The Annals of Statistics
1984, Vol. 12, No. 1, 230–240

CONSISTENCY FOR CROSS-VALIDATED NEAREST NEIGHBOR ESTIMATES IN NONPARAMETRIC REGRESSION

BY KER-CHAU LI[1]

Purdue University

Under suitable conditions, we show that the cross-validated nearest neighbor estimates for the unknown smooth regression function in R^p is asymptotically consistent.

1. Introduction. Let p be a natual number and $\mathscr{X}$ be the closure of an open set in R^p. Consider the case that $\mathscr{X}$ is compact. Suppose n independent observations $y_1, y_2, \cdots, y_n$ are made at levels $\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n \in \mathscr{X}$. Write $\mathbf{y}_n = (y_1, \cdots, y_n)'$. Without loss of generality, assume that $\mathbf{x}_i \neq \mathbf{x}_j$ for $i \neq j$. Consider the model

$$y_i = f(\mathbf{x}_i) + \varepsilon_i, \quad i = 1, 2, \cdots, n, \tag{1.1}$$

where f is continuous on $\mathscr{X}$ and ε_i's are independent random variables with means 0 and variances σ_i^2, $i = 1, \cdots, n$. To estimate the unknown function f, many classes of estimators have been proposed, including the kernel method (Watson, 1964, Nadaraya, 1964, etc.), the nearest neighbor method (Fix and Hodges, 1951; Cover and Hart, 1967; Cover, 1968; Stone, 1977, etc.), and the spline method (particularly for $p = 1$, Reinsch, 1967; Wahba and Wold, 1975; Agarwal and Studden, 1980, etc.) Basically, these estimates are linear in the y_i's. Also, each estimate is associated with an index h (e.g., the bandwidth for the kernel estimate; the number of neighbors for the nearest neighbor estimate; the smoothing parameter for the smoothing spline). The choice of h turns out to be crucial in effectively estimating f. Most studies on the asymptotic aspect have been addressed to the case where h is deterministically chosen. However, for practical use, it is often preferable to have a data-driven h. One such practice is to select h by the cross-validation technique, whose consistency property will be investigated here for the nearest neighbor method.

Given $\mathbf{x}_1, \cdots, \mathbf{x}_n$ let $\mathbf{x}_{i(j)}$ denote the jth nearest neighbor of $\mathbf{x}_i$ in the sense that $\| \mathbf{x}_i - \mathbf{x}_{i(j)} \|$ is the jth smallest number among the n values $\| \mathbf{x}_i - \mathbf{x}_{i'} \|$, $i' = 1, 2, \cdots, n$ (ties can be broken in any systematic manner). Let $H_n = \{1, 2, \cdots, n\}$. For any $h \in H_n$, the h nearest neighbor estimate of $f(\mathbf{x}_i)$ is defined by $\sum_{j=1}^n W_{n,h}(j) y_{i(j)}$ with $W_{n,h}(\cdot)$ being a non-negative weight function satisfying certain conditions to be specified whenever needed. For our development it is easier to represent the estimate of $\mathbf{f}_n = (f(\mathbf{x}_1), f(\mathbf{x}_2), \cdots, f(\mathbf{x}_n))'$ by a matrix form of $\tilde{M}_n(h)\mathbf{y}_n$ where $\tilde{M}_n(h)$ denotes a suitable $n \times n$ matrix with rows being

Received December 1982; revised August 1983.

[1] Research supported in part by the National Science Foundation under grant No. MCS-82-00631.

AMS 1980 *subject classification.* Primary, 62G05.

Key words and phrases. Cross-validation, consistency, nearest neighbor estimates, nonparametric regression.

certain permutations of the vector $(W_{n,h}(1), W_{n,h}(2), \cdots, W_{n,h}(h), 0, \cdots, 0)$. Clearly $\tilde{M}_n(h)\mathbf{y}_n$ can also be used to predict the values of y observations to be made in the future at the same levels $\mathbf{x}_1, \cdots, \mathbf{x}_n$. To assess its prediction performance, a naive estimate seems to be $\| \mathbf{y}_n - \tilde{M}_n(h)\mathbf{y}_n \|^2$. This quantity tends to underestimate the true error since the same data have been used both to construct and to evaluate $\tilde{M}_n(h)\mathbf{y}_n$. Cross-validation circumvents this difficulty by removing each y_i from the data set used in its own prediction.

Precisely, to predict a future y observation at level $\mathbf{x}_i$, $1 \leq i \leq n$, we use only the data $y_1, y_2, \cdots, y_{i-1}, y_{i+1}, \cdots, y_n$; namely $\sum_{j=1}^{h} W_{n,h}(j) y_{i(j+1)}$. To put it in matrix form, we write $M_n(h)\mathbf{y}_n$. Thus $M_n(h)$ is an $n \times n$ matrix with zero diagonals and the $(i, i(j))$th element, $j \neq 1$, being $W_{n,h}(j-1)$. Now the cross-validated assessment of prediction error for $\tilde{M}_n(h)\mathbf{y}_n$ is $\| \mathbf{y}_n - M_n(h)\mathbf{y}_n \|^2$ and the cross-validated choice of $h \in H_n$, denoted by h_n^*, is the minimizer of

$$\inf_{h \in H_n} (1/n) \| (I_n - M_n(h))\mathbf{y}_n \|^2 \tag{1.2}$$

where I_n denotes the $n \times n$ identity matrix.

Although the motivation behind the cross-validation technique is easily understood (see, Allen, 1974; Stone, 1974; and Geisser, 1975), available theorems with regards to its statistical properties seem to be sparse. In classification problems, some interesting properties relating cross-validation with bootstrapping were obtained by Efron (1983). In density estimation, Chow, Geman and Wu (1983) and Hall (1982) established some asymptotic results for the cross-validated kernel estimates. In our nonparametric regression problem with kernel estimates, Wong (1983) proved the consistency in the case that $p = 1$ with the $\mathbf{x}_i$'s being equi-spaced in a bounded interval. For the spline smoothing, generalized cross-validation of Craven and Wahba (1979) was shown to possess a certain asymptotic efficiency property by Speckman (1982).

In this paper, we shall show that as $n \to \infty$,

$$(1/n) \| \mathbf{f}_n - \tilde{M}_n(h_n^*)\mathbf{y}_n \|^2 \to 0, \tag{1.3}$$

in probability for the nearest neighbor estimates (Section 2). Somewhat in line with Wong (1983), our proofs will consist in establishing the following four statements (hereafter, unless otherwise specified, any convergence involving random variables will be interpreted as the convergence in probability):

$$\sup_{h \in H_n} (1/n) | \langle (I - M_n(h))\mathbf{f}_n, \varepsilon_n \rangle | \to 0, \tag{S.1}$$

where $\varepsilon_n = (\varepsilon_1, \varepsilon_2, \cdots, \varepsilon_n)'$ and $\langle \cdot, \cdot \rangle$ denotes the inner product in R^n.

$$\sup_{h \in H_n} (1/n) | \langle M_n(h)\, \varepsilon_n, \varepsilon_n \rangle | \to 0. \tag{S.2}$$

(S.3) There exists a sequence $\{h_n\}$ such that

$$(1/n) \| \mathbf{f}_n - M_n(h_n)\mathbf{y}_n \|^2 \to 0.$$

$$(1/n) \| \tilde{M}_n(h_n^*)\mathbf{y}_n - M_n(h_n^*)\mathbf{y}_n \|^2 \to 0. \tag{S.4}$$

To see that the above four statements imply (1.3), observe that by (1.1),

$$(1/n)\|(I_n - M_n(h))\mathbf{y}_n\|^2$$
$$= (1/n)\|\mathbf{f}_n - M_n(h)\mathbf{y}_n\|^2 + (2/n)\langle(I_n - M_n(h))\mathbf{f}_n, \varepsilon_n\rangle$$
$$- (2/n)\langle M_n(h)\varepsilon_n, \varepsilon_n\rangle + (1/n)\|\varepsilon_n\|^2.$$

Thus by (S.1) – (S.3) and (1.2) we obtain

$$(1/n)\|\mathbf{f}_n - M_n(h_n^*)\mathbf{y}_n\|^2 \to 0 \tag{1.4}$$

which together with (S.4) implies (1.3) as desired.

The following two regularity conditions on the $\mathbf{x}$ sequence will be imposed:

(C.1) There exists a constant λ_1 such that for any $r > 0$, there exists an integer N_r such that for any $n \geq N_r$ and any closed ball $B(\mathbf{x}, r)$ with center $\mathbf{x} \in \mathscr{X}$ and radius r,

$$\#\{\mathbf{x}_i \mid \mathbf{x}_i \in B(\mathbf{x}, r), 1 \leq i \leq n\} \geq \lambda_1 n r^p.$$

(C.2) There exists a constant λ_2 such that

$$\#\{\mathbf{x}_i \mid \mathbf{x}_i \in S, 1 \leq i \leq n\} \leq \lambda_2 n \lambda(S),$$

for any n and Borel set S with Lebesque measure $\lambda(S)$.

(C.1) and (C.2) imply that $\mathbf{x}$ sequence gets dense in $\mathscr{X}$ in a uniform fashion. When $\mathbf{x}_i$'s are the realizations of i.i.d. random vectors with a common density bounded away from both 0 and ∞ on $\mathscr{X}$, (C.1) and (C.2) are satisfied with probability one. For such random $\mathbf{x}$ cases, note that the consistency property (1.3) is conditioned on the $\mathbf{x}$ values.

We shall also assume the following moment condition on the random errors:

(C.3) The fourth moments of ε_i's are no greater than μ^4, with a finite constant $\mu > 0$, and

$$\liminf_{i\to\infty} \sigma_i^2 > \sigma^2 > 0. \tag{1.5}$$

(1.5) is to avoid the trivial case that σ_i^2 are eventually 0, while the finite fourth moment assumption is made to obtain a simple proof of (S.2).

2. Consistency results. Assume the following conditions on the weight functions:

(C.4) $\sum_{i=1}^{h} W_{n,h}(i) = 1$, and $W_{n,h}(i) = 0$ for $i > h$.

(C.5) $W_{n,h}(i)$ is nonincreasing in i.

(C.6) There exists a sequence $\{h_n\}$ such that

(C.6.1) $h_n/n \to 0$,

and

(C.6.2) $W_{n,h_n}(1) \to 0$

as $n \to \infty$.

We now show that under (C.1)–(C.6), (S.1)–(S.3) hold.

PROOF OF (S.1). Given $\delta > 0$, we shall show $P\{\sup_{h\in H_n} (1/n)|\langle (I_n - M_n(h))\mathbf{f}_n, \varepsilon_n\rangle| > \delta\} \to 0$.

First, since f is continuous on $\mathscr{X}$ (this implies the uniform continuity), there exists $\ell > 0$ such that for any $\mathbf{x}$ and $\mathbf{x}' \in \mathscr{X}$ with $\|\mathbf{x} - \mathbf{x}'\| \le \ell$, $\|f(\mathbf{x}) - f(\mathbf{x}')\| \le (1/6)\cdot(\delta/\mu)$. Let k_n be the largest integer in H_n such that $\sup_{1\le i\le n}\|\mathbf{x}_i - \mathbf{x}_{i(k_n)}\| \le \ell$. By (C.1), we have

$$k_n \ge \lambda_1 n\ell^p, \quad \text{for any} \quad n > N_\ell. \tag{2.1}$$

Then, we have

$$\sup_{1\le h\le k_n-1} \frac{1}{n}|\langle (I_n - M_n(h))\mathbf{f}_n, \varepsilon_n\rangle| \le \frac{1}{6}\cdot\frac{\delta}{\mu}\cdot\frac{1}{n}\sum_{i=1}^n |\varepsilon_i|. \tag{2.2}$$

To see this, observe that for $1 \le h \le k_n - 1$ the absolute value of each coordinate of $(I_n - M_n(h))\mathbf{f}_n$ is no greater than $\delta/6\mu$ because of the definition of $M_n(h)$, (C.4) and the definition of k_n. Clearly, (2.2) implies that

$$P\{\sup_{1\le h\le k_n-1} (1/n)\ |\langle (I_n - M_n(h))\mathbf{f}_n, \varepsilon_n\rangle| > \delta\} \to 0.$$

Thus it remains to show that

$$P\{\sup_{k_n\le h\le n} (1/n)|\langle (I_n - M_n(h))\mathbf{f}_n, \varepsilon_n\rangle| > \delta\} \to 0. \tag{2.3}$$

Partition the space $\mathscr{X}$ into $\mathscr{X}_1, \mathscr{X}_2, \cdots, \mathscr{X}_m$ for some fixed number m such that the diameter of each $\mathscr{X}_j$ is no greater than α, a number to be set suitably later on (see (2.5) below). Define $H_n^j = \{i \mid \mathbf{x}_i \in \mathscr{X}_j\}$ for $j = 1, \cdots, m$. Denote the ith coordinate of $M_n(h)\mathbf{f}_n$ by $[M_n(h)\mathbf{f}_n]_i$. We claim that

(2.4) d can be chosen so that for any $n \ge N_\alpha$, any h with $k_n \le h \le n$, and any $i, i' \in H_n^j$,

$$|[M_n(h)\mathbf{f}_n]_i - [M_n(h)\mathbf{f}_n]_{i'}| \le \frac{3}{4}\cdot\frac{\delta}{\mu}.$$

Assuming the validity of (2.4) and letting $\hat{f}_n(j)$ denote $[M_n(h)\mathbf{f}_n]_i$ with i being the smallest integer in H_n^j, we then have

$$\begin{aligned}
&\sup_{k_n\le h\le n} \frac{1}{n}\ |\ \langle (I_n - M_n(h))\mathbf{f}_n, \varepsilon_n\rangle\ | \\
&\quad\le \frac{1}{n}\ |\ \langle \mathbf{f}_n, \varepsilon_n\rangle\ | + \sup_{k_n\le h\le n}\frac{1}{n}\ |\ \langle M_n(h)\mathbf{f}_n, \varepsilon_n\rangle\ | \\
&\quad\le \frac{1}{n}\ |\ \langle \mathbf{f}_n, \varepsilon_n\rangle\ | + \sup_{k_n\le h\le n}\left\{\frac{1}{n}\sum_{j=1}^m |\ \hat{f}_n(j)\ \sum_{i\in H_n^j}\varepsilon_i|\ \right\} + \frac{3}{4}\cdot\frac{\delta}{\mu}\cdot\frac{1}{n}\sum_{i=1}^n |\varepsilon_i| \\
&\quad\le \frac{1}{n}\ |\ \langle \mathbf{f}_n, \varepsilon_n\rangle\ | + |f|_\infty\cdot\sum_{j=1}^m \frac{\#(H_n^j)}{n}\left|\ \frac{1}{\#(H_n^j)}\sum_{i\in H_n^j}\varepsilon_i\ \right| \\
&\qquad + \frac{3}{4}\frac{\delta}{\mu}\cdot\frac{1}{n}\sum_{i=1}^n |\varepsilon_i|
\end{aligned}$$

where $|f|_\infty$ denotes the supremum of $|f(\mathbf{x})|$ over $\mathbf{x} \in \mathscr{X}$. In the last expression, it is clear that the first and the second terms tend to 0 in probability, while the third term is asymptotically no greater than (3/4) δ. Thus (2.3) is established.

Therefore to complete the proof of (S.1), it remains to verify (2.4). The following lemma will be useful. Write $\ell[i]_n = k$ when $i(k) = \ell$; i.e., when $\mathbf{x}_\ell$ is the kth nearest neighbor of $\mathbf{x}_i$.

LEMMA 2.1. *There exists a constant a such that* $|\ell[i]_n - \ell[j]_n| \le a \cdot n \cdot \|\mathbf{x}_i - \mathbf{x}_j\|$ *for any* $1 \le i, j, \ell \le n$.

The proof of this lemma will be given in the Appendix. Now observe that for any $k_n \le h \le n$ and $i, i' \in H_n^j$,

$$
\begin{aligned}
&|[M_n(h)\mathbf{f}_n]_i - [M_n(h)\mathbf{f}_n]_{i'}| \\
&\qquad = |\textstyle\sum_{k=1}^{h} W_{n,h}(k) f(\mathbf{x}_{i(k+1)}) - \sum_{k=1}^{h} W_{n,h}(k) f(\mathbf{x}_{i'(k+1)})| \\
&\qquad \le |\textstyle\sum_{k=1}^{k_n-1} W_{n,h}(k)(f(\mathbf{x}_{i(k+1)}) - f(\mathbf{x}_{i'(k+1)}))| + \sum_{\ell \in A_n} |W_{n,h}(\ell[i]_n) \\
&\qquad\qquad - W_{n,h}(\ell[i']_n)| \cdot |f(\mathbf{x}_\ell)| + \textstyle\sum_{\ell \in B_n} W_{n,h}(\ell[i]_n)|f(\mathbf{x}_\ell)| \\
&\qquad\qquad + \textstyle\sum_{\ell \in C_n} W_{n,h}(\ell[i']_n)|f(\mathbf{x}_\ell)|,
\end{aligned}
$$

where $A_n = \{i(k): k_n \le k \le n\} \cap \{i'(k): k_n \le k \le n\}$, $B_n = \{i(k): k_n \le k \le n\} - A_n$, and $C_n = \{i'(k): k_n \le k \le n\} - A_n$. In the last expression the first term will be no greater than (½) $\cdot$ (δ/μ) supposing than $d \le \ell$. The second term is no greater than

$$
\begin{aligned}
&|f|_\infty \cdot \textstyle\sum_{\ell \in A_n} |W_{n,h}(\ell[i]_n) - W_{n,h}(\ell[i']_n)| \\
&\qquad \le |f|_\infty \cdot \textstyle\sum_{\ell \in A_n} \{(W_{n,h}(\ell[i]_n) - W_{n,h}(\ell[i]_n + ad n)) \\
&\qquad\qquad + (W_{n,h}(\ell[i']_n) - W_{n,h}(\ell[i']_n + ad n))\} \\
&\qquad\qquad\qquad\qquad \text{(by Lemma 2.1 and (C.5))} \\
&\qquad \le 2|f|_\infty \cdot \textstyle\sum_{k=k_n}^{n} (W_{n,h}(k) - W_{n,h}(k + ad n)) \\
&\qquad = 2|f|_\infty \cdot \textstyle\sum_{k=k_n}^{k_n + ad n - 1} W_{n,h}(k) \\
&\qquad \le 2|f|_\infty \cdot \textstyle\sum_{k=\lambda_1 n \ell^p}^{(\lambda_1 \ell^p + ad)n - 1} W_{n,h}(k) \quad \text{(by (2.1))} \\
&\qquad \le 2|f|_\infty \cdot \frac{ad\, n}{(\lambda_1 \ell^p + ad)n} \quad \text{(by (C.5))} \\
&\qquad \le 2|f_\infty| \cdot \frac{ad}{\lambda_1 \ell^p}.
\end{aligned}
$$

(Note that in the above expressions, the term "$ad n$" should be interpreted as the

largest integer $\leq \alpha d n$, whenever necessary). Furthermore, by Lemma 2.1,

$$\sum_{\ell \in B_n} W_{n,h}(\ell[i]_n)\,|f(\mathbf{x}_\ell)| \leq |f|_\infty \sum_{k=k_n-\alpha d n}^{k_n} W_{n,h}(k)$$

$$\leq |f|_\infty \cdot \frac{\alpha d}{\lambda_1 \ell^p} \quad \text{(by (2.1) and (C.5))}.$$

We obtain the same bound for $\sum_{\ell \in C_n} W_{n,h}(\ell[i']_n)\,|f(\mathbf{x}_\ell)|$ in a similar way. Therefore, we have

$$|[M_n(h)\mathbf{f}_n]_i - [M_n(h)\mathbf{f}_n]_{i'}| \leq \frac{1}{2}\frac{\delta}{\mu} + 4\,|f|_\infty \frac{\alpha d}{\lambda_1 \ell^p}.$$

Thus (2.4) holds for

$$d = \min\left\{\ell, \frac{1}{16} \cdot \frac{\lambda_1 \ell^p \delta}{\alpha \mu\,|f|_\infty}\right\}. \tag{2.5}$$

The proof of (S.1) is now complete. We turn to:

PROOF OF (S.2). First, observe that

$$|(1/n)\langle \varepsilon_n, M_n(h)\varepsilon_n\rangle| = (1/n)\,\Big|\sum_{i=1}^n \varepsilon_i \Big(\sum_{\ell=1}^h W_{n,h}(\ell)\varepsilon_{i(\ell+1)}\Big)\Big|$$

$$\leq (1/n) \sum_{\ell=2}^{h+1} W_{n,h}(\ell-1)\,\Big|\sum_{i=1}^n \varepsilon_i \varepsilon_{i(\ell)}\Big|.$$

Therefore, by (C.4) we have

$$P\left\{\sup_{1\leq h\leq n} \frac{1}{n}\,|\langle \varepsilon_n, M_n(h)\varepsilon_n\rangle| > \delta\right\}$$

$$\leq P\{\sup_{2\leq \ell\leq n} \Big|\sum_{i=1}^n \varepsilon_i \varepsilon_{i(\ell)}\Big| > n\delta\}$$

$$\leq \sum_{\ell=2}^n P\{(\sum_{i=1}^n \varepsilon_i \varepsilon_{i(\ell)})^4 > n^4\delta^4\}$$

$$\leq \sum_{\ell=2}^n \frac{E(\sum_{i=1}^n \varepsilon_i \varepsilon_{i(\ell)})^4}{n^4\delta^4}$$

$$\leq \sum_{\ell=2}^n \frac{4n^2\mu^4}{n^4\delta^4} \quad \text{(by some combinatorial arguments)}$$

$$= \frac{4\mu^4}{n\delta^4} \to 0. \quad \square$$

PROOF OF (S.3). For the sequence $\{h_n\}$ of (C.6), we have

$$\sup_{1\leq i\leq n} \|\mathbf{x}_i - \mathbf{x}_{i(h_n)}\| \to 0, \quad \text{as} \quad n \to \infty,$$

because of (C.1) and (C.6.1). Now by the continuity of f, we see that

$$(1/n)\,\|(I_n - M_n(h_n))\mathbf{f}_n\|^2 \to 0.$$

On the other hand, (C.4), (C.5) and (C.6.2) imply that

$$E\ (1/n) \| \ M_n(h_n)\varepsilon_n \|^2 \to 0.$$

Thus $E\ (1/n)\ \| \mathbf{f}_n - M_n(h_n)\mathbf{y}_n \|^2 \to 0$, which implies (S.3). □

To prove (S.4), we shall further assume the following condition:

(C.7) There exists fixed numbers λ_3, $\lambda_4 > 0$ such that

$$W_{n,h}(1) \leq \lambda_3 h^{-(1/2+\lambda_4)}, \quad \text{for any} \quad n \quad \text{and any} \quad h \in H_n.$$

This condition is satisfied by most commonly-used weight functions including (see Stone, 1977):

(i) (uniform weight function) $W_{n,h}(i) = (1/h)$ for $1 \leq i \leq h$.
(ii) (triangular weight function) $W_{n,h}(i) = 2(h - i + 1)/h(h + 1)$ for $1 \leq i \leq h$.
(iii) (quadratic weight function) $W_{n,h}(i) = 6(h^2 - (i - 1)^2)/h(h + 1)(4h - 1)$ for $1 \leq i \leq h$.

In general, given a nonincreasing positive continuous function $W(\cdot)$ on $[0, 1]$, we may construct weight functions satisfying (C.4)—(C.7) by letting $W_{n,h}(i)$ be proportional to $W(i/h)$.

Now, we prove (S.4) by establishing

$$\sup_{1\leq h\leq n}\ (1/n)\ \| \tilde{M}_n(h)\mathbf{f}_n - M_n(h)\mathbf{f}_n \|^2 \to 0, \tag{2.6}$$

and

$$(1/n) \| \tilde{M}_n(h_n^*)\varepsilon_n - M_n(h_n^*)\varepsilon_n \|^2 \to 0. \tag{2.7}$$

PROOF OF (2.6). It suffices to show that given any $\delta > 0$, we have

$$\sup_{1\leq h\leq n}\sup_{1\leq i\leq n} \left| \sum_{\ell=1}^{h} W_{n,h}(\ell) f(\mathbf{x}_{i(\ell)}) - \sum_{\ell=1}^{h} W_{n,h}(\ell) f(\mathbf{x}_{i(\ell+1)}) \right| \leq \delta$$

for large n.

First define b, k_n as in the proof of (S.1) with $\mu = 2/3$. Then,

$$\begin{aligned}
&\sup_{1\leq h\leq k_n-1}\sup_{1\leq i\leq n} \left| \sum_{\ell=1}^{h} W_{n,h}(\ell)(f(\mathbf{x}_{i(\ell)}) - f(\mathbf{x}_{i(\ell+1)})) \right| \\
&\quad \leq \sup_{1\leq i\leq n,\, 1\leq \ell\leq k_n-1} |f(\mathbf{x}_{i(\ell)}) - f(\mathbf{x}_{i(\ell+1)})| \\
&\quad \leq \sup_{1\leq i\leq n,\, 1\leq \ell\leq k_n-1} |f(\mathbf{x}_{i(\ell)}) - f(\mathbf{x}_i)| + |f(\mathbf{x}_i) - f(\mathbf{x}_{i(\ell+1)})| \\
&\quad \leq \frac{\delta}{4} + \frac{\delta}{4} = \frac{\delta}{2}.
\end{aligned}$$

On the other hand,

$$\begin{aligned}
&\sup_{k_n\le h\le n}\sup_{1\le i\le n}\left|\sum_{\ell=1}^{h} W_{n,h}(\ell)f(\mathbf{x}_{i(\ell)}) - \sum_{\ell=1}^{h} W_{n,h}(\ell)f(\mathbf{x}_{i(\ell+1)})\right| \\
&\quad\le \sup_{k_n\le h\le n}\sup_{1\le i\le n}\left|\sum_{\ell=1}^{k_n-1} W_{h,h}(\ell)(f(\mathbf{x}_{i(\ell)}) - f(\mathbf{x}_{i(\ell+1)})\right| \\
&\qquad + \sup_{1\le i\le n}\sup_{k_n\le h\le n} W_{n,h}(k_n)\left|f(\mathbf{x}_{i(k_n)})\right| \\
&\qquad + \sum_{\ell=k_n}^{h}\left(W_{n,h}(\ell) - W_{n,h}(\ell+1)\right)\left|f(\mathbf{x}_{i(\ell+1)})\right| \\
&\quad\le (\delta/2) + \sup_{k_n\le h\le n} W_{n,h}(k_n)\left|f_\infty\right| \\
&\quad\le (\delta/2) + k_n^{-1}\left|f_\infty\right|,
\end{aligned}$$

where the last inequality is due to (C.4) and (C.5). Now, by (2.1), $k_n \to \infty$ and the proof is complete. □

PROOF OF (2.7). Observe that

$$\begin{aligned}
&(1/n)\left\|\tilde{M}_n(h_n^*)\varepsilon_n - M_n(h_n^*)\varepsilon_n\right\|^2 \\
&\quad\le 2(W_{n,h_n^*}(1))^2\left\{\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i^2\right\} + \frac{2}{n}\sum_{i=1}^{n}\left\{\sum_{\ell=2}^{h_n^*+1}\left(W_{n,h_n^*}(\ell) - W_{n,h_n^*}(\ell-1)\right)\varepsilon_{i(\ell)}\right\}^2 \\
&\quad\le 2(W_{n,h_n^*}(1))^2\left\{\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i^2\right\} + \frac{2}{n}\sum_{i=1}^{n}\left\{\sum_{\ell=2}^{h_n^*+1}\left(W_{n,h_n^*}(\ell) - W_{n,h_n^*}(\ell-1)\right)^2\right\} \\
&\qquad\cdot\left\{\sum_{\ell=2}^{h_n^*+1}\varepsilon_{i(\ell)}^2\right\} \\
&\quad\le 2(W_{n,h_n^*}(1))^2\left\{\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i^2\right\} + (W_{n,h_n^*}(1))^2\left\{\frac{2}{n}\sum_{i=1}^{n}\sum_{\ell=2}^{h_n^*+1}\varepsilon_{i(\ell)}^2\right\},
\end{aligned}$$

where the last inequality is due to (C.5). In view of (C.7), the proof of (2.7) will be complete supposing that the following two statements hold:

(2.8) $$h_n^* \to \infty, \quad \text{in probability.}$$

(2.9) There exists a constant λ_5 such that

$$\sum_{i=1}^{n}\sum_{\ell=2}^{h_n^*+1}\varepsilon_{i(\ell)}^2 \le \lambda_5\, h_n^*\sum_{i=1}^{n}\varepsilon_i^2.$$

PROOF OF (2.8). It suffices to show that for any natural number N, $P\{h_n^* = N\} \to 0$. Given any $\delta, \delta' > 0$, (1.4) implies that there exists an N' such that

$$P\{(1/n)\left\|\mathbf{f}_n - M_n(h_n^*)\mathbf{y}_n\right\|^2 > \delta\} \le \delta' \quad \text{for any} \quad n \ge N'.$$

Thus for $n \ge N'$,

$$P\{h_n^* = N\} \le P\{(1/n)\left\|\mathbf{f}_n - M_n(N)\mathbf{y}_n\right\|^2 \le \delta\} + \delta'.$$

Take $\delta \le \sigma^2/2N^2$. We shall show that as n tends to ∞, the first term on the right side of the above inequality tends to 0.

First, due to the continuity of f, it is clear that $(1/n)\|\mathbf{f}_n - M_n(N)\mathbf{f}_n\|^2 \to 0$ as $n \to \infty$. Therefore, it suffices to show that

$$P\left\{\frac{1}{n}\|M_n(N)\varepsilon_n\|^2 \le \frac{\sigma^2}{2N^2}\right\} \to 0. \tag{2.10}$$

Now,

$$\begin{aligned} E(1/n)\|M_n(N)\varepsilon_n\|^2 &= E(1/n)\sum_{i=1}^n \{\sum_{\ell=1}^N W_{n,N}(\ell)\varepsilon_{i(\ell+1)}\}^2 \\ &= (1/n)\sum_{i=1}^n \sum_{\ell=1}^N W_{n,N}^2(\ell)\sigma_{i(\ell+1)}^2 \\ &\ge (1/n)\, W_{n,N}^2(1)\sum_{i=1}^n \sigma_{i(2)}^2 \\ &\ge \frac{1}{N^2 n}\sum_{i=1}^n \sigma_{i(2)}^2 \quad \text{by (C.5).} \end{aligned}$$

By (1.5) and Lemma 2.2 below (taking $h = 1$), we see that $\sum_{i=1}^n \sigma_{i(2)}^2 \ge n\sigma^2$ for n sufficiently large. Thus we have $\liminf_{n\to\infty} E(1/n)\|M_n(N)\varepsilon_n\|^2 \ge \sigma^2/N^2$. On the other hand, with the fourth moment condition of (C.3), one can easily verify that $\{(1/n)\|M_n(N)\varepsilon_n\|^2 - E(1/n)\|M_n(N)\varepsilon_n\|^2\} \to 0$ in probability. Hence (2.11) holds. The proof of (2.8) is now complete. □

PROOF OF (2.9). Recall the notation $\ell[i]_n$ from the paragraph preceding Lemma 2.1. Clearly, (2.9) follows from the following lemma.

LEMMA 2.2 *There exists a universal constant* λ_5 *(depending only on the dimension* p*) such that*

$$\#\{i\colon 2 \le \ell[i]_n \le h + 1\} \le \lambda_5 h, \quad \textit{for any} \quad \ell, h, n.$$

The proof of this lemma will be given in the Appendix. We may take, for instance, $\lambda_5 = 2$ for $p = 1$ and $\lambda_5 = 6$ for $p = 2$.

We summarize our results by the following.

THEOREM. *Under* (C.1)–(C.5) *and* (C.7), (1.3) *holds in probability.*

Here note that (C.6) is implied by (C.7).

REMARK 1. Suppose that instead of $\tilde{M}_n(h_n^*)\ \mathbf{y}_n$, we use $M_n(h_n^*)\mathbf{y}_n$ as our estimate, then the consistency can be proved under (C.1)–(C.6) (see (1.4)). Now, is the estimate $\tilde{M}_n(h_n^*)\mathbf{y}_n$ better than $M_n(h_n^*)\mathbf{y}_n$? Intuitively speaking, the answer seems to be yes because it appears that the estimate $M_n(h_n^*)\mathbf{y}_n$ does not use the full information. For instance, in estimating $f(\mathbf{x}_i)$, the observation y_i seems to have been ignored. However, $M_n(h_n^*)\mathbf{y}_n$ does use y_n in estimating $f(\mathbf{x}_i)$ since h_n^* depends partly on y_i. Moreover, if $\tilde{M}_n(h_n^*)\mathbf{y}_n$ is very much different from $M_n(h_n^*)\mathbf{y}_n$, then the cross-validation method may be questionable for such cases. To warrant the success of cross-validation, it is important that our prescription about the class of estimates to be cross-validated should be appropriate (Stone,

1974). To assess the appropriateness of a prescription, one should at least check whether or not $\tilde{M}_n(h_n^*)\mathbf{y}_n$ and $M_n(h_n^*)\mathbf{y}_n$ are close to each other. The condition (C.7) (or any other similar condition) on the prescription about the weight functions serves the purpose of diminishing the chance of the possible drastic changes from $M_n(h_n^*)\mathbf{y}_n$ to $\tilde{M}_n(h_n^*)\mathbf{y}_n$.

REMARK 2. It is clear that similar arguments apply to the case of cross-validation by the leaving-k-out method with k being a fixed integer.

APPENDIX

PROOF OF LEMMA 2.1. Recall the notation of $\lambda(\cdot)$ and $B(\mathbf{x}, r)$ from (C.2) and (C.1). Let $\alpha = \sup\{\| \mathbf{v} - \mathbf{u} \|: \mathbf{v}, \mathbf{u} \in \mathscr{X}\}$. Suppose $\ell[j]_n \geq \ell[i]_n$. Then by (C.2),

$$\begin{aligned}
\ell[j]_n - \ell[i]_n &\leq \lambda_2 n\{\lambda(B(\mathbf{x}_j, \| \mathbf{x}_\ell - \mathbf{x}_i \| + \| \mathbf{x}_i - \mathbf{x}_j \|)) - \lambda(B(\mathbf{x}_i, \| \mathbf{x}_\ell - \mathbf{x}_i \|))\} \\
&\leq \lambda_2 n C[(\| \mathbf{x}_\ell - \mathbf{x}_i \| + \| \mathbf{x}_i - \mathbf{x}_j \|)^P - \| \mathbf{x}_\ell - \mathbf{x}_i \|^P] \\
&\qquad\qquad \text{(where } C = \lambda(B(0, 1))\text{)} \\
&\leq \lambda_2 n C[(\alpha + \| \mathbf{x}_i - \mathbf{x}_j \|)^P - \alpha^P] \\
&\leq 2^P \alpha^{P-1} \lambda_2 C n \| \mathbf{x}_i - \mathbf{x}_j \|.
\end{aligned}$$

Therefore we may take $\alpha = 2^P \alpha^{P-1} \lambda_2 C$ to complete the proof. □

PROOF OF LEMMA 2.2. Denote $\mathbf{S}(\mathbf{x}, r) = \{\mathbf{v}: \mathbf{v} \in R^P \text{ and } \| \mathbf{v} - \mathbf{x} \| = r\}$ and $\mathbf{O}(\mathbf{x}, r) = \{\mathbf{v}: \mathbf{v} \in R^P \text{ and } \| \mathbf{v} - \mathbf{x} \| < r\}$ for any $\mathbf{x} \in R^P$ and $r > 0$. Since $\mathbf{S}(0, 1)$ is compact, we can find a finite number ($\equiv \lambda_5$) of vectors $\mathbf{v}_1, \cdots, \mathbf{v}_{\lambda_5} \in \mathbf{S}(0, 1)$ such that $\cup_{k=1}^{\lambda_5} \mathbf{O}(\mathbf{v}_k, \frac{1}{2}) \supset \mathbf{S}(0, 1)$. Take $\mathbf{C}_k(\mathbf{x}) = \{r\mathbf{v} + \mathbf{x}: r \geq 0 \text{ and } \mathbf{v} \in \mathbf{O}(\mathbf{v}_k, \frac{1}{2}) \cap \mathbf{S}(0, 1)\}$ for $1 \leq k \leq \lambda_5$. Let $\mathbf{x}_{\ell(j;k,n)}$ denote the jth nearest neighbor of $\mathbf{x}_\ell$ among $\{\mathbf{x}_1, \cdots, \mathbf{x}_n\} \cap \mathbf{C}_k(\mathbf{x}_\ell)$. It suffices to show that

$$\{i: 2 \leq \ell[i]_n \leq h + 1\} \subset \cup_{k=1}^{\lambda_5} \{\ell(j; k, n): 2 \leq j \leq h + 1\}.$$

To see this, observe that for any $1 \leq i \leq n$ such that $\mathbf{x}_i \in \mathbf{C}_k(\mathbf{x}_\ell)$ for some k and $i \notin \{\ell(j; k, n): 2 \leq j \leq h + 1\}$, we have

$$\begin{aligned}
\| \mathbf{x}_i - \mathbf{x}_\ell \| &= \sup\{\| \mathbf{x}_i - \mathbf{x} \|: \mathbf{x} \in \mathbf{C}_k(\mathbf{x}_\ell) \cap B(\mathbf{x}_\ell, \| \mathbf{x}_\ell - \mathbf{x}_{\ell(h+1;k,n)} \|)\} \\
&> \max_{2 \leq j \leq h+1} \| \mathbf{x}_i - \mathbf{x}_{\ell(j;k,n)} \|,
\end{aligned}$$

where the last inequality holds because for j, $2 \leq j \leq h + 1$, $\mathbf{x}_\ell(j; k, n)$ belongs to the set $\mathbf{C}_k(\mathbf{x}_\ell) \cap B(\mathbf{x}_\ell, \| \mathbf{x}_\ell - \mathbf{x}_{\ell(h+1;k,n)} \|)$. This implies that $\ell[i]_n > h + 1$. Thus the proof is complete. □

Acknowledgements. I thank Professor Wing Hung Wong for providing me with an early version of his paper (Wong 1983), which led to the present work.

REFERENCES

AGARWAL, G. G. and STUDDEN, W. J. (1980). Asymptotic integrated mean square error using least squares and bias minimizing splines. *Ann. Statist.* **8** 1307–1325.

ALLEN, D. M. (1974). The relationship between variable selection and data augmentation and a method for prediction. *Technometrics* **16** 125–127.

CRAVEN, P. and WAHBA, G. (1979). Smoothing noisy data with spline functions. *Numer. Math.* **31** 377–404.

CHOW, Y. S., GEMAN, S. and WU, L. D. (1983). Consistent cross-validated density estimation. *Ann. Statist.* **11** 25–38.

COVER, T. M. (1968). Estimation by the nearest neighbor rule. *IEEE Trans. Information Theory* **IT-14** 50–55.

COVER, T. M. and HART, P. E. (1967). Nearest neighbor pattern classification. *IEEE Trans. Information Theory* **IT-13** 21–27.

EFRON, B. (1983). Estimating the error rate of a prediction rule: improvement on cross-validation. *J. Amer. Statist. Assoc.* **78** 316–331.

FIX, E. and HODGES, J. L., JR. (1951). Discriminatory analysis, nonparametric discrimination, consistency properties. Randolph Field, Texas, Project 21-49-004, Report No. 4.

GEISSER, S. (1975). The predictive sample reuse method with applications. *J. Amer. Statist. Assoc.* **70** 320–328.

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

NADARAYA, E. A. (1964). On estimating regression. *Theor. Probability Appl.* **9** 141–142.

REINSCH, C. (1967). Smoothing by spline functions. *Numer. Math.* **10** 177–183.

SPECKMAN, P. (1982). Efficient nonparametric regression with cross-validated smoothing splines. Technical report, Department of Statistics, University of Missouri-Columbia.

STONE, C. J. (1977). Consistent nonparametric regression (with discussion). *Ann. Statist.* **5** 595–645.

STONE, M. (1974). Cross-validatory choice and assessment of statistical predictions. *J. Royal Statist. Soc. Ser. B.* **36** 111–147.

WATSON, G. S. (1964). Smooth regression analysis. *Sankhyā Ser. A.* **26** 359–372.

WAHBA, G. and WOLD, S. (1975). A completely automatic French curve: fitting spline functions by cross-validation. *Communication in Statistics* **4** 1–17.

WONG, W. H. (1983). On the consistency of cross-validation in kernel nonparametric regression. *Ann. Statist.* **11** 1257–1262.

DEPARTMENT OF STATISTICS
PURDUE UNIVERSITY
MATHEMATICAL SCIENCES BUILDING
WEST LAFAYETTE, INDIANA 47907