Statistica Sinica **9**(1999), 759-774

SMOOTHING REGRESSION QUANTILE BY COMBINING k-NN ESTIMATION WITH LOCAL LINEAR KERNEL FITTING

Keming Yu

Lancaster University

Abstract: A two-step nonparametric regression quantile smoothing technique is presented here, combining a standard k-NN technique and a locally linear kernel smoother. There are many advantages to this approach: an asymptotically optimal mean square error (Fan, Hu and Truong (1995)), a ready-made bandwidth selection rule (Yu and Jones (1998)), and simple computation and flexible estimation under variable transformations and distributional assumptions. The method is tested on a simulated example, and applied to data.

Key words and phrases: Bandwidth selection, correlated regression model, double kernel method, k-NN method, local linear kernel smoothing, mean square error, regression quantile.

1. Introduction

Quantile regression is widely used for screening some biometric measurements (height, weight, circumferences and skinfold) against an appropriate covariate (age, time) (Healy, Rasbash and Yang (1988), Cole (1988), Goldstein and Pan (1992), Royston and Altman (1994)). Some extreme (high or low) quantiles of underlying distributions of measurements are particularly useful for industrial applications (Magee, Burbidge and Robb (1991), Hendricks and Koenker (1992)). In this area many advances in theory and application have been made in the last few years, and some nonparametric and semi-parametric techniques (Jones and Hall (1990), Bhattacharya and Gangopadhyay (1990), Cole and Green (1992), Fan, Hu and Truong (1995), Yu and Jones (1998)) are particularly attractive. Jones and Hall's theoretical investigation based on the kernel fitting of the "check-function" $\rho_p(t) = \{|t| + (2p-1)t\}/2$ of Koenker and Bassett (1978) was extended by Fan, Hu and Truong with an advanced locally linear smoother. The asymptotic mean square error (AMSE) for internal and boundary points is given by

$$AMSE(\hat{q}_p(x)) = \mu_2(K)^2(q_p''(x))^2h^4 + \frac{R(K)}{nhg(x)},$$

where $\hat{q}_p(x)$ is the estimator of the true pth $(0 < p < 1)$ quantile function $q_p(x)$ of response Y given covariate $X = x$, K is a symmetric function with

$\mu_2(K) = \int u^2 K(u)du$, $R(K) = \int K^2(u)du$, $g(x)$ is the density function of x, n is the sample size, and h is the bandwidth. As in fitting a mean function, the AMSE of the local linear kernel fitting quantile function depends only on the second derivative of the quantile function and this approach has no boundary modification. Thus the AMSE should be "optimal" under kernel fitting. A novel bandwidth selection rule based on this optimal AMSE has been explored recently by Yu and Jones (1998) and is given by

$$h_p = h_{mean}\{\frac{p(1-p)}{\phi(\Phi^{-1}(p))^2}\}^{1/5},$$

where h_{mean} is the bandwidth for the smoothing estimation of the regression mean and $\phi(\cdot)$ and $\Phi(\cdot)$ are respectively the standard normal density and distribution function.

Obviously, the "check function" based approach may need some specific algorithms to perform practical calculations, whereas Bhattacharya and Gangopadhyay's (1990) k-NN method is usually simple to calculate. The method is as follows.

Let $\{(X_i, Y_i)\}$, $i = 1, \ldots, n$ be independent and identically distributed as (X,Y), and given $X = x_0$, define $Z = |X - x_0|$. Here $\{(Z_i, Y_i)\}$, $i = 1, \ldots, n$, are i.i.d. from (Z, Y). The order statistics of Z are denoted by $Z_{n1} < Z_{n2} < \cdots < Z_{nn}$ and the induced order statistics of Y by $Y_{n1}, \ldots, Y_{nn}$, i.e., $Y_{ni} = Y_j$ if $Z_{ni} = Z_j$.

For any positive integer $k \leq n$, the k-NN estimator $\overline{q}_p(x)$ of the conditional p-quantile $q_p(x)$ of Y given $X = x_0$ is the p-quantile of the empirical distribution of conditionally independent responses $Y_{n1}, \ldots, Y_{nk}$. So

$$\hat{G}_{nk}(y) = k^{-1} \sum_{i=1}^{k} I(Y_{ni} \leq y),$$

is the c.d.f. and

$$\overline{q}_p(x) = the \quad [kp]th \quad order \quad statistic \quad of \quad Y_{n1}, \ldots, Y_{nk},$$

where I (S) denotes the indicator of the event S.

This k-NN estimator has a Bahadur-type expression as the ordinary quantile (Bahadur (1966)), but the practical performance of k-NN regression quantile estimation is not always satisfactory. A Monte Carlo example from Healy, Rasbash and Yang (1988) throws some light on this problem.

The data $\{(X_i, Y_i)\}_1^n$, $n = 500$, are simulated from the model

$$Y_i = {X_i}^2 + 10\epsilon_i, \quad \epsilon \sim N(0; 1), \quad X_i \sim U(0, 10).$$

Figure 1 shows the median curve of the simulated data set based on the k-NN method. Obviously it is prone to local noise for small k, while it has a heavy tail

(right boundary here) for larger k. Moreover, it is almost impossible to find an approximate "optimal" k for good fitting and smoothing in this case.

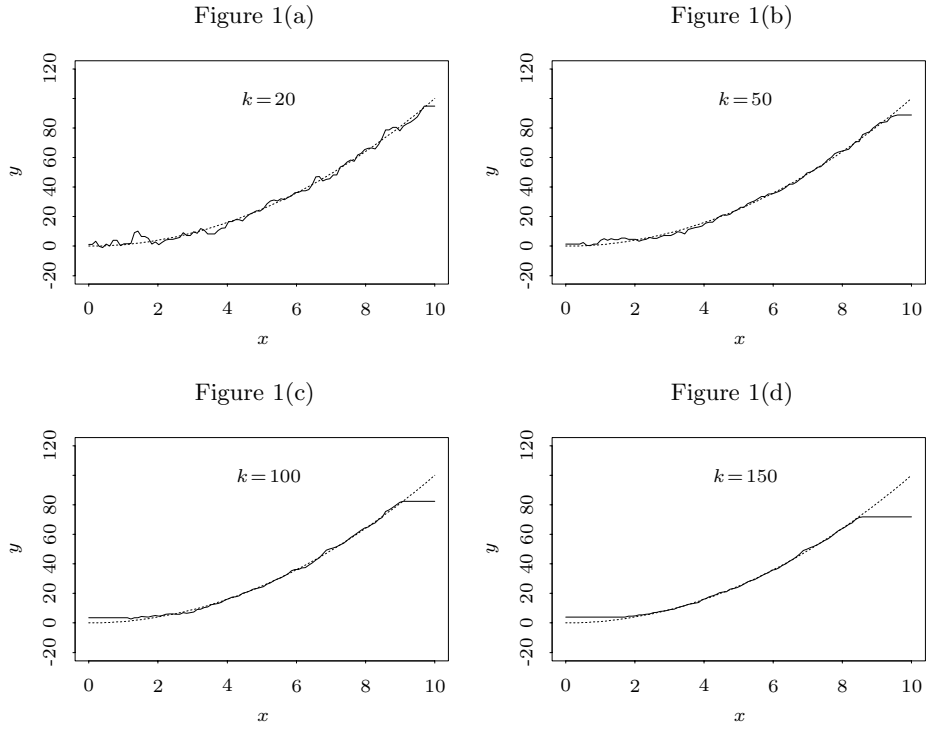

Figure 1. Simulated data ($n = 500$), $y = x^2$ plus Gaussian errors with standard deviation 10. Fitted median by k-NN method. (a) $k = 20$. (b) $k = 50$. (c) $k = 100$. (d) $k = 150$. True median (dotted line) and fitted curves (solid line).

The almost-interpolation feature of the k-NN method which comes from selecting smaller k provides an "initial estimator" of the true quantile even if it is prone either to large variance or boundary bias. One then combines this initial estimator with advanced local linear fitting to get the updated smoothers. For example, take the initial pth quantiles as new samples and smooth them again by some other smoothing techniques. Among other smoothing techniques, the local linear kernel fitting with no boundary modification is a good choice.

An obvious drawback is that the quality of the fitted curves depends on the choice of initial samples, regardless of the smoothing technique applied, and directly smoothing the k-NN points seldom gives a good fit in terms of the quality of k-NN estimators. Healy's rule (Healy, Rasbash and Yang (1988)) for partitioning the covariate value range can help in this context, since this rule successively

and repeatedly takes advantage of original sample information when partitioning the covariate-range into boxes. This rule, called the HRY partition rule in this paper, will be shown to be different from the general age-grouping and binning of data that are usual in statistical analyses.

Section 2 describes this combining method, and Section 3 pcesents a theoretical investigation, including the asymptotic optimal mean square error property. Lastly, Section 4 addresses computational and practical performance issues, and shows that the results are satisfactory and comparable to several existing methods.

2. The Method

The two-step regression quantile smoothing method may be described as follows. First produce a sample of initial quantiles by k-NN at each covariate point. To do this, sort the data by X, denote it by $\{(X_i, Y_i)\}_1^n$ and the sorted $\{Y_i\}_1^n$ can be treated as conditionally independent for $X = x$. The k-NN estimator of the p-quantile $q_p(x)$, for given p and k and for any $X = x$, is based on measurements $\{(X_{i+j-1}, Y_{i+j-1})_{i=1}^n,\ j = 1, \ldots, n-k+1\}$; that is, the HRY rule partitions the covariate range into $n-k+1$ boxes, where the first k points yield the initial estimator at $X = x$ in the first box. Then the procedure is repeated using points 2 to $(k+1)$, 3 to $(k+2), \ldots$, until the entire covariate range has been covered. In short, given p and k, a sample of size n produces $n-k+1$ "initial quantile" samples, but the sample quantiles arising from this first step are irregular and are correlated with each other. Then a local linear kernel fitting based on the $n-k+1$ "initial sample" is used to give the final quantile estimator.

It should be stressed that the selection of k here does not have a big influence on the smoothing results as long as k is not too large ($n-k \to \infty$ when $k, n \to \infty$).

3. The Model and Asymptotic Mean Square Error

For fixed p, let the sequence $\{\xi_{p,i}, \eta_{p,i}\}_{i=1}^m$, $m = n-k+1$, be the pth quantile samples obtained from the k-NN method and HRY rule in the first step, which gives a new set of observations for the p-quantile of the response Y and the corresponding covariate X. The $\{\eta_{p,i}\}_{i=1}^m$ are neither irregular nor smooth, but it is reasonable to assume that they follow a regression model with true function $q_p(x)$:

$$Y = q_p(X) + \epsilon, \tag{1}$$

where $E(\epsilon) = 0$ and $\{\epsilon_i\}$, $i \geq 1$, are correlated errors.

Härdle (1990) summarizes nonparametric regression models for correlated data as follows.

Model (S). There is a stationary sequence $\{(X_i, Y_i), i \geq 1\}$, which may be stochastically dependent, and interest is in estimation of $E(Y|X = x)$.

Model (T). There is a time series $\{Z_i, i \geq 1\}$ and interest is in predicting Z_{n+1} by $E(Z_{n+1}|Z_n = x)$.

Model (C). The observation errors $\{\epsilon_{in}\}$ in the fixed design regression model $Y_{in} = m(i/n) + \epsilon_{in}$ form a sequence of correlated random variables.

Obviously, Model (1) can be approximated by Model (C), but it is not necessarily limited to a fixed design. Among several popular kernel estimators for treating this model, Gasser and Müller's estimator and Priestly and Chao's estimator were investigated respectively by Hart and Wehrly (1986), Hart (1991) and Altman (1990), but neither the Nadaraya and Waston estimator, nor current local polynomial kernel estimators have been analysed so far. Our investigation is restricted to local linear kernel estimators that do not require boundary modification.

For a random vector (X, Y), let g denote the density of X and $f(.|x)$ the conditional density of Y given $X = x$, with corresponding conditional distribution $F(.|x)$. Then

$$F(q_p(x)|x) = p.$$

Given k, for fixed i ($i \in \{1, \ldots, m\}$), let $Y_{k,i}, \ldots, Y_{k,i+k-1}$ be the induced order statistics of $(Z_i, Y_i), \ldots, (Z_{i+k-1}, Y_{i+k-1})$. Then the conditional empirical distribution of Y based on $Y_{k,i}, \ldots, Y_{k,i+k-1}$ is

$$F_{i,k}(y) = 1/k \sum_{j=1}^{k} I(Y_{k,j+i-1} \leq y).$$

Note that the ith p-quantile sample $\eta_{p,i}$, as defined in model (1), is obtained from k conditionally independent samples $\{Y_i, \ldots, Y_{i+k-1}\}$ which are part of the original sample $Y_1, \ldots, Y_n$, $k + m - 1 = n$. Then $\eta_{p,i}$ is the $[kp]$th order statistic of $Y_{k,i}, \ldots, Y_{k,i+k-1}$ and

$$\eta_{p,i} = inf\{y : F_{i,k}(y) \geq [kp]/k\}, \quad i = 1, \ldots, m.$$

Obviously, if $i = 1$, both $\eta_{p,1}$ and $\eta_{p,1+j}$ for $j = 1, \ldots, k-1$ are related to $\{Y_l\}_{l=j+1}^{k}$, and for any $u > k$, $\eta_{p,1}$ and $\eta_{p,u}$ are independent.

On the other hand, from Theorem N1 of Bahattacharya and Ganaopadhyay (1990), $\eta_{p,i}$, $1 \leq i \leq m$, has a Bahadur-type representation as a sum of k independent error random variables.

Write

$$\eta_{p,i} - q_p(x) = \beta(q_p(x) + \frac{1}{kg(q_p(x)|x)} \sum_{j=i}^{i+k-1} W_j(x) + R_k,$$

where

$$\beta(q_p(x)) = -\frac{f(x)F^{2,0}(q_p(x)|x) + 2f'(x)F^{1,0}(q_p(x)|x)}{24f^3(x)g(q_p(x))}$$

with

$$F^{r,0}(q_p(x)|x) = \frac{\partial^r}{\partial x^r}F(y|x)|_{x=x,y=q_p(x)}, \quad r = 1, 2.$$

Then asymptotically

$$max_{k\in N}|R_k| = O(n^{-3/5}logn),$$

and for each k, the $\{W_j(x_0)\}_{j=i}^{i+k-1}$ are independent random variables with mean 0 and variance $p(1-p)$.

Now define

$$\epsilon_i = \frac{1}{kg(q_p(x)|x)} \sum_{j=i}^{i+k-1} W_j(x).$$

Then clearly

$$\text{Var}\,(\epsilon_i) = \frac{p(1-p)}{kg^2(q_p(x)|x)}$$

and, for any $\nu = 0, 1, \ldots$, the covariance of ϵ_i and $\epsilon_{i+\nu}$ depends only on ν . In fact when $i = 1$,

$$\text{Cov}\,(\epsilon_1, \epsilon_{1+\nu}) = \begin{cases} \frac{(k-\nu)p(1-p)}{k^2g^2(q_p(x)|x)}, & \nu = 0, \ldots, k-1, \\ 0, & \nu \geq k. \end{cases}$$

For any k, and sufficiently large n,

$$\sum_{\nu=1}^{\infty} \text{Cov}\,(\epsilon_1, \epsilon_{1+\nu}) = \sum_{\nu=1}^{k-1} \text{Cov}\,(\epsilon_1, \epsilon_{1+\nu}) = \frac{(k-1)p(1-p)}{2kg^2(q_p(x)|x)}.$$

This completes the proof of Theorem 1 below.

Theorem 1. *Let the sample $\{\xi_{p,i}, \eta_{p,i}\}_{i=1}^m$ of model (1) be generated from a random sample of n ordered pairs $\{(X_i, Y_i)\}_1^n$ by the HRY partition rule and the k-NN method. Then the model random errors constitute a stationary process with covariance function*

$$E\{\epsilon_i, \epsilon_j\} = \sigma^2(x)\rho(|i-j|), \tag{2}$$

where $\sigma^2(x)$ is given by $\frac{p(1-p)}{kg^2(q_p(x)|x)}$ and

$$\rho_\nu = \begin{cases} \frac{k-\nu}{k}, & \nu = 0, \ldots, k-1, \\ 0, & \nu \geq k. \end{cases}$$

Further, we have

Theorem 2. *Given i.i.d. observations* $\{(X_i, Y_i)\}_1^n$, *under the conditions of Theorem* 1, *if* $h \to 0$ *and* $nh \to \infty$, *the local linear kernel estimator* $\hat{q}_p(x)$ *of* $q_p(x)$ *based on regression model* (1) *with* 2*nd order symmetric kernel* K *satisfies the following.*

(i) *Interior property:*

$$E(\hat{q}_p(x) - q_p(x))^2 \approx 1/4(q_p''(x))^2\mu_2^2(K)h^4 + \frac{R(K)}{mh}\left(\frac{p(1-p)}{kg(q_p(x)|x)^2} + 2\sum_{\nu}\rho(\nu)\right)$$
$$= 1/4(q_p''(x))^2\mu_2^2(K)h^4 + \frac{R(K)p(1-p)}{mhg(q_p(x)|x)^2}, \tag{3}$$

where $\mu_2(K) = \int u^2K(u)du$, $R(K) = \int K^2(u)du$.

(ii) *Boundary behavior: Assume* $x \in [0,1]$, *then for left-boundary points* $x = ch$ *with* $c > 0$,

$$E(\hat{q}_p(x) - q_p(x))^2 \approx 1/4(q_p''(0+))^2\left\{\frac{s_{2,c}^2 - s_{1,c}s_{3,c}}{s_{2,c}s_{0,c} - s_{1,c}^2}\right\}^2h^4$$
$$+\frac{\int_{-\infty}^c[s_{2,c} - us_{1,c}]^2K^2(u)du}{[s_{2,c}s_{0,c} - s_{1,c}^2]^2}\left(\frac{p(1-p)}{kg(q_p(0+)|0+)^2} + 2\sum_{\nu}\rho(\nu)\right)$$
$$= 1/4(q_p''(0+))^2\left\{\frac{s_{2,c}^2 - s_{1,c}s_{3,c}}{s_{2,c}s_{0,c} - s_{1,c}^2}\right\}^2h^4$$
$$+\frac{\int_{-\infty}^c[s_{2,c} - us_{1,c}]^2K^2(u)du}{[s_{2,c}s_{0,c} - s_{1,c}^2]^2}\frac{p(1-p)}{mhg(q_p(0+)|0+)^2}, \tag{4}$$

where $s_{l,c} = \int_{-\infty}^c K(u)u^l du, \quad l = 0,1,2,3,\ m = n - k + 1.$

Proof of Theorem 2. To prove this theorem the following lemma is required. (See Lemma 4 and Lemma 2 of Fan and Gijbels (1992).)

Lemma. *Assume that* $g(\cdot)$, $K(\cdot)$ *and* $S(\cdot)$ *are bounded and continuous functions in* $[0,1]$ *and right continuous at* $x = 0$. *Suppose further that* $\limsup_{u\to-\infty}|K(u)u^{l+2}| < \infty$ *for a nonnegative integer* l.

(i) *For interior points* $x \in [0,1]$,

$$\sum_{j=1}^n K\left(\frac{x - X_j}{h}\right)S(X_j)(x - X_j)^l = nh^{l+1}S(x)g(x)\int_{-\infty}^{+\infty}K(u)u^l du(1 + o_P(1)).$$

(ii) *For left boundary points* $x = ch_n$, *when* $h_n \to 0$,

$$\sum_{j=1}^n K\left(\frac{x - X_j}{h}\right)S(X_j)(x_n - X_j)^l = nh^{l+1}S(0+)g(0+)\ int_{-\infty}^c K(u)u^l du(1 + o_P(1)).$$

Since

$$\hat{q}_p(x) = \sum_{j=1}^{m} w_{h,m}(x,j)\eta_{p,j},$$

where the weights are local linear kernel fitting weights,

$$w_{h,m}(x,j) = \frac{K(\frac{x-\xi_{p,j}}{h})(S_{m,2} - (x-\xi_{p,j})S_{m,1})}{S_{m,2}S_{m,0} - S_{m,1}^2},$$

with

$$S_{m,l} = \sum_{1}^{m} K(\frac{x-\xi_{p,j}}{h})(x-\xi_{p,j})^l, \quad l = 0,1,2.$$

Conditioning on covariates $\xi_{p,j}$, $j = 1,\ldots,m$, and letting $\Sigma_m(\cdot)$ be the covariance of the observations, $w_{h,m}(x,\cdot)$ the column vector $w_{h,m}(x,j)$, and $w_{h,m}(x,\cdot)^T$ the transpose of $w_{h,m}(x,\cdot)$, the mean squares error of $\hat{q}_p(x)$ is

$$MSE(x,h,m,p) = (w_{h,m}(x,\cdot)^T q_p(\cdot) - q_p(x))^2 + w_{h,m}(x,\cdot)^T \Sigma_m(\cdot) w_{h,m}(x,\cdot).$$

Since the bias term $(w_{h,m}(x,\cdot)^T q_p(\cdot) - q_p(x))$ is not affected by the correlation structure and has the same asymptotic form as the bias provided by Fan (1993), it follows that

(i) for an interior point

$$w_{h,m}(x,\cdot)^T q_p(\cdot) - q_p(x) = -1/2h^2\mu_2(K)q_p''(x) + o(h^2) + o(1/mh),$$

(ii) for a boundary point

$$w_{h,m}(0+,\cdot)^T q_p(\cdot) - q_p(0+) = -1/2h^2\alpha(K,c)^2 q_p''(0+) + o(h^2) + o(1/mh),$$

with $\alpha(K,c) = \frac{s_{2,c}^2 - s_{1,c}s_{3,c}}{s_{2,c}s_{0,c} - s_{1,c}^2}$.

To derive the variance term, apply the lemma with $S(\cdot) = \sigma^2(\cdot)(1+2\sum_{\nu=1}^{\infty}\rho(\nu))$. Note that the correlation function $\rho(\cdot)$ is independent of the covariate, and $\sum_{\nu}^{m}|\rho(\nu)|$ converges as $m \to \infty$. Thus

$$\begin{aligned}
&|w_{h,m}(x,\cdot)^T\Sigma_m(\cdot)w_{h,m}(x,\cdot) - \frac{R(K)}{nh}\sigma^2(\cdot)(1+2\sum_{\nu}\rho(\nu))| \\
&\le |w_{h,m}(x,\cdot)^T\Sigma_m(\cdot)w_{h,m}(x,\cdot) - w_{h,m}(x,\cdot)^T\sigma^2(\cdot)(1+2\sum_{\nu}\rho(\nu))w_{h,m}(x,\cdot)| \\
&\quad + |w_{h,m}(x,\cdot)^T\sigma^2(\cdot)(1+2\sum_{\nu}\rho(\nu))w_{h,m}(x,\cdot) - \frac{R(K)}{nh}\sigma^2(\cdot)(1+2\sum_{\nu}\rho(\nu))| \\
&\le o(1/mh) + \sigma^2(\cdot)(1+2\sum_{\nu}\rho_\nu)|w_{h,m}(x,\cdot)^T w_{h,m}(x,\cdot) - \frac{R(K)}{nh}|.
\end{aligned}$$

Then from the above lemma,

$$|w_{h,m}(x,\cdot)^T w_{h,m}(x,\cdot) - \frac{R(K)}{mh}| = o(1/mh),$$

and at interior points

$$MSE(x,h,m,p) = 1/4h^4\mu_2(K)^2 q_p''(x)^2 + \frac{R(K)}{mh}\sigma^2(1+2\sum_{\nu}\rho(\nu)) + o(1/mh) + o(h^4),$$

and hence (i).

Part (ii) for boundary points $x = ch$ with $c > 0$ and $h \to 0$ can be proved along same lines as (i).

Remark 1. It is interesting and important that the asymptotic pointwise mean square error (AMSE) shows that this two-step regression quantile smoothing method yields the same result as the direct minimization of the "check function" by local linear kernel fitting (Fan, Hu and Trong (1995)).

Remark 2. The pointwise mean square error at interior points in Theorem 2 still holds for Gasser and Müller smoothing (Hart and Wehrly (1986), Hart (1991)) and Priestly and Chao smoothing (Altman (1990)). These smoothers can also be used in practice but with boundary kernel modification.

Remark 3. Another interesting feature of this method is that the AMSE is independent of k, but asymptotically it is required that $n \to \infty, k \to \infty$, and $n - k \to \infty$.

4. Bandwidth Selection and Numerical Examples

Theorem 2 shows that the asymptotically optimal bandwidth for interior points is

$$h^5 = \frac{R(K)p(1-p)}{g^2(q_p(x)|x)q_p''(x))^2\mu_2^2(K)m}. \tag{5}$$

As mentioned in Remark 1 above, this method of combining k-NN estimation with local linear kernel mean fitting for smooth conditional p-quantile based on n independent samples is asymptotically equivalent to the local linear kernel weighting "check-function" based on m independent samples, in the sense of asymptotic mean square error. Recall the rule-of-thumb for bandwidth selection rule derived from Yu and Jones (1998):

(a) use ready–made, and sophisticated, methods to select h_{mean} such as the technique suggested by Ruppert, Sheather and Wand (1995);

(b) use $h_p = h_{mean}\{\frac{p(1-p)}{\phi(\Phi^{-1}(p))^2}\}^{1/5}(\frac{n}{m})^{1/5}$ to obtain all other h_p's from h_{mean}.

Clearly, for fixed p the bandwidth selection relates to the values of m and n, and h_p here is generally a bit bigger than that used in local linear kernel fitting.

The Associate Editor pointed out a deficiency: using the global bandwidth in problems where the designs are not uniform might undersmooth the quantile curves in one area but oversmooth in another. This is because a single k that appeals to the k-NN method might produce many points in one area but few in another area. However, this is not a big concern, as the design-adaptive local linear kernel smoothing technology is employed in the final estimation.

On the other hand, if the deficiency above occurs in practice, then various k in the k-NN method may be used to avoid the shortcoming.

4.1. Smoothing quantile curves

The method is applied to fit the median using different values of k.

First, data are simulated from the model of Section 1 with sample size $n = 500$. A normal kernel with $h_{mean} = 1$, selected subjectively, is used to fit the median with $k = 10, 20, 50$ and 100. Figure 2 is based on 100 simulations. It is seen from Figure 2 that $k \geq 50$ is not necessary for this method and that changing k has very little smoothing effect on the fitted curve. The same conclusions can be drawn for fitting other quantiles.

Figure 2(a) Figure 2(b)

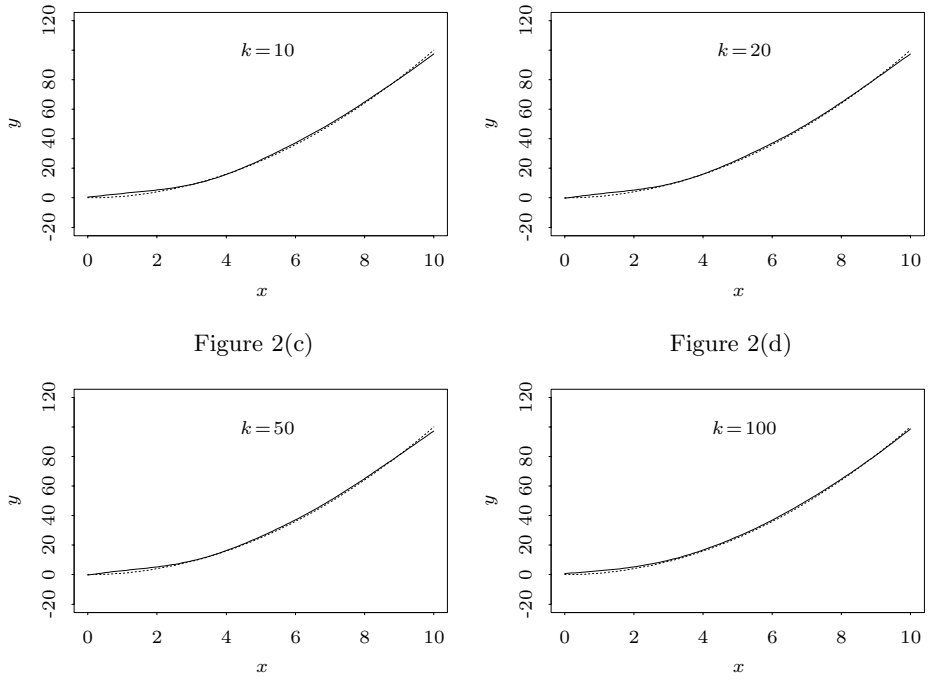

Figure 2(c) Figure 2(d)

Figure 2. Simulated data ($n = 500$), $y = x^2$ plus Gaussian errors with standard deviation 10. Fitted median by two-step method with 100 simulations.

(a) $k = 10$. (b) $k = 20$. (c) $k = 50$. (d) $k = 100$. True median (dotted line) and fitted curves (solid line).

Secondly, two practical datasets are used here, one is approximately normally distributed with $n = 298$ while the other is skew, with $n = 4011$. We will refer to them as the serum concentration data and the US girls' weight data (Yu and Jones (1998)). We employ the rule δ (a) and (b) derived from equation (5) for h_p selection.

(i) US girls weight data with $h_{mean} = 1.8$ for $k = 30, 50$. The seven fitted quantiles are

$\{p = 0.5, 0.25, 0.75, 0.9, 0.1, 0.97, 0.03\}$ (Figure 3).

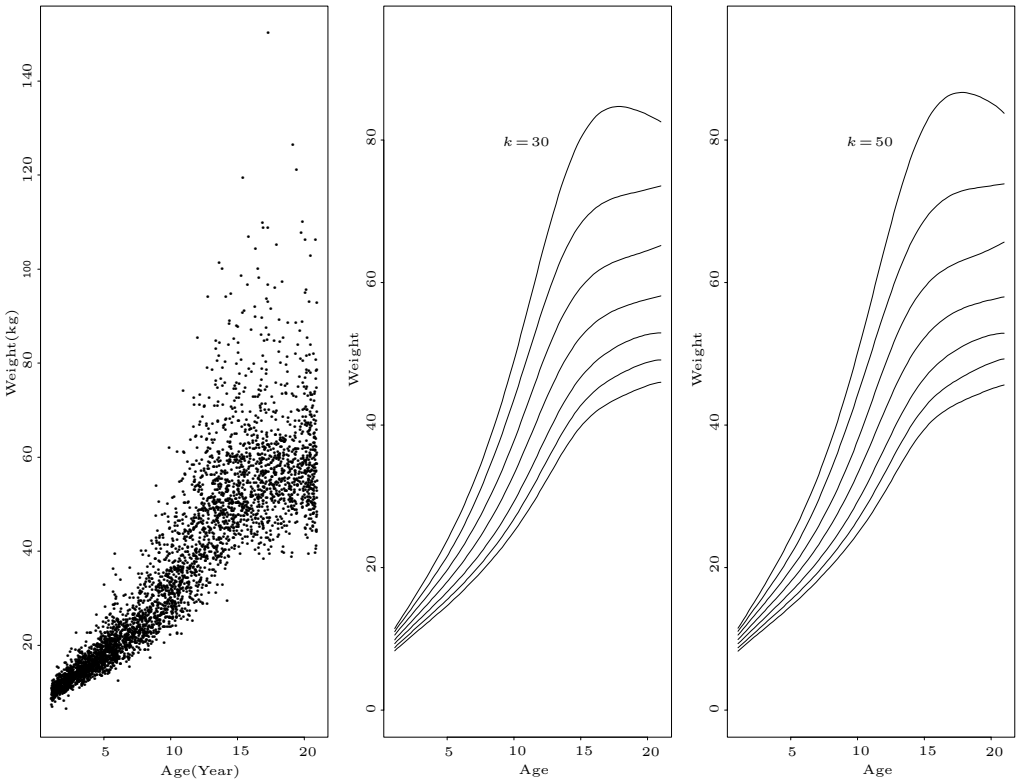

Figure 3. Scatter plots of body weight of 4011 U.S. girls aged between 1 and 21 years. (a) Smoothed reference centile curves for the US girls' weight data at 3rd, 10th, 25th, 50th, 75th, 90th and 97th quantiles. (b) $k = 30$. (c) $k = 50$.

(ii) Serum concentration data (IgG) with $h_{mean} = 0.5$ and $k = 20, 30$. The seven fitted quantiles are

$\{p = 0.5, 0.25, 0.75, 0.9, 0.1, 0.95, 0.05\}$ (Figure 4).

Figure 4(a) Figure 4(b)

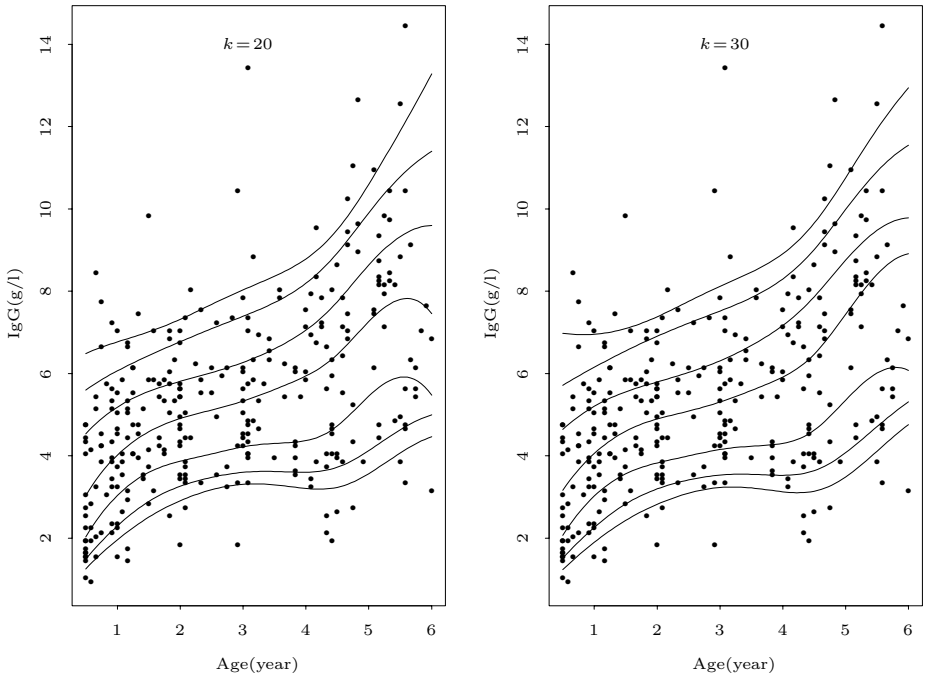

Figure 4. Smoothed reference centile curves for the immunoglobin–G data at 5th, 10th, 25th, 50th, 75th, 90th and 95th quantiles. (a) $k = 20$. (b) $k = 30$.

Clearly, the differences for moderate centile curves and extreme centile curves based on different k values are small for the two data sets.

4.2. A comparison based on simulation

As we have seen, this proposed two-step method for smoothing regression quantiles is related to three existing methods: the HRY method (Healy, Rasbash and Yang (1988)), the k-NN method (Bhattacharya and Gangopadhyay (1990)) and local linear kernel fitting using the "check function" (Fan, Hu and Truong (1995)).

Within the "check function" group, several smoothing techniques, in the category of spline smoothing with roughness penalty have been discussed by Koenker, Ng and Portong (1994), He and Shi (1994) and He (1997). Kernel smoothing and spline smoothing are the two main smoothing techniques, and it should be noted that the reason kernel smoothing, not spline smoothing, is used here, is to try to use the ready-made smoothing parameter selection rule.

Whatever the smoothing method, the key issue in estimation lies in smoothing parameter selection. As is the case with kernel smoothing ($\sum_i \psi(Y_i - a)K(\frac{x-X_i}{h})$) and spline smoothing ($\sum_i \psi(Y_i - g(x_i)) + \lambda \int (|g''(x)|^r)^{1/r} dt$ with ψ the quadratic function or the check function), the efficiency of kernel smoothing and spline smoothing for regression modelling is the same if both smoothing parameters are correctly choosen. As a matter of fact, the two smoothing techniques, in terms of the asymptotic relationship of their smoothing parameters, are equivalent (Cox (1983), Silverman (1984)).

It would be useful to compare the new methodology with these, and also to the double-kernel method (Yu and Jones (1998)) and the restricted regression quantile method (RRQ) (He (1997)).

To highlight the advantages and disadvantages of each approach, we carry out a simulation comparison based on the model

$$Y = 2 + 2\ \cos(X) + \exp(-4X^2) + e,$$

with $X \sim N(0,1)$ and $e \sim E(1)$.

Clearly, the general expression of the pth regression quantile for this model is

$$q_p(x) = 2 + 2\ \cos(x) + \exp(-4x^2) - \log(1-p).$$

The simulations are based on the estimation of three regression quantiles: $p = 0.5, p = 0.9$ and $p = 0.1$, with sample size $n = 500$. Only 100 simulations were performed here.

The regression quantiles in the interval $x \in [-1, 1]$ are estimated. The integrated square error (ISE) is computed as

$$ISE_p = \int_{-1}^{1} (\hat{q}_p(x) - q_p(x))^2 dx.$$

For the HRY method, $k = 50$ is selected subjectively in the first step calculation, and an even polynomial is chosen in the second step:

$$y_p = a_{0p} + a_{1p}x + a_{2p}x^2 + \cdots + a_{Tp}x^T,$$

where the coefficients $\{a_{jp}\}_{j=0}^{T}$ are given by

$$a_{ip} = b_{p0} + b_{p1}\Phi^{-1}(p) + \cdots + b_{pg_i}(\Phi^{-1}(p))^{g_i}.$$

The higher the degree of polynomial y_p, the more accurate the fitting. Here $T = 8$, $g_0 = 1$ and $g_1 = g_2 = 0$.

Regression B-splines are chosen to fit the RRQ. In the other methods, k or h was selected to minimize asymptotic MISE. For example, the $k = [\hat{k}]$ which minimizes the AMISE of k-NN method from Section 3 is

$$AMISE = (k/n)^4 \int_{-1}^{1} \left(\frac{g(x)F^{2,0}(q_p(x)|x) + 2g'(x)F^{1,0}(q_p(x)|x)}{24g^3(x)f(q_p(x)|x)}\right)^2 dx$$

$$+\frac{p(1-p)}{k}\int_{-1}^{1}\frac{1}{f(q_p(x)|x)}dx.$$

The h_p for applying the "check function" method is

$$h_p{}^5=\frac{\frac{R(K)p(1-p)}{n\mu_2^2(K)}\int_{-1}^{1}\frac{1}{g^2(q_p(x)|x)f(x)}dx}{\int_{-1}^{1}(-\frac{F^{2,0}(q_p(x)|x)}{g(q_p(x)|x)}+2\frac{f'(x)q_p'(x)}{f(x)})^2dx}.$$

The h_p for applying the two-step method is

$$h_p{}^5=\frac{\frac{R(K)p(1-p)}{m\mu_2^2(K)}\int_{-1}^{1}\frac{1}{g^2(q_p(x)|x)f(x)}dx}{\int_{-1}^{1}(q_p''(x))^2dx}.$$

The $(h_p, h_{2,p})$ for the applying the double-kernel method is to minimize

$$\begin{aligned}&1/4\int_{-1}^{1}(h^2\mu_2(K)F^{2,0}(q_p(x)|x)/g(q_p(x)|x)\\&+\ h_2^2\mu_2(W)g'(q_p(x)|x)/g(q_p(x)|x))^2dx\\&+\ \int_{-1}^{1}\frac{R(K)}{nhg(x)g^2(q_p(x)|x)}(p(1-p)-h_2g(q_p(x)|x)\alpha(W))dx,\end{aligned}$$

with $W(u)=1/2I(|u|\le 1)$, $\alpha(W)=\int\Omega(t)(1-\Omega(t))dt$ and $\Omega(t)=\int_{-\infty}^{t}W(u)du$.

Table 1 lists the results.

Table 1. ISE of three quantiles estimators by six methods.

method	$p=0.1$	$p=0.5$	$p=0.9$
HRY	0.12	0.09	0.13
k-NN	0.02	0.03	0.29
check function	0.01	0.032	0.18
two-step	0.008	0.023	0.19
double-kernel	0.0069	0.023	0.12
RRQ	0.0068	0.015	0.12

Clearly, the HRY method does not perform as well as the others, although there is no big difference between fitting median and extreme quantiles in terms of ISE. It seems that there is not much difference between the other three methods, but the proposed two-step method has general advantages over each except the double-kernel method and the RRQ method in this simulation. Also, the proposed method's calculations are much quicker than those of the "check function" and double-kernel methods, as the two-step method does not require any iterative calculations.

Two-step estimating or sampling ideas, and methods such as two-step regression and bootstrapping, have existed for a long time in the statistical literature, and have proved very successful in classical statistical analysis and advanced estimation theory. In this paper similar techniques are demonstrated for regression quantile estimation.

Acknowledgements

Thanks are due to Chris Jones and Tim Cole for their helpful comments, and the referee and Associate Editor for useful comments on the earlier version of the paper.

References

Altman, N. S. (1990). Kernel smoothing of data with correlated errors. *J. Amer. Statist. Assoc.* **85**, 749-759.

Bahadur, R. R. (1966). A note on quantiles in large samples. *Ann. Math. Statist.* **37**, 577-580.

Bhattacharya, P. K. and Gangopadhyay, A. K. (1990). Kernel and nearest-neighbor estimation of a conditional quantile. *Ann. Statist.* **18**, 1400-1415.

Cole, T. J. (1988). Fitting smoothed centile curves to reference data, (with discussion). *J. Roy. Statist. Soc. Ser. A* **151**, 385-418.

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

Fan, J. (1993). Local linear regression smoothing and their minimax efficiencies. *Ann. Statist.* **21**, 196-216.

Fan, J. and Gijbels, I. (1992). Variable bandwidth and local linear regression smoothers. *Ann. Statist.* **20,** 2008-2036.

Fan, J., Hu, T.-C. and Truong, Y. K. (1995). Robust nonparametric function estimation. *Scan. J. Statist.* **22**, 433-446.

Goldstein, H. and Pan, H. (1992). Percentile smoothing using piecewise polynomials, with Covariates. *Biometrics* **48**, 1057-1068.

Härdle, W. (1990). *Applied Nonparametric Regression.* Cambridge University Press, Cambridge.

Hart, J. D. and Wehrly, T. E. (1986). Kernel regression estimation using repeated measurements data. *J. Amer. Statist. Assoc.* **81**, 1080-1088.

He, X. and Shi, P. (1994). Covergence rate of B-spline estimators of nonparametric conditional quantile functions. *Nonparametric Statistics* **3**, 299-308.

He, X. (1997). Quantile curves without crossing. *Amer. Statist.* **51**, 186-192.

Healy, M. J. R., Rasbash, J. and Yang, M. (1988). Distribution-free estimation of age-related centiles. *Ann. Human Bio.* **15**, 17-22.

Hendricks, W. and Koenker, R. (1992). Hierarchical spline models for conditional quantiles and the demand for electricity. *J. Amer. Statist. Assoc.* **81**, 58-68.

Jones, M. C. and Hall, P. (1990). Mean square error properties of kernel estimates of regression quantiles. *Statist. Probab. Lett.* **10**, 283-289.

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

Koenker, R., Portnoy, S. and Ng, P. (1992). Nonparametric estimation of conditional quantile functions. In L_1*–Statistical Analysis and Related Methods* (Edited by Y. Dodge), 217-229. Elsevier, Amsterdam.

Koenker, R., Ng, P. and Portnoy, S. (1994). Quantile smoothing spline. *Biometrika* **81**, 673-680.

Magee, L., Burbidge, J. B. and Robb, A. L. (1991). Computing kernel-smoothed conditional quantiles from many observations. *J. Amer. Statist. Assoc.* **86**, 673-677.

Masry, E. and Fan, J. (1994). Local polynomial estimation of regression functions. Technical Report 2311. Chapel Hill, North Carolina.

Pan, H., Goldstein, H. and Yang, Q. (1990). Nonparametric estimation of age-related centiles over wide age ranges. *Ann. Human Bio.* **17**, 475-81.

Royston, P. and Altman, D. G. (1994). Regression using fractional polynomials of continuous covariates: parsimonious parametric modelling, (with discussion). *Appl. Statist.* **43**, 429-467.

Silverman, B. W. (1984). Spline smoothing: the equivalent variable kernel method. *Ann. Statist.* **12**, 898-916.

Yu, K. and Jones, M. C. (1998). Local linear quantile smoothing. *J. Amer. Statist. Assoc.* **93**, 228-238.

Yu, K. and Jones, M. C. (1997). A comparsion of local constant and local linear regresion quantile estimators. *J. Comput. Statist. Data Analy.* **25**, 159-166.

Department of Mathematics and Statistics, University of Lancaster, Lancaster LA1 4YF, U.K.
E-mail:k.yu@lancaster.ac.uk

(Received September 1997; accepted October 1998)