The Korean Communications in Statistics **Vol. 14** No. 1, 2007, pp. 195–203

Iterative Support Vector Quantile Regression for Censored Data*

Jooyong Shim[1] Dug Hun Hong[2] Dal Ho Kim[3] and Changha Hwang[4]

Abstract

In this paper we propose support vector quantile regression (SVQR) for randomly right censored data. The proposed procedure basically utilizes iterative method based on the empirical distribution functions of the censored times and the sample quantiles of the observed variables, and applies support vector regression for the estimation of the quantile function. Experimental results are then presented to indicate the performance of the proposed procedure.

Keywords: Censoring; empirical distribution function; quantile regression; support vector regression.

1. Introduction

The median is a simple and meaningful measure of the center of the thick tailed distribution of the survival times, and can be well estimated even for not too heavy censoring. But usually large or small quantile depends on the input differently from the median, which leads to consider the quantile regression approach. Koenker and Bassett (1978) introduced the quantile regression model. Quantile regression is gradually evolving into an ensemble of practical statistical methods for estimating and conducting inference about models for conditional quantile functions. In the linear quantile regressin model the quantile function

* This work was supported by the Korea Science and Engineering Foundation (KOSEF) grant funded by the Korea government (MOST) (No. R01-2006-000-10226-0).

1) Adjunct Professor, Department of Applied Statistics, Catholic University of Daegu, Kyungbuk 712-702, Korea.

2) Professor, Department of Mathematics, Myongji University, Kyunggido 449-72, Korea.

3) Professor, Department of Statistics, Kyungbuk National University, Daegu 702-701, Korea.

4) Professor, Division of Information and Computer Science, Dankook University, Seoul 140-714, Korea.
Correspondence : chwang@dankook.ac.kr

of the response y_i for a given $\boldsymbol{x}_i$ is assumed to be linearly related to the input vector $\boldsymbol{x}_i$ as follows

$$q(\theta|\boldsymbol{x}_i) = \boldsymbol{x}_i^t\boldsymbol{\beta}(\theta) \quad \text{for } \theta \in (0,1), \tag{1.1}$$

where $\boldsymbol{\beta}(\theta)$ is the θ^{th} regression quantile and its estimator is defined as any solution to minimize the objective function,

$$\sum_{i=1}^{n} \rho_\theta(y_i - \boldsymbol{x}_i^t\boldsymbol{\beta}(\theta)) \quad \text{for } \theta \in (0,1), \tag{1.2}$$

where $\rho_\theta(\cdot)$ is the check function defined as

$$\rho_\theta(r) = \theta r I(r \geq 0) + (\theta - 1) r I(r < 0), \tag{1.3}$$

where $I(\cdot)$ is the indicator function. The median regression estimator is easily seen to be a special case of $\theta = 1/2$.

Powell (1986) studied the censored quantile regression, where observations could not be observed below the fixed level 0 in the regression model. The censored regression quantile estimator is defined as the value of $\boldsymbol{\beta}$ minimizing the objective function,

$$\sum_{i=1}^{n} \rho_\theta[y_i - \max\{0, \boldsymbol{x}_i^t\boldsymbol{\beta}(\theta)\}]. \tag{1.4}$$

Lindgren (1997) suggested a way to estimate a parametric quantile function with local Kaplan-Meier (1958) estimates of the survival functions for more general censored case, which is to estimate the θ^{th} quantile function of response variables by transforming the estimation of the θ^{th} quantile function of response variables into the estimation of the corresponding p^{th} quantile function of the observed variables in iterative method when response variables are censored.

In this paper we propose new nonlinear quantile regression method for censored data using support vector quantile regression (SVQR) of Hwang and Shim (2005), which is called SVQRC. We also derive SVQRC using iterative reweighted least squares (IRWLS) procedure based on modified check function and obtain generalized approximate cross validation (GACV) function in a simple form for selecting hyper-parameters related.

2. Iterative SVQRC Using QP

In this section we present iterative SVQR for censored data using quadratic programming (QP). This is called SVQRC. First, we are going to illustrate SVQR in Hwang and Shim (2005).

Let T_i be the response variable corresponding to input vector $\boldsymbol{x}_i$ or transformation on it, where $i = 1, 2, \ldots, n$. Here $\boldsymbol{x}_i$ is $(p+1)$-dimensional vector of which the first component is set to 1. Let $q(\theta|\boldsymbol{x}_i)$ be the θ^{th} quantile function of T_i given $\boldsymbol{x}_i$ for θ in (0,1) then

$$q(\theta|\boldsymbol{x}_i) = \inf \{\text{t:P}(T_i \leq t|\boldsymbol{x}_i) \geq \theta\}. \tag{2.1}$$

Assume that $q(\theta|\boldsymbol{x}_i)$ is nonlinearly related to input vector $\boldsymbol{x}_i$ as

$$q(\theta|\boldsymbol{x}_i) = \boldsymbol{w}(\theta)^t \boldsymbol{\Phi}(\boldsymbol{x}_i) \quad \text{for } i = 1, 2, \ldots, n, \tag{2.2}$$

where $\boldsymbol{\Phi}(\boldsymbol{x}_i)$ is a nonlinear feature mapping function which is used to allow for the case of nonlinear quantile regression. The input vectors are nonlinearly transformed into a potentially higher dimensional feature space by a nonlinear mapping function $\boldsymbol{\Phi}$ and then a linear quantile regression is performed there. For this nonlinear quantile regression, the solution requires the computations of dot products $\boldsymbol{\Phi}(\boldsymbol{x}_k)^t \boldsymbol{\Phi}(\boldsymbol{x}_l)$, $k, l = 1, \ldots, n$ in a potentially higher dimensional feature space. Under certain conditions (Mercer, 1909), these demanding computations can be reduced significantly by introducing a kernel function $\boldsymbol{K}$ such that $\boldsymbol{\Phi}(\boldsymbol{x}_k)^t \boldsymbol{\Phi}(\boldsymbol{x}_l) = \boldsymbol{K}(\boldsymbol{x}_k, \boldsymbol{x}_l)$. Several choices of kernel functions are possible. Gaussian kernel is the most frequently used kernel. The linear quantile regression can be regarded as special case of nonlinear quantile regression with an identity feature mapping function $\boldsymbol{\Phi}$ in nonlinear regression, that is, $\boldsymbol{K}(\boldsymbol{x}_k, \boldsymbol{x}_l) = \boldsymbol{x}_k^t \boldsymbol{x}_l$.

Let $\boldsymbol{K}$ be $n \times n$ matrix with entries $\boldsymbol{K}(\boldsymbol{x}_k, \boldsymbol{x}_l)$ and $\boldsymbol{K}_i$ be the i^{th} row of $\boldsymbol{K}$. Then, in SVQR the quantile function (2.2) can be rewritten as

$$q(\theta, \boldsymbol{x}_i) = \boldsymbol{K}_i(\boldsymbol{\alpha} - \boldsymbol{\alpha}^*) \quad \text{for } i = 1, 2, \ldots, n, \tag{2.3}$$

where $\boldsymbol{\alpha}$ and $\boldsymbol{\alpha}^*$ are vectors of the solutions α_i and α_i^* for the optimization problem of SVQR using QP with $(t_i, \boldsymbol{x}_i)$, $i = 1, 2, \ldots, n$,

$$\begin{aligned} &\min \frac{1}{2}(\boldsymbol{\alpha} - \boldsymbol{\alpha}^*)^t \boldsymbol{K}(\boldsymbol{\alpha} - \boldsymbol{\alpha}^*) - (\boldsymbol{\alpha} - \boldsymbol{\alpha}^*)^t \boldsymbol{t} \\ &\text{subject to} \quad 0 \leq \alpha_i \leq \theta C \text{ and } 0 \leq \alpha_i^* \leq (1-\theta)C. \end{aligned} \tag{2.4}$$

Here C is a regularization parameter.

We are now going to describe SVQRC. In fact, we can not observe T_i's but can observe $Y_i = \min(T_i, C_i)$ and $\delta_i = I(T_i \leq C_i)$, where C_i is the censoring variable corresponding to $\boldsymbol{x}_i$ for $i = 1, 2, \ldots, n$. C_i's are assumed to be independently

distributed with unknown distribution function G. Since T_i and C_i are assumed to be independent given $\boldsymbol{x}_i$,

$$\begin{aligned} \mathrm{P}(Y_i \le q(\theta|\boldsymbol{x}_i)|\boldsymbol{x}_i) &= 1 - \mathrm{P}(T_i > q(\theta|\boldsymbol{x}_i)|\boldsymbol{x}_i)\mathrm{P}(C_i > q(\theta|\boldsymbol{x}_i)|\boldsymbol{x}_i) \\ &\ge 1 - (1-\theta)\{1 - G(q(\theta|\boldsymbol{x}_i))\}. \end{aligned} \tag{2.5}$$

The right-hand side of (2.5) depends on θ and $G(q(\theta|\boldsymbol{x}_i))$, but does not depend on the distribution of the response variable. Let us denote it by p_i. Then $q(\theta|\boldsymbol{x}_i)$ can be set to satisfy

$$q(\theta|\boldsymbol{x}_i) = \inf\{y : \mathrm{P}(Y_i \le y|\boldsymbol{x}_i) \ge p_i\}, \tag{2.6}$$

that is, $q(\theta|\boldsymbol{x}_i)$ can be the p_i^{th} quantile function of Y_i given $\boldsymbol{x}_i$. If the censoring distribution G is known, the θ^{th} quantile function, $q(\theta|\boldsymbol{x}_i)$, can be obtained by minimizing the objective function iteratively,

$$\sum_{i=1}^{n} \rho_{p_i}(y_i - q(\theta|x_i)), \tag{2.7}$$

where $p_i = 1 - (1-\theta)\{1 - G(q(\theta|\boldsymbol{x}_i))\}$.

It can be extended to the nonlinear case where $q(\theta|\boldsymbol{x}_i) = \boldsymbol{K}_i(\boldsymbol{\alpha} - \boldsymbol{\alpha}^*), i = 1, \ldots, n$, which can be estimated by iterative method as follows:

1. Set $(\boldsymbol{\alpha}, \boldsymbol{\alpha}^*)$ to the initial value $(\boldsymbol{\alpha}^{(0)}, \boldsymbol{\alpha}^{*(0)})$ and $q(\theta|\boldsymbol{x}_i) = \boldsymbol{K}_i(\boldsymbol{\alpha}^{(0)} - \boldsymbol{\alpha}^{*(0)})$.
2. Obtain $p_i^{(l)} = 1 - (1-\theta)\{1 - \widehat{G}(q(\theta|\boldsymbol{x}_i))\}$.
3. Find $(\boldsymbol{\alpha}^{(l+1)}, \boldsymbol{\alpha}^{*(l+1)})$ which is the solution of $(\boldsymbol{\alpha}, \boldsymbol{\alpha}^*)$ for the optimization problem of SVQR with $(y_i, \boldsymbol{x}_i)$, $i = 1, 2, \ldots, n$,

$$\begin{aligned} &\min \frac{1}{2}(\boldsymbol{\alpha} - \boldsymbol{\alpha}^*)^t \boldsymbol{K}(\boldsymbol{\alpha} - \boldsymbol{\alpha}^*) - (\boldsymbol{\alpha} - \boldsymbol{\alpha}^*)^t \boldsymbol{t} \\ &\text{subject to} \quad 0 \le \alpha_i \le C p_i^{(l)},\ 0 \le \alpha_i^* \le C(1 - p_i^{(l)}),\ i = 1, 2, \ldots, n. \end{aligned} \tag{2.8}$$

4. Iterate Steps 2-3 until convergence.

Here, the Kaplan-Meier estimate $\widehat{G}$ of distribution function G of C_i's can be obtained as,

$$1 - \widehat{G}(y) = \begin{cases} \prod_{i:y_{(i)} \le y} \left(\frac{n-i}{n-i+1}\right)^{1-\delta_{(i)}}, & \text{if} \quad y \le y_{(n)}, \\ 0, & \text{otherwise}, \end{cases} \tag{2.9}$$

where $(y_{(i)}, \delta_{(i)})$ is (y_i, δ_i) ordered on y_i for $i = 1, \ldots, n$.

3. Iterative SVQRC Using IRWLS

If we consider a differentiable modified check function instead of check function (1.3), then we can obtain SVQRC using iterative reweighted least squares (IRWLS) procedure based on modified check function, which is much faster in computing and can also have easy derivation of GACV function. In this paper we use the modified check function $\rho_{p,\delta}(\cdot)$ which is attained by providing the differentiability at 0 by differing from the original check function $\rho_p(\cdot)$ in the small interval $(-\delta, \delta)$

$$\rho_{p,\delta}(r) = p\frac{r^2}{\delta}I(r \geq 0) + (1-p)\frac{r^2}{\delta}I(r < 0). \tag{3.1}$$

Now the problem (2.8) becomes obtaining $\boldsymbol{\beta}$ to minimize

$$L(\boldsymbol{\beta}) = \frac{1}{2}\boldsymbol{\beta}^t \boldsymbol{K}\boldsymbol{\beta} + C\sum_{i=1}^{n} \rho_{p_i,\delta}(y_i - \boldsymbol{K}_i\boldsymbol{\beta}). \tag{3.2}$$

Taking partial derivatives of (3.2) with regard to $\boldsymbol{\beta}$ leads to the optimal value of $\boldsymbol{\beta}$ to be the solution to

$$\mathbf{0} = \boldsymbol{K}\boldsymbol{\beta} - C\boldsymbol{K}\boldsymbol{W}\boldsymbol{y} + C\boldsymbol{K}\boldsymbol{W}\boldsymbol{K}\boldsymbol{\beta}. \tag{3.3}$$

Here $\boldsymbol{W}$ is a diagonal matrix with the i^{th} diagonal element w_{ii} obtained from the derivative of the modified check function as

$$w_{ii} = \begin{cases} \frac{2p_i}{\delta} & \text{if} \quad 0 < r_i < \delta, \\ \frac{p_i}{r_i} & \text{if} \quad r_i > \delta, \\ \frac{2(1-p_i)}{\delta} & \text{if} \quad -\delta < r_i < 0, \\ \frac{(p_i - 1)}{r_i} & \text{if} \quad r_i < -\delta, \end{cases} \tag{3.4}$$

where $r_i = y_i - \boldsymbol{K}_i\boldsymbol{\beta}$. The solution to (2.8) can be obtained with $\boldsymbol{W}$ which is composed of the values of $p_i^{(l)}$ and $\boldsymbol{\beta}^{(l)}$ which were obtained in previous steps. Thus, $q(\theta|\boldsymbol{x}_i) = \boldsymbol{K}_i\boldsymbol{\beta}$, $i = 1, \ldots, n$ can be estimated by iterative method as follows:

1. Set $\boldsymbol{\beta}$ to the initial value $\boldsymbol{\beta}^{(0)}$ and $q(\theta|\boldsymbol{x}_i) = \boldsymbol{K}_i\boldsymbol{\beta}^{(0)}$.
2. Obtain $p_i^{(l)} = 1 - (1-\theta)(1 - \widehat{G}(q(\theta|\boldsymbol{x}_i)))$ and obtain $\boldsymbol{W}$ from (2.6) using δ, $p_i^{(l)}$ and $r_i^{(l)} = y_i - \boldsymbol{K}_i\boldsymbol{\beta}^{(l)}$.

3. Obtain $\boldsymbol{\beta}^{(l+1)}$ from $\boldsymbol{\beta}^{(l+1)} = (\boldsymbol{KWK} + \boldsymbol{K}/C)^{-1}\boldsymbol{KWy}$.

4. Iterate Steps 2-3 until convergence.

The functional structure of SVQRC is characterized by regularization parameter C and kernel parameter. The cross validation (CV) technique used in SVR cannot be used in SVQR, since the loss function used in SVQR is not a function of residuals as SVR. To select the hyper-parameters of SVQRC we consider the cross validation (CV) function as follows:

$$CV(\boldsymbol{\lambda}) = \frac{1}{n}\sum_{i=1}^{n}\rho_{\theta,\delta}(y_i - q_{(i)}(\theta|\boldsymbol{x}_i)), \tag{3.5}$$

where $\boldsymbol{\lambda}$ is the set of hyper-parameters and $q_{(i)}(\theta|\boldsymbol{x}_i)$ indicates the quntile function estimated without the i^{th} observation. Since for each candidates of parameters $q_{(i)}(\theta|\boldsymbol{x}_i)$ for $i = 1, \ldots, n$ should be evaluated, selecting parameters using CV function is computationally formidable. Thus we apply GACV function (Yuan, 2006) to select the set of hyper-parameters $\boldsymbol{\lambda}$ for SVQRC as follows,

$$GACV(\boldsymbol{\lambda}) = \frac{\sum_{i=1}^{n}\rho_{\theta,\delta}(y_i - q(\theta|\boldsymbol{x}_i))}{n - \text{trace}(\boldsymbol{H})}, \tag{3.6}$$

where $\boldsymbol{H}$ is the hat matrix such that $q(\theta|\boldsymbol{x}) = \boldsymbol{Hy}$ with the $(i,j)^{th}$ element $h_{ij} = \partial q(\theta|\boldsymbol{x}_i)/\partial y_j$. GACV function cannot be applied to SVQRC using QP since $\boldsymbol{H}$ is not computable. But for SVQRC using IRWLS, C and kernel parameter can be selected by applying (3.6), where $\boldsymbol{H}$ is obtained easily as

$$\boldsymbol{H} = \boldsymbol{K}(\boldsymbol{KWK} + \boldsymbol{K}/C)^{-1}\boldsymbol{KW}. \tag{3.7}$$

4. Numerical Studies

We illustrate the performance of the censored regression method using SVQRC with modified check function through the simulated example on the nonlinear regression cases and real data in Heuchenne and Keilegom (2005). For the nonlinear censored regression case, 100 of x's are equally spaced ranging from 0 to 1, and (t, c)'s are generated as follows.

$$t_i = \sin(2\pi x_i) + 0.5 + \epsilon_{t_i}, \qquad c_i = \sin(2\pi x_i) + 0.5 + \epsilon_{c_i},\ i = 1, \ldots, 100,$$

where ϵ_{t_i}'s and ϵ_{c_i}'s are generated from normal distributions, $N(0, 0.1)$ and $N(cc, 0.1)$, respectively. cc is chosen for 20% censoring proportion. Then the θ^{th} quatile function of given can be modelled as

$$q(\theta|x) = 0.5 + \sin(2\pi x) + \sqrt{0.1}\Phi(\theta)^{-1}.$$

We set $\delta = 0.001$ in the modified check function. By the estimation procedure in Section 3, we have the estimated quantile functions given x. The Gaussian kernel is utilized in this example, which is

$$K(x_1, x_2) = \exp(-\frac{1}{\sigma^2}(x_1 - x_2)^2).$$

For each data set, the regularization parameter C and the kernel parameter σ^2 are obtained from GACV function (3.6).

Figure 4.1 shows the true and estimated quantile functions corresponding to $\theta = 0.25, 0.5$ and 0.75 for one of 50 data sets. The true and estimated quantile functions are represented by solid and dashed curves, respectively. Uncensored and censored data points are denoted by "*" and "o", respectively. From Figure 4.1 we can recognize that the estimated quantile functions behave similarly well

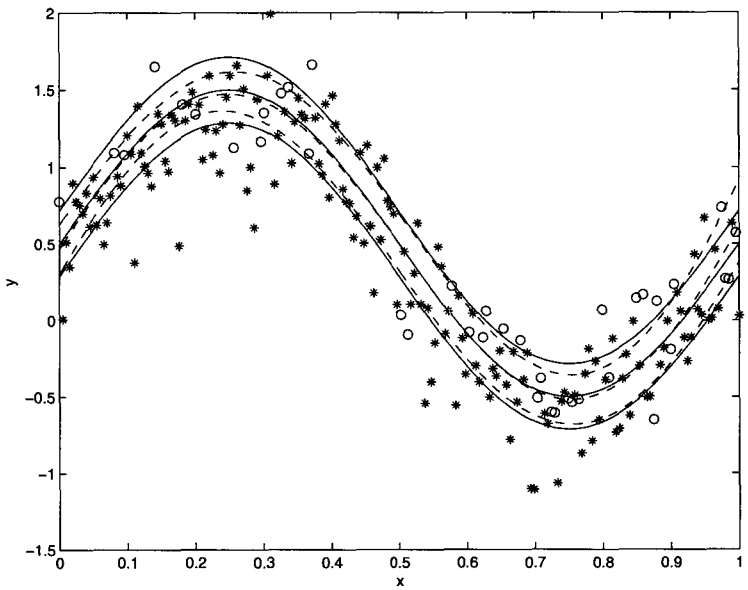

Figure 4.1: The true and the estimated quantile functions corresponding to $\theta = 0.25, 0.5, 0.75$ for 20% censored data

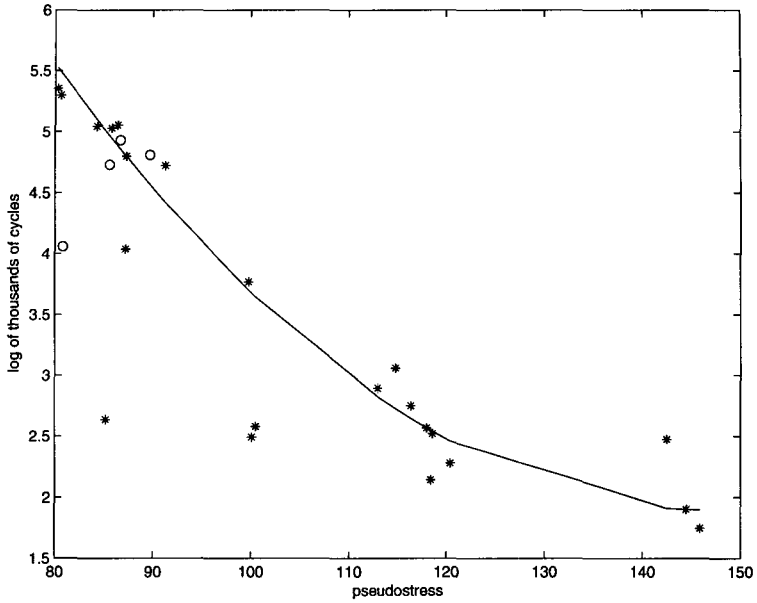

Figure 4.2: The estimated median function for fatigue data by SVQRC

as the true estimated quantile functions do. The mean squared error (MSE) is used for the performance metric,

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(\hat{q}(\theta|x_i) - q(\theta|x_i))^2,$$

where x_i is the input variable, $i = 1, \ldots, n$. From 50 data sets we obtained the average of MSEs as 0.0141, 0.0162, 0.0263, respectively, which indicates that the proposed procedure provides satisfying results.

The median function, that is, 0.5^{th} quantile function, is estimated from the low-cycle fatigue data (Heuchenne and Keilegom, 2005) for a strain-controlled test on 26 cylindrical specimens of nickel-base superalloy, which include 4 censored data. The polynomial kernel with degree 2 is utilized in this example. The regularization parameter is obtained as 5 from GACV function (3.6). Figure 4.2 shows that the logarithms of thousands of cycles before fatigue against pseudostress. The estimated median function for pseudostress by the proposed procedure shows similar values as the estimated mean functions by Heuchenne and Keilegom (2005).

5. Concluding Remarks

In this paper, we dealt with estimating the quantile functions of the censored regression model using SVQRC and obtained GACV function for the proposed procedure. By using GACV function the model selection becomes easier and faster than that by a leave-one-out cross validation. Through two examples we showed that the proposed procedure provides the satisfying results and is attractive approach to modelling of the quantile regression with censored data. We found that the model is not much sensitive to the choice of the regularization parameter C, but it is sensitive to the choice of the kernel parameter σ^2. Thus a consideration such as standardization of input vectors is required for the choice of the kernel parameter for nonlinear cases.

References

Heuchenne, C and Van Keilegom, I. (2005). *Nonlinear Regression with Censored Data.* Technical Report 520, Universite catholique de Louvain.

Hwang, C. and Shim, J. (2005). A simple quantile regression via support vector machine. *Lecture Notes in Computer Science*, **3610**, 512–520.

Kaplan, E. L. and Meier, P. (1958). Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association*, **53**, 457–481.

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

Lindgren, A. (1997). Quantile regression with censored data using generalized L_1 minimization. *Computational Statistics & Data Analysis*, **23**, 509–524.

Mercer, J. (1909). Functions of positive and negative type, and their connection with theory of integral equations. *Philosophical Transactions of the Royal Society*, Ser. A, **209**, 415–446.

Powell, J. L. (1986). Censored regression quantiles. *Journal of Econometrics*, **32**, 143–155.

Yuan, M. (2006). GACV for quantile smoothing splines. *Computational Statistics & Data Analysis*, **50**, 813–829.

[Received November 2006, Accepted February 2007]