Statistica Sinica **8**(1998), 907-921

CROSS-VALIDATION AND MEDIAN CRITERION

Zhong guo Zheng*† and Ying Yang†

*Hong Kong Baptist University and †Peking University

Abstract: In this paper a new method of selecting the smoothing parameter in nonparametric regression called median cross validation (MCV) is suggested. This method is applied to choose the number of nearest neighbors used in estimating the regression function by local sample medians. Uniform strong consistency is obtained under reasonable conditions. MCV is effective in dealing with outliers in the data. Simulation results are given to demonstrate its superiority over other methods.

Key words and phrases: Cross-validation, median, nearest neighbor median estimates, uniformly strong consistency.

1. Introduction

Consider the nonparametric regression model

$$Y_i = g(x_i) + e_i, \quad i \geq 1, \tag{1.1}$$

where $g(\cdot)$ is a smooth function to be estimated; $\{x_i, i \geq 1\}$ are non-random design points in the interval $[0,1]$; $\{e_i, i \geq 1\}$ are i.i.d. random errors and Y_i is the observation at x_i $(i \geq 1)$. Take a subseries $\{Y_i, 1 \leq i \leq n\}$ from the infinite series $\{Y_i, i \geq 1\}$, and let $x_{i(j)}^{(n)}$ denote the jth nearest neighbor of x_i, i.e., $x_{i(j)}^{(n)}$ is a member of $\{x_1, \ldots, x_n\}$ satisfying the following relation: $|x_i - x_{i(j)}^{(n)}|$ is the jth smallest value among $|x_i - x_{i'}|$, $i' = 1, \ldots, n$. Later on, the superscript (n) will be dropped if there is no confusion. We define

$$\widetilde{g}_{n,h}(x_i) = m(Y_{i(1)}, \ldots, Y_{i(h)}) = \text{median of } Y_{i(1)}, \ldots, Y_{i(h)} \tag{1.2}$$

as the nearest neighbor median estimator of $g(x_i)$. The number h of neighbors plays the role of a smoothing parameter, which has to be selected properly. To obtain asymptotic normality results, one usually needs the condition $h = O(n^{\lambda})$, for some $\lambda \in (0,1)$. Stute (1986) considered the case $\lambda = \frac{2}{3}$, while Bhattacharya and Mark (1987) studied the case $\lambda = \frac{4}{5}$. Although in most theoretical studies, h is deterministically specified, in practice it is better to choose h based on information from the data. One of the most common methods of choosing the smoothing parameter h is cross-validation. Li (1984) proved the consistency for the L_2 cross-validated nearest neighbor estimator in nonparametric regression. Marron (1987,

1989) gave a detailed review of various methods for selecting the smoothing parameter. Yang and Zheng (1992) introduced the L_1 cross-validation method and obtained the weak consistency of the L_1 cross-validated nearest neighbor median estimator under the existence of the first moment of e_i. Gangopadhyay and Sen (1990) also suggested the L_1 cross-validation method. The condition $E(e_i^2) < \infty$ (respectively, $E(|e_i|) < \infty$) is necessary for establishing of weak and strong consistency for L_2 (respectively, L_1) cross-validated estimators. However, when there are outliers in the Y observations (or if the distribution of random errors has a heavy tail so that $E(|e_i|) = \infty$), then it becomes very difficult to obtain good asymptotic results for the $L_2(L_1)$ cross-validation criterion. To overcome such disadvantages, a new method called median cross-validation (abbr. MCV) is introduced. In this paper, the uniform strong consistency of MCV will be established under very mild conditions.

In Section 2, the motivation of median cross validation is presented and some simulation results are given to demonstrate the superiority of the MCV method over the L_1criterion. The main theoretical results are given in Section 3. Technical proofs are given in the Appendix.

2. Motivations and Simulations

In constructing $\widetilde{g}_{n,h}$, the number h of neighbors plays an important role. One way of choosing h is by minimizing the asymptotic mean squared error of the estimate $\widetilde{g}_{n.h}$. Under some regularity conditions, Yang (1996) obtained the following Bahadur type representation

$$\widetilde{g}_{n,h}(x) = g(x) + \frac{1}{2}g''(x)M(x)\left(\frac{h}{n}\right)^2 + \frac{1}{hf(0)}\sum_{i=1}^{h}(\frac{1}{2} - I_{\{e_i(x)\le 0\}}) + R_{nh}, \quad (2.1)$$

where $\max_{h\in H_n(a,b)}|R_{nh}| = o(n^{-1/2}\log n) + O(n^{-3\beta/4}\log n) + o(n^{-2(1-\beta)})$, a.s., $H_n(a,b) = \{k : k_0 = [an^\beta] \le k \le [bn^\beta] = k_1\}, 0 < a < b < \infty, 0 < \beta < 1, M(x) > 0$, is a function dependent on the design points satisfying $0 < m = \inf_{x\in[0,1]} M(x) \le \sup_{x\in[0,1]} M(x) = M < \infty, x \in [0,1]$, $f(x)$ is the density function of $e_i, f(0) > 0$ and the median of e_i is 0. Ignoring the remainder term R_{nh} in (2.1), we get the asymptotic mean squared error (AMSE) of the nearest neighbor median estimate

$$\begin{aligned}\mathrm{AMSE}(h) &= \mathrm{AMSE}(\widetilde{g}_{n,h}(x) - g(x))^2 \\ &= (g''(x))^2M^2(x)(h/n)^4/4 + 1/(4f^2(0)h). \end{aligned} \quad (2.2)$$

From (2.2), we see that, theoretically,

$$h_n(x) = n^{4/5}(4f^2(0)(g''(x))^2M^2(x))^{-1/5}$$

is the optimal choice of h, which minimizes the AMSE of the estimator $\widetilde{g}_{n,h}(x)$. But it cannot be directly applied because it depends on unknown quantities such as $f(0), g''(x)$.

For practical use, it is often preferable to have a data-driven h. One such method is to select h by the cross-validation technique.

The motivation behind cross validation is easily understood (see, Allen (1974), Stone (1974)). In recent years, results on its statistical properties have become available. In density estimation, Chow, Geman and Wu (1983) and Hall (1982) established some asymptotic results for cross validated kernel estimates. In nonparametric regression, Wong (1983), and Li (1984) proved the consistency of the cross validated estimates, for the kernel and the nearest neighbor estimates respectively. (For more references on smoothing parameter selection, see Marron (1987, 1989) and Härdle and Chen(1995)). All the aformentioned results are based on the L_2 norm. On the other hand, Ganganbongan and Sen (1990) suggested the use of the L_1-cross-validation technique to select the smoothing parameter. Yang and Zheng (1992) also considered the L_1-cross validation technique. They proved the weak consistency for L_1-cross-validated nearest neighbor median estimates under the assumption of the finite first moment of the random error.

Let H_n be an index set which will be specified later. The L_2 cross-validated choice of $h \in H_n$ for the nearest neighbor median estimates, based on the average squared prediction error, denoted by $h_2^* = h_2^*(n)$, is the minimizer of

$$\inf_{h\in H_n} \mathrm{cv}_2(h) = \inf_{h\in H_n} \frac{1}{n}\sum_{i=1}^{n}[g(x_i) - \widetilde{g}_{n,h,-1}(x_i)]^2,$$

where $\widetilde{g}_{n,h,-1}(x_i)$ is the delete-one estimate of $g(x_i)$, i.e. $\widetilde{g}_{n,h,-1}(x_i) = m(Y_{i(2)}, \ldots, Y_{i(h)})$. The cross-validation function $\mathrm{cv}_2(h)$ measures the average ability of $\widetilde{g}_{n,h,-1}(x_i)$ to predict the "new" observation $Y_{i(1)} = Y_i$.

The L_1-cross-validation criterion is defined as follows. Let $\mathrm{cv}_1(h) = \frac{1}{n}\sum_{i=1}^{n} |g(x_i) - \widetilde{g}_{n,h,-1}(x_i)|$. Choose h_1^* that minimizes $\mathrm{cv}_1(h)$, i.e., $h_1^* = \arg\ \min_{h\in H_n} \mathrm{cv}_1(h)$.

If $E(e_i^2) = \infty, (E(|e_i|) = \infty$, respectively), then $\mathrm{cv}_2(h) \to \infty (\mathrm{cv}_1(h) \to \infty$, respectively) in probability for all $h \in H_n$. Therefore, it is difficult to give a consistency result for $L_2(L_1$, respectively)-cross-validation.

But recall that the consistency of the deterministically chosen nearest neighbor estimate $\widetilde{g}_{n,h}(x)$ dose not require the error to have finite first moment. Thus both L_1 and L_2 cross-validation criteria are not entirely appropriate. One alternative is to consider the median cross validation criteria defined as follows: Let $\mathrm{cv}_m(h) = m(|Y_1 - \widetilde{g}_{n,h,-1}(x_1)|, \ldots, |Y_n - \widetilde{g}_{n,h,-1}(x_n)|)$ and select h_n^* by $h_n^* = \arg\inf_{h\in H_n} \mathrm{cv}_m(h)$.

A simulation was carried out to show the differences among the three criteria. In Figure 1, the circles denote the data and the solid curve denotes the true curve. The data $(x_1, Y_1), \ldots, (x_{200}, Y_{200})$ come from the nonparametric regression model $Y_i = g(x_i) + e_i, 1 = 1, \ldots, 200$, where the true curve is

$$g(x) = \begin{cases} 100x^3, 0 \le x \le 0.3, \\ 2.7 - 8(x - 0.3), 0.3 < x \le 0.6, \\ 3 - 67.5(x - 0.8)^2, 0.6 < x \le 1, \end{cases}$$

$x_i = i/200$, $i = 1, \ldots, 200$, and the observation errors e_i are assumed to be i.i.d. $(1 - \epsilon)\Phi(x) + \epsilon\Phi(\frac{x}{10})$ with $\epsilon = 2(1 - \Phi(1)) = 0.317$, where $\Phi(x)$ is the standard normal distribution function.

A Monte Carlo simulation was carried out to compare the three criteria. We present the 3 nearest neighbor median estimates with smoothing parameter h^* selected by the L_2, L_1 and MCV criterion respectively. In Figure 2 and Figure 3 the estimates are obtained by L_2 and L_1 criteria for this data set, respectively, and in Figure 4 the curve is fitted by the MCV criterion. For this example, MCV appears better than the L_2 and L_1 criteria for most of the curve.

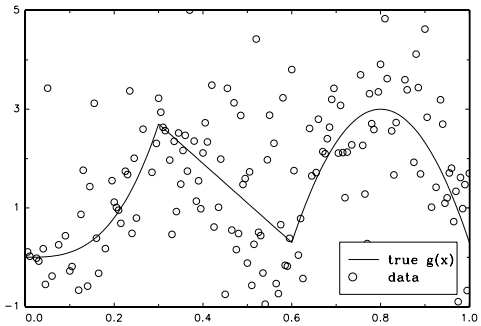

Figure 1. Data and true curve $g(x)$

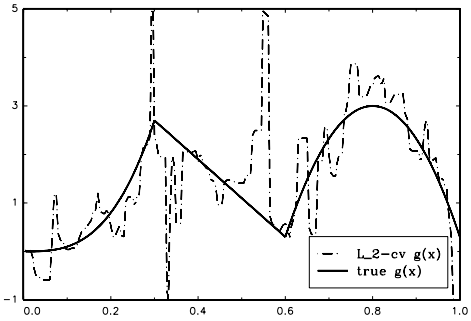

Figure 2. true $g(x)$ and L_2-cv estimate

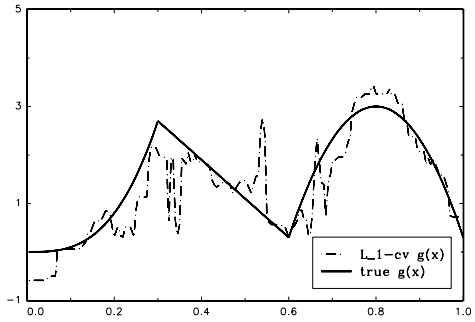

Figure 3. true $g(x)$ and L_1-cv estimate

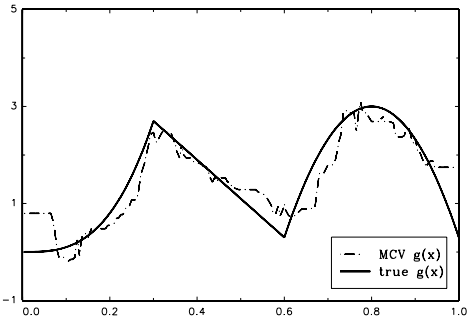

Figure 4. true $g(x)$ and MCV estimate

We repeat this experiment 1000 times. A summary of the result is given in table 1. Here,

$$d(h^*) = \max_{1\le i\le 200} |g(x_i) - \widetilde{g}_{nh^*}(x_i)|, \qquad p_r = P\{d(h^*) > r\}.$$

Table 1. Comparison of the L_2-, L_1-cv and MCV criteria when the error is $(1-0.317)N(0,1)+0.317N(0,10^2)$

criteria	$Ed(h^*)$	std$d(h^*)$	$p_{0.6}$	$p_{0.8}$	p_1	$p_{1.2}$	p_2
L_2-cv	1.653	1.514	0.998	0.904	0.694	0.476	0.151
L_1-cv	1.220	0.674	0.997	0.880	0.609	0.366	0.056
MCV	1.147	0.508	0.991	0.863	0.565	0.312	0.035

Table 2. Comparison of L_2-, L_1-cv and MCV criteria when the error is $N(0,1)$

criteria	$Ed(h^*)$	std$d(h^*)$	$p_{0.6}$	$p_{0.8}$	p_1	$p_{1.2}$	p_2
L_2-cv	0.811	0.171	0.919	0.485	0.148	0.024	0
L_1-cv	0.829	0.173	0.937	0.515	0.164	0.030	0
MCV	0.897	0.234	0.941	0.612	0.268	0.087	0.002

The MCV criterion is better than the L_2- and L_1- CV criteria in gauging the heavy tail distributions. We also study the case that the error follow the standard normal distribution. Table 2 shows that as expected, the L_2-criterion is the best.

3. Main Results

The following are conditions for model (1.1):

(I) $e_1, e_2, \ldots$ are i.i.d. random variables defined on probability space $(\Omega, \mathcal{F}, P)$ with a common distribution function F, $f(x) = F'(x)$ is continuous and positive on $R = (-\infty, \infty)$, $F(0) = \frac{1}{2}$, and $f(x)$ is symmetric about 0 and nonincreasing on $[0, \infty)$. There exist positive numbers c_1 and $\delta(< m(|e|))$ such that $f(x) - f(y) \ge c_1(y - x)$ for all $x < y$, $x, y \in U$, where $U = (m(|e|) - \delta, m(|e|) + \delta)$ and $m(|e|)$ stands for the median of $|e|$;

(II) the function $g(x)$ satisfies a Lipschitz condition of order α ($\alpha \in (0, 1]$), i.e., there exists an $L > 0$ such that $|g(x) - g(y)| \le L|x - y|^\alpha$ for all $x, y \in [0, 1]$;

(III) there exists $c_2 > 0$ such that $\max_{1\le i\le n} |x_i - x_{i(h)}| \le c_2 n^{-1} h \ln n, 1 \le h \le n, n \ge 2$;

(IV) $H_n = \{2, 3, \ldots, \mu_n\}, \mu_n = n/(b_n \ln n) \to \infty, b_n \to \infty$;

Theorem 2.1. *Under conditions* (I)-(IV), *there exists a positive constant* μ *such that for almost all* $\omega \in \Omega$, $h_n^* \ge h_n$ *for all* n *sufficiently large, where* $h_n = [\mu(n(\ln n)^{-2})^{\alpha/(2\alpha+1)}]$, *and* $[x]$ *denotes the integer part of* x.

Theorem 2.2. *Under conditions* (I)-(IV), *the following holds for the Maximum Error* (ME) *of the cross-validated estimate* $\widetilde{g}_{n,h_n^*}$,

$$\mathrm{ME}(h_n^*) \stackrel{\Delta}{=} \max_{1\le i\le n} |\widetilde{g}_{n,h_n^*}(x_i) - g(x_i)| \to 0 \qquad \text{a.s. } n \to \infty.$$

The conditions of Theorem 2.2 are weaker than the conditions for consistency of the L_2 and L_1 cross validation methods but the conclusion is stronger since in the L_2(or L_1) criterion the consistency is defined by $\frac{1}{n}\sum_{i=1}^n (g(x_i)-\widetilde{g}_{n,h^*}(x_i))^2 \to 0$, or $\frac{1}{n}\sum_{i=1}^n |g(x_i) - \widetilde{g}_{n,h^*}(x_i)| \to 0$, which is weaker than $\max_{1\le i\le n} |g(x_i) - \widetilde{g}_{n,h^*}(x_i)| \to 0$.

Remark 1. According to Cheng's results (1984), if $x_1, x_2, \ldots \sim$ i.i.d. $U[0,1]$, condition (III) is satisfied for almost all sequences of sample $x_1, x_2, \ldots$

Remark 2. Condition (IV) is weaker than the algebraic order $\mu_n = n^\lambda, \lambda \in (0,1)$.

Acknowledgement

The authors are very grateful to the referees and editors for invaluable and detailed comments and suggestions which led to the improved form of the manuscript. The research was supported by NSFC and the Doctoral Foundation of Education of China.

Appendix. Proof of Theorem 2.1 and Theorem 2.2.

Let

$$\begin{aligned} a_{i(j)} &= g(x_{i(j)}) - g(x_i),\ 1 \le i \le n,\ j \in H_n, \\ B_{i,h} &= m(a_{i(1)} + e_{i(1)}, \ldots, a_{i(h)} + e_{i(h)}), \\ B_{i,h,-1} &= m(a_{i(2)} + e_{i(2)}, \ldots, a_{i(h)} + e_{i(h)}). \end{aligned}$$

Using this notation, we obtain

$$|Y_i - \widetilde{g}_{n,h,-1}(x_i)| = |e_i - m(a_{i(2)} + e_{i(2)}, \ldots, a_{i(h)} + e_{i(h)})| = |e_i - B_{i,h,-1}|.$$

Let $A_{i,h}$ and $A_{i,h,-1}$ be the unique root of following equations respectively,

$$\sum_{j=1}^{h} F(A_{i,h} - a_{i(j)}) = hF(0), \quad \sum_{j=2}^{h} F(A_{i,h,-1} - a_{i(j)}) = (h-1)F(0).$$

Lemma A.1. (Yang and Zheng (1992)) *Let* $e_1, e_2, \ldots \sim$ *i.i.d.* F *with its density function* $f(x)$ *satisfying the following condition:* f *is bounded away from zero on*

any compact set, (i.e., for every bounded K, there exists $\delta = \delta_K > 0$ such that $f(x) > \delta$ for all $x \in K$). Then for every $M > 0$, there exists $c_0 = c_0(M) > 0$ such that

$$P\{|m(e_1 + a_1, \ldots, e_n + a_n) - A_n| \geq \epsilon\} \leq 2\exp\{-c_0\epsilon^2 n\} \tag{A.1}$$

holds for all $\epsilon > 0$, $|a_i| \leq M$, $i = 1, \ldots, n$, $n \geq 1$, where A_n is the unique solution of the following equation $\sum_{i=1}^n F(x - a_i) = nF(0)$.

Lemma A.2. *Under conditions* (I)-(IV), *there exists a series $\{h'_n : n \geq 1, h'_n \in H_n\}$, such that $cv(h'_n) \to m(|e_1|)$. a.s.*

Proof. Note that

$$\begin{aligned}
|\mathrm{cv}(h) - m(|e_1|)| &= |m(|Y_i - \widetilde{g}_{n,h,-1}(x_i)|, i = 1, \ldots, n) - m(|e_1|)| \\
&= |m(|e_i - B_{i,h,-1}|, i = 1, \ldots, n) - m(|e_1|)| \\
&\leq |m(|e_i|, i = 1, \ldots, n) - m(|e_1|)| \\
&\quad + \max_{1\leq i\leq n} |B_{i,h,-1} - A_{i,h,-1}| + \max_{1\leq i\leq n} |A_{i,h,-1}|.
\end{aligned} \tag{A.2}$$

By the property of the median, we know that

$$|m(|e_i|, i = 1, \ldots, n) - m(|e_1|)| \to 0. \quad \text{a.s.} \tag{A.3}$$

Taking $h'_n = [(\ln n)^2]$, where $[x]$ denotes the integer part of x, in view of (A.1), for every $\epsilon > 0$, we have

$$\begin{aligned}
\sum_{n=2}^{\infty} P\Big\{ \max_{1\leq i\leq n} |B_{i,h'_n,-1} - A_{i,h'_n,-1}| > \epsilon \Big\} &\leq \sum_{n=2}^{\infty} n \max_{1\leq i\leq n} P\{|B_{i,h'_n,-1} - A_{i,h'_n,-1}| > \epsilon\} \\
&\leq \sum_{n=2}^{\infty} n \max_{1\leq i\leq n} 2\exp\{-c_0\epsilon^2(\ln n)^2\} < \infty,
\end{aligned}$$

which implies, by the Borel-Cantelli lemma,

$$\max_{1\leq i\leq n} |B_{i,h'_n,-1} - A_{i,h'_n,-1}| \to 0, \quad \text{a.s.} \tag{A.4}$$

Also, by the definition of $A_{i,h'_n,-1}$ and conditions (II) and (III), we have

$$\begin{aligned}
\max_{1\leq i\leq n} |A_{i,h'_n,-1}| &\leq \max_{1\leq i\leq n} \max_{1\leq j\leq h'_n} |g(x_i) - g(x_{i(j)})| \\
&\leq \max_{1\leq i\leq n} \max_{1\leq j\leq h'_n} L|x_i - x_{i(j)}|^\alpha \\
&= \max_{1\leq i\leq n} L|x_i - x_{i(h'_n)}|^\alpha \leq L\left(c_2 n^{-1} h'_n \ln n\right)^\alpha \\
&= Lc_2^\alpha \left(n^{-1}(\ln n)^3\right)^\alpha \to 0.
\end{aligned} \tag{A.5}$$

(A.2)-(A.5) show that $\mathrm{cv}(h'_n) = \mathrm{cv}([(\ln n)^2]) \to m(|e_1|)$, a.s., which completes the proof.

Lemma A.3. *Under conditions* (I)-(IV),

$$h^*_n \to \infty, \quad \text{a.s.}, n \to \infty. \tag{A.6}$$

Proof. Set $A = \{\omega : h^*_n(\omega) \to \infty\}$. Suppose that, on the contrary, (A.6) does not hold, i.e., $P(A^c) > 0$, where A^c is the complement of A. Thus, for every $\omega \in A^c$, there exists a monotonic increasing index sequence $n_k(\omega)$ satisfying $h^*_{n_k(\omega)}(\omega) \le M(\omega)$ for a certain constant M, i.e., $\{h^*_{n_k(\omega)}(\omega), k \ge 1\}$ is a bounded subsequence. Without loss of generality, we assume that $h^*_{n_k(\omega)}(\omega) \to M(\omega)$, where $M(\omega)$ is a function with range $\{2, 3, \cdots\}$ and domain A^c. It is easy to verify

$$\begin{aligned}\mathrm{cv}(h^*_{n_k(\omega)}(\omega)) &= m\{|Y_i - m(Y_{i(2)}, \ldots, Y_{i(h^*_{n_k(\omega)}(\omega))})|, i = 1, \ldots, n\} \\ &\to m(|e_i - m(e_{i(2)}, \ldots, e_{i(M)})|)|_{M=M(\omega)} \quad \text{a.s., on } A^c.\end{aligned}$$

In view of Lemma A.2, we know that there exists an $h'_n \in H_n (n \ge 2)$ such that $\mathrm{cv}(h'_n) \to m(|e_1|)$ a.s. From the definition of $h^*_{n_k}$, we have $\mathrm{cv}(h'_{n_k(\omega)}(\omega)) \ge \mathrm{cv}(h^*_{n_k(\omega)}(\omega))$, which shows that

$$m(|e_1|) \ge m(|e_1 - m(e_2, \ldots, e_h)|)|_{h=M(\omega)}, \text{ on } A^c. \tag{A.7}$$

Using condition (I) and Andersons's lemma (Anderson (1955) or Ibragimov and Has'minskii (1981), p.155), we obtain $m(|e_1|) < m(|e_1 - m(e_2, \ldots, e_h)|)$, for all $h \ge 2$, which contradicts with (A.7). Therefore (A.6) holds.

Lemma A.3 only shows that h^*_n tends to infinity and does not reflect the rate of the convergence. Intuitively, if $g(x) \not\equiv$ constant, the number of nearest neighbors h should be neither too small nor too large. If h is too small, the influence of the random error plays the main role in the estimator $\widetilde{g}_{n,h}(x)$; on the contrary, if h is too large, the deviation of $g(x)$ at the neighbor of x will have influence on the value of the estimator. Theorem 1 gives the lower bound of h^*_n.

Lemma A.4. *Under condition* (I), *there exists a constant* $c_3 > 0$ *such that*

$$P\{|e_1 - m(e_2, \ldots, e_h)| \le m(|e_1|)\} \le 1/2 - c_3/h \quad \textit{for all } h \textit{ sufficiently large.}$$

Proof. Let $\xi_h = 2f(0)\sqrt{h}m(e_2, \ldots, e_h)$ and Φ_h be the distribution function of ξ_h. By the Central Limit Theorem (CLT) of a median (Serfling (1980), p.77, Corollary A), we have $\Phi_h \to \Phi$, as $h \to \infty$, where Φ is the distribution function

of standard normal random variable. The following inequalities complete the proof,

$$\begin{aligned}
&1/2 - P\{|e_1 - m(e_2,\ldots,e_h)| \le m(|e_1|)\} \\
&= \int_{-\infty}^{\infty} \left(P\{|e_1| \le m(|e_1|)\} - P\left\{\left|e_1 - x/(2f(0)\sqrt{h})\right| \le m(|e_1|)\right\}\right) d\Phi_h(x) \\
&\ge c_1 \int_{0\le x/(2f(0)\sqrt{h})\le\delta} d\Phi_h \int_{m-x/(2f(0)\sqrt{h})}^{m} \frac{x}{2f(0)\sqrt{h}} dt \\
&= \frac{c_1}{4f^2(0)} \frac{1}{h} \int_0^{2\delta f(0)\sqrt{h}} x^2 d\Phi_h(x) \ge \frac{c_1}{4f^2(0)} \frac{1}{h} \int_0^{2\delta f(0)} x^2 d\Phi_h(x) \\
&\ge \frac{c_1}{8f^2(0)} \frac{1}{h} \int_0^{2\delta f(0)} x^2 d\Phi(x), \quad \text{for large enough } h.
\end{aligned}$$

Lemma A.5. *Under conditions* (I)-(IV), *there exist* $\overline{h}_n \in H_n$ *and* $K(n) > 0$ *such that*

$$\sum_{n=2}^{\infty} P\{cv(\overline{h}_n) \ge m(|e_1|) + K(n)\} < \infty. \tag{A.8}$$

Proof. Observe that for every $h \in H_n$, we have

$$\begin{aligned}
\mathrm{cv}(h) &= m\{|e_i - B_{i,h,-1}|, 1 \le i \le n\} \\
&\le m\{|e_i|, 1 \le i \le n\} + \max_{1\le i\le n} |A_{i,h,-1} - B_{i,h,-1}| + \max_{1\le i\le n} |A_{i,h,-1}|.
\end{aligned} \tag{A.9}$$

By the definition of $A_{i,h,-1}$, we have

$$\begin{aligned}
\max_{1\le i\le n} |A_{i,h,-1}| &\le \max_{1\le i\le n} \max_{2\le j\le h} |g(x_{i(1)}) - g(x_{i(j)})| \\
&\le \max_{1\le i\le n} \max_{2\le j\le h} L|x_{i(1)} - x_{i(j)}|^{\alpha} \le L c_2^{\alpha} \left(n^{-1} h \ln n\right)^{\alpha}.
\end{aligned} \tag{A.10}$$

Also, by lemma A.1, for every $b_n > 0$,

$$\begin{aligned}
&\sum_{n=2}^{\infty} P\{\max_{1\le i\le n} |A_{i,h,-1} - B_{i,h,-1}| \ge b_n\} \\
&\le \sum_{n=2}^{\infty} n \max_{1\le i\le n} P\{|A_{i,h,-1} - B_{i,h,-1}| \ge b_n\} \le \sum_{n=2}^{\infty} 2n \exp\{-c_0 b_n^2 h\}.
\end{aligned} \tag{A.11}$$

To insure the convergence of the series on both sides of (A.11), we select h, b_n such that $n\exp\{-c_0 b_n^2 h\} = n^{-1-\beta}$ for all n sufficiently large, where β is a positive constant. Taking $b_n = \sqrt{(2+\beta)c_0^{-1}h^{-1}\ln n}$, we obtain from (A.11) that for almost all sample series, and large enough n,

$$\max_{1\le i\le n} |A_{i,h,-1} - B_{i,h,-1}| \le b_n = \sqrt{(2+\beta)c_0^{-1}h^{-1}\ln n}. \tag{A.12}$$

By the property of the median (Serfling (1980), p.96, Lemma B), we obtain

$$|m\{|e_i|, 1 \le i \le n\} - m(|e_1|)| \le 2/(f(m(|e_1|)))\sqrt{n^{-1}\ln n}$$

for all n sufficiently large. Therefore

$$m\{|e_i|, 1 \le i \le n\} \le m(|e_1|) + 2/(f(m(|e_1|)))\sqrt{n^{-1}\ln n}. \tag{A.13}$$

In view of (A.9)-(A.13), for all sufficiently large n, we have

$$\begin{aligned}\mathrm{cv}(h) \le\ & m(|e_1|) + 2/(f(m(|e_1|)))\sqrt{n^{-1}\ln n} + Lc_2^{\alpha}(n^{-1}\ln n)^{\alpha}h^{\alpha} \\ & + \sqrt{(2+\beta)/c_0 \ln n}/\sqrt{h}.\end{aligned} \tag{A.14}$$

When

$$h = \overline{h}_n = ((2+\beta)/(4c_0))^{1/(2\alpha+1)} (L\alpha c_2^{\alpha})^{-2/(1+2\alpha)} n^{2\alpha/(1+2\alpha)} (\ln n)^{(1-2\alpha)/(1+2\alpha)}, \tag{A.15}$$

the right hand side of (A.14) reaches the minimum value

$$m(|e_1|) + 2/(f(m(|e_1|)))\sqrt{n^{-1}\ln n} + c_4 n^{-\alpha/(1+2\alpha)} (\ln n)^{2\alpha/(1+2\alpha)},$$

where

$$c_4 = c_4(\alpha, \beta, c_0, c_2) = (2+\alpha^{-1})(L\alpha)^{1/(1+2\alpha)}((2c_2+\beta c_2)/(4c_0))^{2\alpha/(1+2\alpha)}.$$

Therefore

$$\begin{aligned}\mathrm{cv}(\overline{h}_n) &\le m(|e_1|) + 2/(f(m(|e_1|)))\sqrt{n^{-1}\ln n} + c_4 n^{-\alpha/(1+2\alpha)} (\ln n)^{2\alpha/(1+2\alpha)} \\ &= m(|e_1|) + K(n), \quad \text{for all } n \text{ sufficiently large,}\end{aligned}$$

where

$$K(n) = 2/(f(m(|e_1|)))\sqrt{n^{-1}\ln n} + c_4 n^{-\alpha/(1+2\alpha)} (\ln n)^{2\alpha/(1+2\alpha)}. \tag{A.16}$$

In fact, by checking (A.10), (A.11) and (A.13) (see Serfling (1980), p.96), we obtain the following inequalities

$$\sum_{n=2}^{\infty} P\Big\{m(|e_i|, 1 \le i \le n) \ge m(|e_1|) + 2/(f(m(|e_1|)))\sqrt{n^{-1}\ln n}\Big\} < \infty, \tag{A.17}$$

$$\sum_{n=2}^{\infty} P\Big\{ \max_{1\le i\le n} |A_{i,\overline{h}_n,-1} - B_{i,\overline{h}_n,-1}| + \max_{1\le i\le n} |A_{i,\overline{h}_n,-1}| \ge c_4 n^{-\frac{\alpha}{1+2\alpha}} (\ln n)^{\frac{2\alpha}{1+2\alpha}} \Big\} < \infty. \tag{A.18}$$

From (A.9), (A.17) and (A.18),

$$\sum_{n=2}^{\infty} P\Big\{\mathrm{cv}(\overline{h}_n) \geq m(|e_1|) + K(n)\Big\} < \infty,$$

where $\overline{h}_n$ and $K(n)$ are defined by (A.15) and (A.16) respectively.

Proof of Theorem 2.1. It suffices to verify

$$\sum_{n=2}^{\infty} P\{h_n^* \leq h_n\} < \infty.$$

Put $S_n = \{\mathrm{cv}(\overline{h}_n) \geq m(|e_1|) + K(n)\}$, where $\overline{h}_n$ and $K(n)$ are determined by (A.15) and (A.16) respectively. Note that

$$\sum_{n=2}^{\infty} P\{h_n^* \leq h_n\} \leq \sum_{n=2}^{\infty} P\{h_n^* \leq h_n, S_n^c\} + \sum_{n=2}^{\infty} P\{h_n^* \leq h_n, S_n\}.$$

By Lemma A.5, we need only show that $\sum_{n=2}^{\infty} P\{h_n^* \leq h_n, S_n^c\} < \infty$. By the definition of h_n^*,

$$\begin{aligned}
\sum_{n=2}^{\infty} P\{h_n^* \leq h_n, S_n^c\} &\leq \sum_{n=2}^{\infty} P\{\min_{2\leq h\leq h_n} \mathrm{cv}(h) \leq \mathrm{cv}(\overline{h}_n), S_n^c\} \\
&\leq \sum_{n=2}^{\infty} \sum_{h=2}^{h_n} P\{\mathrm{cv}(h) \leq \mathrm{cv}(\overline{h}_n), S_n^c\} \\
&= \sum_{n=2}^{\infty} \sum_{h=2}^{h_n} P\{m(|Y_i - \widetilde{g}_{n,h,-1}(x_i)|, i = 1, \ldots, n) \leq \mathrm{cv}(\overline{h}_n), S_n^c\} \\
&= \sum_{n=2}^{\infty} \sum_{h=2}^{h_n} P\Big\{\frac{1}{n}\sum_{i=1}^{n} I(|e_i - B_{i,h,-1}| \leq \mathrm{cv}(\overline{h}_n)) \geq \frac{1}{2}, S_n^c\Big\}, \quad \text{(A.19)}
\end{aligned}$$

where $I\{A\}$ denotes the indicator of A. In the following we will transform $\frac{1}{n}\sum_{i=1}^{n} I\{|e_i - B_{i,h,-1}| \leq \mathrm{cv}(\overline{h}_n)\}$ into the form of a sum of independent random variables. Set

$$D_j = \{i : x_i = x_{(l(2h+1)+j)}, 1 \leq i \leq n, 0 \leq l \leq [n/(2h+1)]\},\ j = 0, 1, \ldots, 2h,$$

where $x_{(i)}, i = 1, \ldots, n$, are order statistics of $x_i, i = 1, \ldots, n$. The index set $\{1, \ldots, n\}$ becomes the union of the disjoint subsets $D_j, j = 0,\ 1, 2, \ldots, 2h$, and $\{|e_i - B_{i,h,-1}|, i \in D_j\}$ is a set of independent random variables for each $j \in \{0, 1, 2, \ldots, 2h\}$. From now on we treat $n/(2h+1)$ as an integer to avoid unnecessary complications.

Recalling (A.19), we obtain

$$\begin{aligned}
&\sum_{n=2}^{\infty} P\{h_n^* \le h_n, S_n^c\} \le \sum_{n=2}^{\infty}\sum_{h=2}^{h_n} P\Big\{\sum_{j=0}^{2h}\sum_{i\in D_j} I\{|e_i - B_{i,h,-1}| \le \mathrm{cv}(\overline{h}_n)\} \ge \frac{n}{2}, S_n^c\Big\} \\
&\le \sum_{n=2}^{\infty}\sum_{h=2}^{h_n}\sum_{j=0}^{2h} P\Big\{\sum_{i\in D_j} I\{|e_i - B_{i,h,-1}| \le \mathrm{cv}(\overline{h}_n)\} \ge \frac{n}{4h+2}, S_n^c\Big\}. \\
&\le \sum_{n=2}^{\infty}\sum_{h=2}^{h_n}\sum_{j=0}^{2h} P\Big\{\sum_{i\in D_j} I\{|e_i - B_{i,h,-1}| \le m(|e_1|) + K(n)\} \ge \frac{n}{4h+2}\Big\} \\
&= \sum_{n=2}^{\infty}\sum_{h=2}^{h_n}\sum_{j=0}^{2h} P\Big\{\frac{2h+1}{n}\sum_{i\in D_j}\Big\{I\{|e_i - B_{i,h,-1}| \le m(|e_1|) + K(n)\} \\
&\qquad\qquad -P\{|e_i - B_{i,h,-1}| \le m(|e_1|) + K(n)\}\Big\} \\
&\qquad\qquad \ge \frac{1}{2} - \frac{2h+1}{n}\sum_{i\in D_j} P\{|e_i - B_{i,h,-1}| \le m(|e_1|) + K(n)\}\Big\} \\
&\le \sum_{n=2}^{\infty}\sum_{h=2}^{h_n}\sum_{j=0}^{2h} \exp\Big\{-\frac{2n\lambda_h^2}{2h+1}\Big\} \text{(by Hoeffding's inequality (1963),)} \qquad \text{(A.20)}
\end{aligned}$$

where

$$\lambda_h = 1/2 - (2h+1)/n \sum_{i\in D_j} P\{|e_i - B_{i,h,-1}| \le m(|e_1|) + K(n)\}.$$

In the above reasoning, $\lambda_h > 0$ for $2 \le h \le h_n$ is required for the Hoeffding inequality. In the following we deal with λ_h. Note that

$$\begin{aligned}
&P\{|e_i - B_{i,h,-1}| \le m(|e_1|) + K(n)\} \\
&\le P\{|e_i - m(e_{i(2)}, \ldots, e_{i(h)})| - |m(e_{i(2)}, \ldots, e_{i(h)}) - B_{i,h,-1}| \le m(|e_1|) + K(n)\} \\
&\le P\Big\{|e_i - m(e_{i(2)}, \ldots, e_{i(h)})| \le m(|e_1|) + K(n) + \max_{2\le j\le h} |a_{i(j)}|\Big\} \\
&\le P\Big\{|e_i - m(e_{i(2)}, \ldots, e_{i(h)})| \le m(|e_1|) + K(n) + \max_{1\le i\le n} L|x_i - x_{i(h)}|^\alpha\Big\} \\
&\le P\{|e_i - m(e_{i(2)}, \ldots, e_{i(h)})| < m(|e_1|) + K(n) + L(c_2 n^{-1} h \ln n)^\alpha\} \\
&= P\{|e_i - m(e_{i(2)}, \ldots, e_{i(h)})| < m(|e_1|)\} \\
&\quad + P\{m(|e_1|) \le |e_i - m(e_{i(2)}, \ldots, e_{i(h)})| \le m(|e_1|) + K(n) + L(c_2 n^{-1} h \ln n)^\alpha\} \\
&\le P\{|e_1 - m(e_2, \ldots, e_h)| < m(|e_1|)\} + c_6(K(n) + L(c_2 n^{-1} h \ln n)^\alpha) \\
&\le 1/2 - c_3/h + c_6(K(n) + L(c_2 n^{-1} \ln n)^\alpha h^\alpha), \quad \text{(by Lemma A.4)}
\end{aligned}$$

where c_6 is a positive constant depending only on $f(x)$, the density function of e_1. Therefore,

$$\begin{aligned}\lambda_h &\geq c_3/h - c_6(K(n) + L(c_2 n^{-1}\ln n)^{\alpha} h^{\alpha}) \\ &= c_3/h - c_6\Big(\frac{2\sqrt{n^{-1}\ln n}}{f(m(|e_1|))} + c_4 n^{-\alpha/(1+2\alpha)}(\ln n)^{\frac{2\alpha}{1+2\alpha}} + L(c_2 n^{-1} h \ln n)^{\alpha}\Big). \quad \text{(A.21)}\end{aligned}$$

Note that the right of (A.21) is a monotonically decreasing function of h. It is easy to show that for $h_n = [\mu(\ln n)^2/n]^{-\alpha/(1+2\alpha)}$,

$$\lambda_h \geq c_3/h_n - 3c_6 c_4 n^{-\alpha/(1+2\alpha)}(\ln n)^{2\alpha/(1+2\alpha)} > 0, \ 2 \leq h \leq h_n$$

holds for large n and small μ, which shows that by (A.20)

$$\begin{aligned}&\sum_{n=2}^{\infty} P\{h_n^* \leq h_n, S_n^c\} \\ &\leq \sum_{n=2}^{\infty}\sum_{h=2}^{h_n} 2h_n \exp\{-2n\lambda_n^2/(2h_n+1)\} \leq \sum_{n=2}^{\infty} 2h_n^2 \exp\{-2n\lambda_{h_n}^2/(2h_n+1)\} \\ &\leq \sum_{n=2}^{\infty} 2h_n^2 \exp\Big\{ - nh_n^{-1}\Big(c_3/h_n - c_6(2/(f(m(|e_1|)))\sqrt{n^{-1}\ln n} \\ &\qquad\qquad + c_4 n^{-\frac{\alpha}{1+2\alpha}}(\ln n)^{\frac{2\alpha}{1+2\alpha}} + L(c_2 n^{-1}\ln n)^{\alpha} h_n^{\alpha})\Big)^2\Big\} < \infty.\end{aligned}$$

By the Borel-Cantelli lemma and Lemma A.5, we know that $h_n^* \geq h_n = \mu(n(\ln n)^{-2})^{\alpha/(1+2\alpha)}$ a.s., which completes the proof of Theorem 1.

Proof of Theorem 2.2. Note that

$$\begin{aligned}\mathrm{ME}(h_n^*) &= \max_{1\leq i\leq n} |g(x_i) - \widetilde{g}_{n,h_n^*}(x_i)| = \max_{1\leq i\leq n} |B_{i,h_n^*}| \\ &\leq \max_{1\leq i\leq n} |B_{i,h_n^*} - A_{i,h_n^*}| + \max_{1\leq i\leq n} |A_{i,h_n^*}| \qquad \text{(A.22)}\end{aligned}$$

From Theorem 2.1 and Lemma A.1, for all $\epsilon > 0$,

$$\begin{aligned}&\sum_{n=2}^{\infty} P\Big\{ \max_{1\leq i\leq n} |B_{i,h_n^*} - A_{i,h_n^*}| \geq \epsilon\Big\} \\ &\leq \sum_{n=2}^{\infty} P\Big\{ \max_{1\leq i\leq n} |B_{i,h_n^*} - A_{i,h_n^*}| \geq \epsilon, h_n^* \geq h_n\Big\} + \sum_{n=2}^{\infty} P\{h_n^* \leq h_n\} \\ &\leq \sum_{n=2}^{\infty} n \max_{1\leq i\leq n} P\{|B_{i,h_n^*} - A_{i,h_n^*}| \geq \epsilon, h_n^* \geq h_n\} + \sum_{n=2}^{\infty} P\{h_n^* \leq h_n\} \\ &= \sum_{n=2}^{\infty} n \max_{1\leq i\leq n} \sum_{h=h_n}^{\mu_n} P\{|B_{i,h_n^*} - A_{i,h_n^*}| \geq \epsilon, h_n^* = h\} + \sum_{n=2}^{\infty} P\{h_n^* \leq h_n\}\end{aligned}$$

$$\le \sum_{n=2}^{\infty} n \max_{1\le i\le n} \sum_{h=h_n}^{\mu_n} P\{|B_{i,h} - A_{i,h}| \ge \epsilon\} + \sum_{n=2}^{\infty} P\{h_n^* \le h_n\}$$
$$\le \sum_{n=2}^{\infty} n \max_{1\le i\le n} \sum_{h=h_n}^{\mu_n} 2\exp\{-c_0 h\epsilon^2\} + \sum_{n=2}^{\infty} P\{h_n^* \le h_n\}$$
$$\le \sum_{n=2}^{\infty} 2n(\mu_n - h_n)\exp\{-c_0 h_n \epsilon^2\} + \sum_{n=2}^{\infty} P\{h_n^* \le h_n\}$$
$$\le \sum_{n=2}^{\infty} 2n^2 \exp\Big\{ - c_0\epsilon^2 \mu (n(\ln n)^{-2})^{\frac{\alpha}{1+2\alpha}}\Big\} + \sum_{n=2}^{\infty} P\{h_n^* \le h_n\} < \infty.$$

By the Borel-Cantelli lemma,

$$\max_{1\le i\le n} |B_{i,h_n^*} - A_{i,h_n^*}| \to 0 \quad \text{a.s.} \tag{A.23}$$

Also, by the definition of $A_{i,h}$, for all $h \in H_n, n \ge 2$,

$$\max_{1\le i\le n} |A_{i,h}| \le \max_{1\le i\le n} \max_{1\le j\le h} |g(x_{i(1)}) - g(x_{i(j)})| \le \max_{1\le i\le n} \max_{1\le j\le h} L|x_i - x_{i(j)}|^{\alpha}$$
$$\le \max_{1\le i\le n} L|x_i - x_{i(h)}|^{\alpha} \le L c_2^{\alpha} \left(n^{-1}\ln n\right)^{\alpha} h^{\alpha},$$

Therefore,

$$\max_{1\le i\le n} |A_{i,h_n^*}| \le L c_2^{\alpha} (n^{-1}\ln n)^{\alpha} (h_n^*)^{\alpha} \le L c_2^{\alpha} (n^{-1}\ln n)^{\alpha} (n/(b_n \ln n))^{\alpha}$$
$$= L c_2^{\alpha} b_n^{-\alpha} \to 0. \tag{A.24}$$

From (A.22)$\sim$(A.24), we obtain $\mathrm{ME}(h_n^*) \to 0$ a.s., which completes the proof of Theorem 2.2.

References

Allen, D. M. (1974). The relationship between variables selection and data augmentation and a method of prediction. *Technometrics* **16,** 125-127.

Anderson, T. W. (1955). The integral of a symmetric unimodal function. *Proc. Amer. Math. Soc.* **6**, 170-176.

Bhattacharya, P. K. and Gangopadhyay, A. K. (1989). Kernel and nearest neighbor estimation of a conditional quantile. *Ann. Statist.* **18**, 1400-1415.

Bhattacharya, P. K. and Mark, Y. P. (1987). Weak convergence of $k-NN$ density and regression estimators with varying k and applications. *Ann. Statist.* **15**, 976-994.

Cheng, S. H. (1984). On a problem concerning spacings. *Z. Wahrsch. Verw. Gebiete.* **66**, 245-258.

Chow, Y. S., Geman, S. and Wu, L. D. (1983). Consistent cross validated density estimation. *Ann. Statist.* **11**, 25-38.

Gangopadhyay, A. K. and Pranab K. Sen (1990). Bootstrap confidence intervals for conditional quantile functions. *Sankhya Ser. A* **52**, 346-363.

Hall, P. (1982). Cross validation in density estimation. *Biometrika* **69**, 383-390.

Härdle, W. and Chen, R. (1995). Nonparametric time series analysis, a selective review with examples. In *Proceedings of the* 50*th ISI Session* (Beijing, 1995) vol.1, 375-394.

Hoeffding, W. (1963). Probability inequalities for sums of bounded random variables. *J. Amer. Statist. Assoc.* **58**, 13-30.

Ibragimov, I. A. and Has'minskii, R. Z. (1981). *Statistical Estimation: Asymptotic Theory.* Springer-Verlag, New York.

Li, K. C. (1984). Consistency for cross-validated nearest neighbor estimates in nonparametric regression. *Ann. Statist.* **12**, 230-240.

Marron, J. S. (1987). What dose optimal bandwidth selection means for nonparametric regression estimation? In *Statistical Data Analysis Based on the* L_1*-Norm and Related Methods* (Edited by Y. Dodge), 379-391. North Holland, Amsterdam.

Marron, J. S. (1989). Automatic smoothing parameter selection: a survey. *Empirical Econom.* **13**, 187-208.

Serfling, R. J. (1980). *Approximation Theorems of Mathematical Statistics.* John Wiley, New York.

Stone, M. (1974). Cross-validatory choice and assessment of statistical predictions. *J. Roy. Statist. Soc. B* **36**, 111-147.

Stute, W. (1986). Conditional empirical processes. *Ann. Statist.* **14**, 638-647.

Wong, W. H. (1983). On the consistency of cross-validation in kernel nonparametric regression. *Ann. Statist.* **11**, 1136-1141.

Yang Y. (1996). *Nearest Neighbor Median Estimate and the Selection of Smoothing Parameter.* Ph. D. Dissertation, Dept. Probab & Statist., Peking University.

Yang, Y. and Zheng, Z. (1992). Asymptotic properties for cross-validated nearest neighbor median estimates in nonparametric regression: the L_1-view. In *Probability and Statistics* (Edited by Z. Jiang, S. Yan, P. Cheng and R. Wu), 242-257. World Scientific, Singapore.

Department of Probability and Statistics, Peking University, Beijing 100871.

E-mail: zgzheng@statms.stat.pku.edu.cn

E-mail: yyang@statms.stat.pku.edu.cn

(Received July 1995; accepted July 1997)