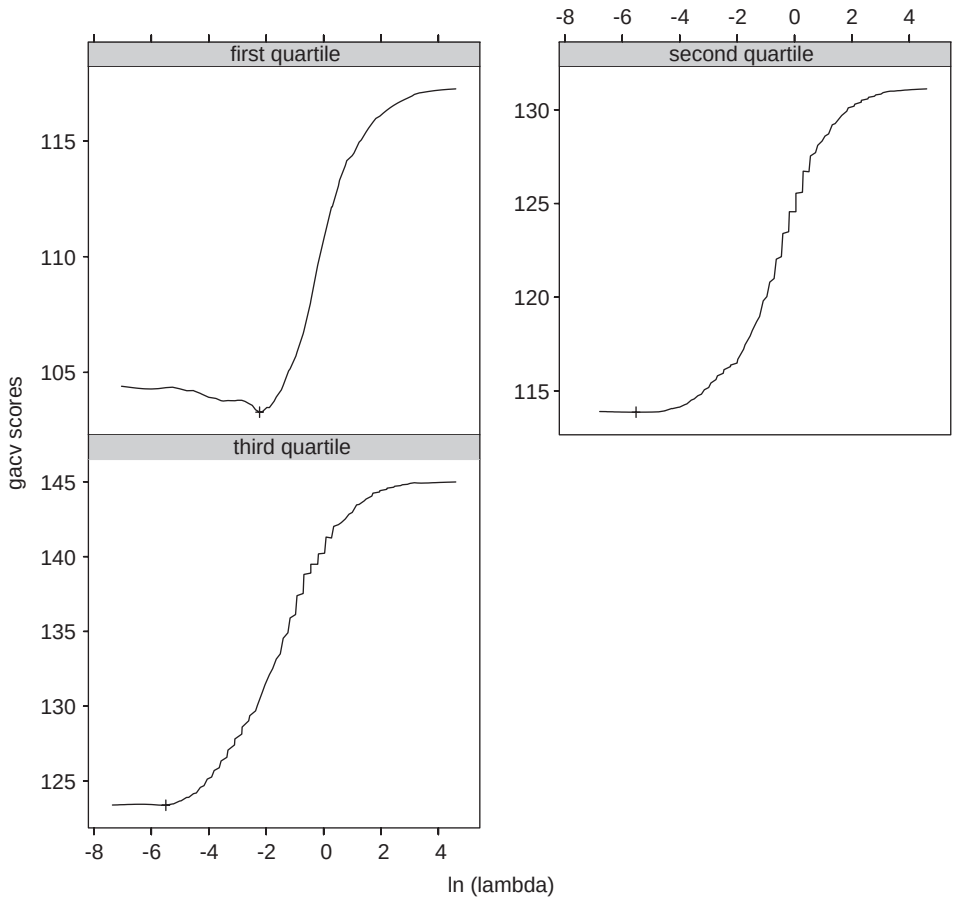

Fig. 7. *GACV for baseball data*: This figure gives the GACV curves for three quartiles from the Baseball example. The minimizers are marked by crosses.

estimate, we calculated the corresponding MSE. Among these 100 quantile smoothing splines, we plotted the true test function together with the 5th, 50th and 95th best fits in terms of MSE in Fig. 5. From this figure, we can see that the GACV-based tuning method has a fairly high statistical efficiency. Even the 95th fit is very close to the true value.

5. Real data analysis

In this section, we considered a real application. We evaluated the annual salary (in thousands of dollars) of baseball players as a function of performance and seniority. The data was obtained from He et al. (1998). It consists of records of 263 North American Major League players for the 1986 season. Following He et al. (1998), we used the number of

home runs in the latest year to measure performance, and the number of years played as the seniority variable. The bottom right panel of Fig. 6 gives the scatterplot of this dataset.

Quantile smoothing splines were fitted to the dataset for the first quartile, median and third quartile. Fig. 7 gives the GACV curves corresponding to each quartile. Fig. 6 gives the quantile smoothing spline estimates with smoothing parameters picked by GACV.

From the fitted quantile smoothing splines we can see that lower income players tended to get higher salary if the performance they had in the last season is better. Also, for lower paid players there is a positive effect of the seniority. Players with more experience made more money. However, for higher paid players, the income pattern is somewhat different. They got better paid if they had played for 10–15 years. This agrees with our intuition. For players with higher performance, they usually get paid better at their "golden ages".

ACV- and SIC-based tuning methods have also been experimented on this dataset. Although the shape of the fitted surfaces are roughly the same, AIC selects rather smaller tuning parameters and gives more wiggly estimate. This echoes our findings in the simulations. SIC performs more similarly to GACV except that it selects the tuning parameter slightly larger than GACV.

Acknowledgements

This research is partially supported by NSF Grant DMS-0772292. The author is grateful to the editor, associate editor and an anonymous referee for their insightful comments which helped him greatly improve both the content and the presentation of this paper.

References

Craven, P., Wahba, G., 1979. Smoothing noisy data with spline functions. Estimating the correct degree of smoothing by the method of generalized cross-validation. Numer. Math 31, 377–403.

He, X., Ng, P., Portney, S., 1998. Bivariate quantile smoothing splines. J. Roy. Statist. Soc. Ser. B 60 (3), 537–550.

Koenker, R., Bassett Jr., G., 1978. Regression quantiles. Econometrica 46 (1), 33–50.

Koenker, R., Hallock, K.F., 2001. Quantile regression. J. Econom. Perspect. 15 (4), 143–156.

Koenker, R., Mizera, I., 2002. Elastic and plastic splines: some experimental comparisons. In: Dodge, Y. (Ed.), Statistical Data Analysis Based on the $L1$-norm and Related Methods. Birkhauser, Basel, pp. 405–414.

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

Nychka, D., Gray, G., Haaland, P., Martin, D., O'Connell, M., 1995. A nonparametric regression approach to syringe grading for quality improvement. J. Amer. Statist. Assoc. 90 (432), 1171–1178.

Oh, H., Nychka, D., Brown, T., Charbonneau, P., 2002. Period analysis of variable stars by robust smoothing. J. Roy. Statist. Soc. Ser. A 53, 15–30.

Xiang, D., Wahba, G., 1996. A generalized approximate cross validation for smoothing splines with non-Gaussian data. Statist. Sinica 6 (3), 675–692.