OPTIMAL BANDWIDTH SELECTION IN NONPARAMETRIC REGRESSION FUNCTION ESTIMATION

Wolfgang Härdle
Universität Heidelberg and
University of North Carolina at Chapel Hill

James Stephen Marron
University of North Carolina at Chapel Hill

Keywords and Phrases: nonparametric regression estimation, kernel estimators, optimal bandwidth, smoothing parameter, cross-validation

AMS 1980 subject classifications: Primary 62G05, Secondary 62G20

ABSTRACT

In the setting of nonparametric regression estimation where the independent variables are random, kernel estimators are considered. It is seen that a certain cross-validated choice of the bandwidth is asymptotically equivalent to the bandwidth which minimizes a version of the Mean Integrated Square Error. Since no precise assumptions are made on the amount of "smoothness" of the unknown regression function, the estimators of this paper settle an open problem raised by Stone (1982).

1. INTRODUCTION

This paper presents a solution to the univariate version of an open problem raised in the Special Invited Paper (to the Annals of Statistics) of Stone (1982) (see his Question 3). In fact it will be seen that the results of this paper reach somewhat deeper than the level of Stone's question.

The setting is that of nonparametric regression estimation. Let (X,Y), (X_1,Y_1), (X_2,Y_2), ... be independent random vectors with a common joint density function, $f_{X,Y}(x,y)$. Let $f(x)$ be the marginal density of X. Denote the regression curve of Y on X by

$$m(x) = E[Y|X=x] = \int y f_{X,Y}(x,y)\,dy/f(x).$$

The results of Stone (1982) may be interpreted, in the present setting, as follows. If very precise "smoothness" assumptions are made on $m(x)$, then there is an estimator of $m(x)$, depending on the "smoothness" of m, which optimizes (in a minimax sense, as $n \to \infty$) the exponent of the algebraic rate of convergence of an L^2 error criterion. Stone says such an estimator "achieves the optimal rate of convergence." In question 3, Stone asks if there exists a single estimator which achieves the optimal rate of convergence uniformly over a certain continuum of different smoothness classes.

In this paper not only is an affirmative answer to this question provided, but in fact the results presented here go somewhat further. This is because not only is the exponent of algebraic convergence optimized, but in fact the constant coefficient is in some sense optimized as well.

The results of this paper use kernel estimators, which are defined as follows. Given a positive integer n, a "kernel function" $K(x)$, and a "bandwidth" $h > 0$, define, for $i=1,\ldots,n$, the kernel weights,

$$\alpha_i(x) = n^{-1}h^{-1}\,K\left(\frac{x-X_i}{h}\right).$$

Then $m(x)$ is estimated by

the following weighted average of $Y_1,\ldots,Y_n$, as proposed by Nadaraya (1964) and Watson (1964),

$$m^*(x) = \sum_{i=1}^{n} \alpha_i(x)Y_i/\hat{f}(x) \quad ,$$

where $\hat{f}(x)$ is the familiar Rosenblatt-Parzen estimator of $f(x)$ given by

$$\hat{f}(x) = \sum_{i=1}^{n} \alpha_i(x) \quad .$$

In the case where the marginal density $f(x)$ is known, another reasonable estimate, studied by Johnston (1982), is given by

$$\hat{m}(x)/f(x) = \sum_{i=1}^{n} \alpha_i(x)Y_i/f(x) \quad .$$

It should be noted that this estimator has asymptotic behavior which is, in general, slightly inferior to that of m^*. It is studied here because the nonrandom denominator makes it more tractable.

Given any estimator $m_n(x)$, a popular means of assessing the performance of $m_n(x)$ is the Mean Integrated Square Error, defined by

$$\text{MISE} = E\int [m_n(x) - m(x)]^2 w(x)dx \quad ,$$

where $w(x)$ is a nonnegative "weight function". In the case $m_n = \hat{m}/f$, MISE may be easily analyzed by a variance/bias^2 decomposition. First define

$$S(x) = E(Y^2|X=x) \quad , \tag{1.1}$$

and assume S and f are uniformly continuous. Now, by straightforward computations very similar to those of Rosenblatt (1971), as $n \to \infty$, with $h = h(n) \to 0$,

$$\begin{aligned} E\hat{m}(x) &= \int K(u)m(x-hu)f(x-hu)du, \\ \text{var } \hat{m}(x) &= n^{-1}h^{-1}S(x)f(x)\int K(u)^2du + o(n^{-1}h^{-1}) \quad . \end{aligned} \tag{1.2}$$

Hence,

$$\text{MISE} = E\int [\hat{m}(x)/f(x) - m(x)]^2 w(x)\,dx =$$

(1.3)

$$= n^{-1}h^{-1}\int S(x) f(x)^{-1} w(x)\,dx \int K(u)^2\,du + o(n^{-1}h^{-1}) + s_1(h) \ ,$$

where the bias^2 contribution has been denoted

(1.4) $$s_1(h) = \int [\int K(u) m(x-hu) f(x-hu)\,du - m(x) f(x)]^2 f(x)^{-2} w(x)\,dx.$$

Many authors have dealt with quantities similar to $s_1(h)$ by approximations which arise from assuming that K has some vanishing moments and that m and f have a Taylor expansion. This technique is inadequate for the results of this paper because it gives only upper bounds on the bias^2 part of the MISE. The advantage of $s_1(h)$ is that it measures precisely the rate of convergence of the bias^2. Hence, it is apparent that $s_1(h)$ provides a measure of the quantity called "smoothness" which is perhaps superior to that of Stone (1982) and previous authors.

In the case $m_n = m^*$, some modification of MISE is required because the moments of m^* need not exist (see Rosenblatt (1969)). Since the results of this paper can be more easily described in terms of the estimator $\hat{m}/f$, the more difficult case of m^* will be discussed in Section 2.

From (1.3) it is clear that if the bandwidth h is chosen deterministically to asymptotically minimize MISE (for the estimator $\hat{m}/f$) then use must be made of functionals of the unknown m(x). The theorems of this paper show that this difficulty may be overcome by using the data to specify h through cross-validation. This technique was introduced in the setting of regression function estimation using splines by Wahba and Wold (1975). The idea is to try to choose h to make m(X)/f(X) (or $m^*(X)$) an effective predictor of Y.

This is accomplished as follows. First, for j=1,...,n, define the "leave-one-out" estimators

$$\hat{m}_j(x) = \sum_{i \neq j} \alpha_i(x) Y_i \ ,$$

(1.5)

$$m_j^*(x) = \hat{m}_j(x)/\hat{f}(x) \ .$$

Then form the estimated Residual Sums of Squares

$$\hat{RSS} = n^{-1} \sum_{j=1}^{n} [Y_j - \hat{m}_j(X_j)/f(X_j)]^2 \ ,$$

$$\hat{RSS}^* = n^{-1} \sum_{j=1}^{n} [Y_j - m_j^*(X_j)]^2 \ ,$$

and take $\hat{h}$ (or $\hat{h}^*$) to minimize $\hat{RSS}$ ($\hat{RSS}^*$ respectively). The reason for employing the leave-one-out estimators is that otherwise $\hat{RSS}$ is trivially minimized at h = 0, and $\hat{RSS}$ will yield similar pathological behavior.

In section 3 theorems are stated which show that a slight modification of the cross-validated bandwidths $\hat{h}$ and $\hat{h}^*$ have excellent asymptotic properties. In particular it is shown that choosing h to minimize $\hat{RSS}$ is asymptotically equivalent to choosing h to minimize MISE. A similar result will also be established for $\hat{RSS}^*$, where MISE for m^* is appropriately defined in section 2.

Sections 3 and 4 contain the proofs of the optimality theorems for $\hat{h}$ and $\hat{h}^*$ respectively.

2. AN ERROR CRITERION FOR m^*

As noted in section 1, MISE may not be a meaningful error criterion for the estimator $m_n = m^*$, because the moments of m^* may fail to exist. This difficulty will be overcome by restricting expectation to an event whose

probability tends to one.

First let $\{\underline{h}_n\}$ denote a sequence for which there is an $\epsilon > 0$ so that

$$\lim_{n\to\infty} \underline{h}_n n^{\frac{1}{2}-\epsilon}/\log n = 0, \quad \lim_{n\to\infty} \underline{h}_n n^{\frac{1}{2}-\epsilon} = \infty . \tag{2.1}$$

and let $\{\bar{h}_n\}$ denote a sequence for which

$$\lim_{n\to\infty} \bar{h}_n = 0 \quad , \quad \lim_{n\to\infty} \bar{h}_n \log n = \infty . \tag{2.2}$$

Intuitively, $\underline{h}_n$ tends to 0 "just slower than" $n^{-\frac{1}{2}}$ and $\bar{h}_n$ "just barely" tends to 0. Next suppose that the marginal density $f(x)$ and the kernel $K(x)$ satisfy the assumptions of theorem A of Silverman (1978). Note that the proof of that theorem may be easily extended to show

$$\sup_{\underline{h}_n \le h \le \bar{h}_n} \sup_x |\hat{f}(x) - f(x)| \to 0 \qquad \text{a.s.} \tag{2.3}$$

Next assume that there is a constant $\gamma > 0$ so that for $x \in \text{supp}(w)$, the support of the weight function $w(x)$,

$$f(x) > \gamma .$$

For $n = 1,2,\ldots$ define the event

$$U_n = \{\hat{f}(x) > \gamma/2 \text{ for } h \in [\underline{h}_n, \bar{h}_n],\ x \in \text{supp}(w)\} ,$$

and let U_n^c denote the complement of U_n. Note that, by (2.3)

$$\lim_{n\to\infty} P[U_n] = 1.$$

Also note that on the event U_n there is no difficulty about existence of moments of m^*.

From the above it follows that, on the event U_n,

$$m^*(x)-m(x) = [\hat{m}(x)-m(x)\hat{f}(x)]/\hat{f}(x) =$$

$$(2.4) \qquad = [\hat{m}(x)-m(x)\hat{f}(x)]/f(x)+[\hat{m}(x)-m(x)\hat{f}(x)][f(x)-\hat{f}(x)]/f(x)\hat{f}(x)$$

$$= [\hat{m}(x)-m(x)\hat{f}(x)]/f(x)+o_p(\hat{m}(x)-m(x)\hat{f}(x)) \ ,$$

uniformly over $h \in [\underline{h}_n, \bar{h}_n]$, $x \in \text{supp}(w)$.

Now for i=1,2,... define the residuals

$$(2.5) \qquad \epsilon_i = Y_i - m(X_i) \ .$$

Note that

$$(2.6) \qquad \hat{m}(x)-m(x)\hat{f}(x)= \sum_{i=1}^{n} \alpha_i(x)\epsilon_i + \sum_{i=1}^{n} \alpha_i(x)[m(X_i)-m(x)] \ ,$$

Next, following the notation (1.1), let

$$V(x) = S(x)-\bar{m}(x)^2 \ ,$$

and assume V, f and m are uniformly continuous. A computation very similar to that leading to (1.2) yields

$$E[\hat{m}(x)-m(x)\hat{f}(x)]^2 = n^{-1}h^{-1}V(x)f(x)\int K(u)^2du +$$

$$+ [\int K(u)[m(x-hu)-m(x)]f(x-hu)du]^2 + o(n^{-1}h^{-1}) \ .$$

Next for n=1,2,... let E* denote expectation over the event U_n. It follows from the above that

(2.7) $$MISE^* = E^*\int [m^*(x)-m(x)]^2 w(x)dx =$$

$$= n^{-1}h^{-1}\int V(x) f(x)^{-1} w(x)dx \int K(u)^2 du + s_2(h) + o(n^{-1}h^{-1}),$$

where

(2.8) $$s_2(h) = \int [\int K(u)[m(x-hu)-m(x)]f(x-hu)du]^2 f(x)^{-2} w(x)dx.$$

MISE* is the error criterion that will be used for m* in the optimality theorem of section 3.

3. ASSUMPTIONS AND THEOREMS

The theorems of this paper require the following set of assumptions.

(A.1) Let $\{\underline{h}_n\}$ and $\{\bar{h}_n\}$ satisfy (2.1) and (2.2).

(A.2) There exists a $c < \infty$ and a sequence of positive constants $\{a_n\}$ such that (letting f_Y denote the marginal density of Y)

$$\sup_{\underline{h}_n \le h \le \bar{h}_n} h^{-3} \int_{|y|>a_n} y^2 f_Y(y)dy \le c \ , \quad \text{for } n = 1,2,\ldots,$$

$$\lim_{n\to\infty} \sup_{0\le x\le 1} \int_{|y|>a_n} y^2 f(x,y)dy = 0,$$

$$\lim_{n\to\infty} \sup_{\underline{h}_n \le h \le \bar{h}_n} n^{-\frac{1}{2}} h^{-\frac{1}{2}} a_n (\log n)^2 = 0,$$

$$g_n(x) = \int_{-a_n}^{a_n} y^2 f(x,y)dy \ge \eta > 0 \quad \text{for all } x \in [0,1],\ n=1,2,\ldots$$

$$\int |d_u g_n(uh)| = o((-\log h)^{\frac{1}{2}}) \text{ uniformly over } h \in [\underline{h}_n, \bar{h}_n]$$

(A.3) There are constants $M > 0$, and $\xi > \frac{1}{2} + \epsilon$ (see (2.1)) so that, for any real numbers x and t,

$$|m(x)-m(t)| \le M|x-t|^{\xi} ,$$

$$|S(x)-S(t)| \le M|x-t|^{\xi} ,$$

$$|f(x)-f(t)| \le M|x-t|^{\xi}$$

(A.4) Both $S(x)$ and $f(x)$ are of bounded variation.

(A.5) There is a constant $\gamma > 0$, so that for $x \in [0,1]$,

$$f(x) \ge \gamma .$$

(A.6) The kernel function K has compact support, has a derivative which is of bounded variation and satisfies $\int K(x)dx = 1$.

Note that by (1.4), (2.8), (A.3), (A.4) and (A.5), as $h \to 0$,

$$s_1(h) = o(h) \quad \text{and} \quad s_2(h) = o(h) .$$

Hence, from (1.3) and (2.7) the optimal h for both MISE and MISE* is (asymptotically) contained in $[\underline{h}_n, \bar{h}_n]$.

It should be noted that, by taking $a_n = n^{\frac{1}{4}}(\log n)^{-3}$, a sufficient condition for (A.2) is that Y has a moment of order $8 + \eta$ (some $\eta > 0$). This is substantially weaker than the boundedness conditions on Y that have been imposed by a number of authors, starting with Nadaraya (1964).

It was seen in section 1 that the boundedness of f above 0 is very convenient. The choice of the interval [0,1] in (A.5) is without loss of generality (by a simple rescaling argument). It should also be noted that Stone (1982) has made a similar assumption.

The reader may be surprised at the lack of "vanishing moment" assumptions on K such as those introduced by Parzen (1962). The theorems of this paper are true under assumptions of this type, but such assumptions are not necessary.

Now since f is known to be bounded above 0 only on the interval [0,1], define

$$J = \{j=1,\ldots,n: X_j \in [0,1]\} .$$

Next redefine the estimated Residual Sums of Squares

$$\hat{RSS} = n^{-1} \sum_{j\in J} [Y_j - \hat{m}_j(X_j)/f(X_j)]^2 ,$$

(3.1)

$$\hat{RSS}^* = n^{-1} \sum_{j\in J} [Y_j - m_j^*(X_j)]^2 .$$

Using the notation (2.5), the actual Residual Sum of Squares may be written as

$$RSS = n^{-1} \sum_{j\in J} \epsilon_j^2 . \tag{3.2}$$

It is important to note that RSS is independent of h.

The main theorems of this paper may now be stated

Theorem 1: Under the assumptions (A.1)-(A.6),

$$\hat{RSS} = RSS + MISE + o_p(MISE),$$

uniformly over $h \in [\underline{h}_n, \bar{h}_n]$; where the weight function (in MISE) is taken to be

$$w(x) = 1_{[0,1]}(x) f(x) . \tag{3.3}$$

Theorem 2: Under the assumptions (A.1)-(A.6),

$$\hat{RSS}^* = RSS + MISE^* + o_p(MISE^*),$$

uniformly over $h \in [\underline{h}_n, \bar{h}_n]$; where the weight function (in MISE*) is taken to be

$$w(x) = 1_{[0,1]}(x)f(x) \ .$$

From these theorems it follows that choosing $h \in [\underline{h}_n, \bar{h}_n]$ to minimize $\hat{RSS}$ (or $\hat{RSS}$*) is asymptotically the same as minimizing MISE (or MISE* respectively). Thus, as mentioned in section 1, not only is the exponent of algebraic convergence optimized, but in fact the constant coefficient is the best possible for the given kernel K and weight function w.

To see how this provides an answer to question 3 of Stone (1982), assume that for some $k \in \mathbb{Z}$, the kernel K satisfies, for $j=1,\ldots,\ell$, where $\ell \geq k$

$$\int z^j K(z)dz = 0.$$

Then it is apparent from (1.4), (2.8) and Taylor's theorem that if both f and m satisfy Stone's smoothness condition (1.2) with $p = k+\beta \leq \ell$, then

$$s_1(h) = O(h^{2p}) \quad , \quad s_2(h) = O(h^{2p})$$

uniformly (over functions satisfying (1.2)). Hence, by (1.3) and (2.7), letting $r = 2p/(2p+1)$,

$$MISE = O(n^{-r}), \quad MISE^* = O(n^{-r}) \ ,$$

when $h \sim n^{-1/(2p+1)}$. Now define the Integrated Square Error,

$$ISE = \int_0^1 [m_n(x)-m(x)]f(x)dx \ ,$$

where $m_n(x)$ is either $\hat{m}(x)/f(x)$ or $m^*(x)$. In either case, it follows from the Markov Inequality that

$$\lim_{c\to\infty} \lim_{n\to\infty} \sup_{m,f} P[ISE \geq c\, n^{-r}] = 0.$$

Now using computations very similar to (3.11) - (3.13) of Stone (1982), this may be sharpened to

$$\limsup_{n\to\infty} \sup_{m,f} P[ISE \geq c\, n^{-r}] = 0,$$

for some $c > 0$. This is the statement of question 3 of Stone (1982).

The fact that MISE and MISE* are restricted to [0,1] is also consistent with the results of Stone (1982). The choice of w(x) used here diverges slightly from that of Stone but is very natural because here MISE is proportional to the conditional expected square error

$$E^*[(\hat{m}(X)/\hat{f}(X) - m(X))^2 | X \in [0,1]] ,$$

and similarly for MISE*. It is apparent from (1.3) and (2.7) that the choice of w(x) is irrelevant to optimizing the exponent of algebraic convergence. Hence, the estimators of this paper provide a solution to question 3 of Stone (1982).

It should also be noted that the results of this paper concern a fixed regression function, m, while Stone (1982) works uniformly over classes of regression functions. However, the sup taken over the class Θ_r appearing in Stone's question 3 is a consequence of the results of this paper, as long as his constant K_2 may be considered independent of r.

At first glance, the reader who is familiar with the literature on nonparametric regression estimation may be disturbed by the fact that the weight function, w(x), is truncated off the interval [0,1]. In somewhat similar settings Gasser and Müller (1979) and Rice and Rosenblatt (1983) have reported that an untruncated MISE can be drastically influenced by an "endpoint effect". This is caused by inflation of the bias^2 part of the mean square error near the endpoints of the interval of support of f(x). This effect makes MISE a poor measure of the performance of an estimator. Indeed, choosing h to optimize such a MISE gives an estimator which is seen to be quite suboptimal everywhere except near the endpoints. Despite the discouraging fact, one may see with

very little effort that this endpoint effect does not occur in the present setting. This is because here, unlike in the settings treated by the above authors, the marginal density is assumed to extend (and be "smooth") beyond the interval [0,1] and data points from outside the interval are used in the estimators of this paper. Hence, in the present setting, MISE provides a very reasonable assessment of the performance of the estimator on the entire interval [0,1].

Another approach to the above endpoint difficulties has been taken in the unpublished manuscript by Rice (1982), who assumes a somewhat restrictive "circular design", i.e.: $m(x)$ and its first two derivatives agree at the endpoints of the support of $f(t)$. The advantage of this assumption is that MISE may now be taken over the entire support of f, instead of only over a subinterval as done here. The setting of that paper is also somewhat different from the present because there the independent variables are deterministic and in fact equally spaced. Rice's asymptotics avoid the smoothness questions raised by Stone (1982) but his paper contains some interesting Monte Carlo comparisons of several estimators which appear to be indistinguishable using the asymptotics of this paper.

Finally, it is noted that the fact that the results of this paper require minimization be performed over $h \in [\underline{h},\bar{h}]$ may be somewhat disturbing to the experimenter with a fixed sample size. An interesting and worthwhile extension of the results of this paper would be to show that the cross-validated $\hat{h}$ (and $\hat{h}^*$) satisfies this restriction.

4. PROOF OF THEOREM 1

First it is convenient to define the estimated Mean Integrated Square Error

$$(4.1) \qquad \hat{MISE} = n^{-1} \sum_{j \in J} [\hat{m}(X_j)/f(X_j) - m(X_j)]^2 .$$

Wegman (1972) has used a density estimation analog of $\hat{MISE}$ for Monte Carlo comparisons of estimators. The fact that this would be a reasonable procedure in the present setting is established by

Lemma 1: Under the assumptions of theorem 1,

$$\hat{MISE} = MISE + o_p(MISE) ,$$

uniformly over $h \in [\underline{h}_n, \bar{h}_n]$.

The proof of Lemma 1 is not given here because this is Theorem 1 of Härdle (1983). Theorem 1 of this paper is an easy consequence of Lemma 1 and

Lemma 2: Under the assumptions of Theorem 1,

$$\hat{RSS} = RSS + \hat{MISE} + o_p(MISE),$$

uniformly over $h \in [\underline{h}_n, \bar{h}_n]$.

Proof of Lemma 2

From (3.1) and (3.2), by the addition and subtraction of $m(X_j)$ note that,

$$(4.2) \qquad \hat{RSS} = RSS + A_n + B_n$$

where

$$A_n = n^{-1} \sum_{j \in J} 2\epsilon_j [m(X_j) - \hat{m}_j(X_j)/f(X_j)] ,$$

$$B_n = n^{-1} \sum_{j \in J} [m(X_j) - \hat{m}_j(X_j)/f(X_j)]^2 .$$

These quantities will be approximated in turn.

By (1.5) and (2.5), A_n may be decomposed by

$$-\tfrac{1}{2}A_n = A_{1n} + A_{2n} \quad , \tag{4.3}$$

where

$$A_{1n} = n^{-1} \sum_{j\in J} \epsilon_j f(X_j)^{-1} \sum_{i\neq j} \alpha_i(X_j)\epsilon_i \ ,$$

$$A_{2n} = n^{-1} \sum_{j\in J} \epsilon_j f(X_j)^{-1} [\sum_{i\neq j} \alpha_i(X_j)m(X_i) - m(X_j)f(X_j)] .$$

To approximate the term A_{1n}, note that by conditioning on $\{X_1,\ldots,X_n\}$

$$EA_{1n} = 0 .$$

For $j\in J$, define

$$Z_j = \sum_{i\neq j} \alpha_i(X_j) f(X_j)^{-1} \epsilon_i \epsilon_j \ .$$

Note that

$$Z_j^2 = \sum_{i\neq j} \alpha_i(X_j)^2 f(X_j)^{-2} \epsilon_i^2 \epsilon_j^2 +$$

$$+ \sum_{\substack{i,i'\neq j \\ i\neq i'}} \alpha_i(X_j)\alpha_{i'}(X_j) f(X_j)^{-2} \epsilon_i \epsilon_{i'} \epsilon_j^2 \quad ,$$

and so, by the independence of the residuals $\{\epsilon_j\}$, (A.3), (A.5) and a computation similar to (1.2), uniformly over $h \in [\underline{h}_n, \bar{h}_n]$

$$E[Z_j^2|J] = \sum_{i\neq j} E[\alpha_i(X_j)^2 f(X_j)^{-2} \epsilon_i^2 \epsilon_j^2] \leq$$

$$\leq \gamma^{-2} \sup_{0\leq x\leq 1} S^2(x) \sum_{i\neq j} E[\alpha_i(X_j)^2] =$$

$$= O(n^{-1}h^{-1}) \ .$$

By similar methods it is apparent that, for $j \neq j'$, uniformly over $h \in [\underline{h}_n, \bar{h}_n]$,

$$E[Z_j Z_{j'} | J] = 2E\alpha_{j'}(X_j)\alpha_j(X_{j'})f(X_j)^{-1}f(X_{j'})^{-1}\epsilon_j^2\epsilon_{j'}^2, =$$

$$= \int_0^1\int_0^1 n^{-2}h^{-2}K(\tfrac{x-y}{h})K(\tfrac{y-x}{h})V(x)V(y)P[X\epsilon[0,1]]^{-2}dxdy$$

$$= O(n^{-2}h^{-1}) .$$

It follows from the above that

$$E(A_{1n})^2 = n^{-2}E[\sum_{j\in J} E[Z_j^2|J] + \sum_{\substack{j,j'\in J \\ j\neq j'}} E[Z_j Z_{j'}|J]$$

$$= O(n^{-2}h^{-1}) ,$$

and hence, by (1.3), uniformly over $h \in [\underline{h}_n, \bar{h}_n]$,

$$A_{1n} = O_p(n^{-1}h^{-\frac{1}{2}}) = o_p(\text{MISE}) . \tag{4.4}$$

To bound A_{2n} a decomposition which is somewhat similar to the usual variance/bias^2 is used,

$$A_{2n} = A_{2n}^v + A_{2n}^c + A_{2n}^b , \tag{4.5}$$

where

$$A_{2n}^v = n^{-1}\sum_{j\in J}\epsilon_j f(X_j)^{-1}[\sum_{i\neq j}\alpha_i(X_j)m(X_i) - E[\sum_{i\neq j}\alpha_i(X_j)m(X_i)|J,X_j]] ,$$

$$A_{2n}^c = n^{-1}\sum_{j\in J}\epsilon_j f(X_j)^{-1}[E[\sum_{i\neq j}\alpha_i(X_j)m(X_i)|J,X_j] - \int K(u)f(X_j-hu)m(X_j-hu)du] ,$$

$$A_{2n}^b = n^{-1}\sum_{j\in J}\epsilon_j f(X_j)^{-1}[\int K(u)f(X_j-hu)m(X_j-hu)du - m(X_j)f(X_j)] .$$

Now by Chebyshev's Inequality, (A.3), (A.5) and appropriate conditioning arguments (in particular using $E[\epsilon_j|X_j]=0$) together with the notation

$$\bar{S} = \sup_{0\le x\le 1} S(x),$$

$$P[|A_{2n}^{v}| > n^{-1}h^{-1}\eta] \le$$

$$\le n^2h^2\eta^{-2}n^{-2}E[\sum_{j\in J}\epsilon_j^2 f(X_j)^{-2}[\sum_{i\neq j}\alpha_i(X_j)m(X_i) - E[\sum_{i\neq j}\alpha_i(X_j)m(X_i)|J,X_j]]^2] \le$$

$$\le h^2\eta^{-2}\bar{S}_Y^2 E[\sum_{j\in J}\sum_{i\neq j} E[(\alpha_i(X_j)m(X_i) - E[\sum_{i\neq j}\alpha_i(X_j)m(X_i)|J,X_j])^2|J,X_j]] .$$

But by computations very similar to (1.2) it is seen that, uniformly in x and in $h \in [\underline{h}_n, \bar{h}_n]$,

$$E[(\alpha_i(X_j)m(X_i))^2|X_j=x] = O(n^{-2}h^{-1}) .$$

It follows from this that uniformly over $h \in [\underline{h}_n, \bar{h}_n]$,

$$A_{2n}^{v} = o_p(n^{-1}h^{-1}). \tag{4.6}$$

A similar technique will now be used to approximate A_{2n}^{c} ,

$$P[|A_{2n}^{c}| > n^{-1}h^{-1}\eta] \le$$

$$\le h^2\eta^{-2}\bar{S}_Y^2 E[\sum_{j\in J}(E[\sum_{i\neq j}\alpha_i(X_j)m(X_j)|J,X_j] - \int K(u)f(X_j-hu)m(X_j-hu)du)^2] .$$

Next let #(J) denote the cardinality of J, let $C = \mathbb{R}\setminus[0,1]$, and note that

$$E[\sum_{i\neq j}\alpha_i(X_j)m(X_j)|J,X_j] =$$

$$= \frac{\#(J)-1}{n}\int_0^1 K(u)m(X_j-hu)\frac{f(X_j-hu)}{P[X\in[0,1]]}du + \frac{n-\#(J)}{n}\int_C K(u)m(X_j-hu)\frac{f(X_j-hu)}{P[X\in C]}du.$$

But by the Central Limit Theorem,

$$\#(J)/n = P[X\epsilon[0,1]] + O_p(n^{-\frac{1}{2}}) \quad ,$$

$$1 - \#(J)/n = P[X\epsilon C] + O_p(n^{-\frac{1}{2}}) \quad .$$

It now follows from the above that, uniformly over $h \in [\underline{h}_n, \bar{h}_n]$,

$$A_{2n}^c = o_p(n^{-1}h^{-1}) \quad . \tag{4.7}$$

Using the same technique on A_{2n}^b yields

$$P[|A_{2n}^b| > s_1(h)^{\frac{1}{2}}(nh)^{-\frac{1}{2}}\eta] \leq$$

$$\leq s_1(h)^{-1}n^{-1}h\eta^{-2}\bar{SE}[\sum_{j\in J} E[(\int K(u)f(X_j-hu)m(X_j-hu)du - m(X_j)f(X_j))^2 f(X_j)^{-2}|J]] =$$

$$= s_1(h)^{-1}n^{-1}h\eta^{-2}\bar{SE}[\sum_{j\in J}\int_0^1 (\int K(u)[f(x-hu)m(x-hu)-f(x)m(x)]du)^2 \frac{f(x)^{-1}}{P[X\epsilon[0,1]]}dx] \quad .$$

Thus, from (1.4) and (3.3), uniformly over $h \in [\underline{h}_n, \bar{h}_n]$,

$$A_{2n}^b = o_p(s_1(h)^{\frac{1}{2}}(nh)^{-\frac{1}{2}}) \quad .$$

It follows from this together with (1.3), (4.5), (4.6) and (4.7) that, uniformly over $h \in [\underline{h}_n, \bar{h}_n]$,

$$A_{2n} = o_p(\text{MISE}).$$

This, (4.3) and (4.4) imply that, uniformly over $h \in [\underline{h}_n, \bar{h}_n]$,

$$A_n = o_p(\text{MISE}) \quad . \tag{4.8}$$

Now for the term B_n of (4.2), by another addition and subtraction, using the notation (4.1) write

$$B_n = \hat{\text{MISE}} + B_{1n} + B_{2n} \quad , \tag{4.9}$$

where

$$B_{1n} = 2n^{-1} \sum_{j\in J} [m(X_j) - \hat{m}_j(X_j)/f(X_j)][\hat{m}(X_j)-\hat{m}_j(X_j)]f(X_j)^{-1} ,$$

$$B_{2n} = n^{-1} \sum_{j\in J} [\hat{m}(X_j)-\hat{m}_j(X_j)]^2 f(X_j)^{-2} .$$

Now using Prop. 4 of Mack and Silverman (1982), uniformly over $h \in [\underline{h}_n, \bar{h}_n]$

$$\sup_j |\hat{m}(X_j)-m(X_j)f(X_j)| = o_p(1).$$

Thus, by the fact that

$$\hat{m}(X_j)-\hat{m}_j(X_j) = (nh)^{-1}K(0)Y_j$$

and by the assumptions (A.3) and (A.5), uniformly over $h \in [\underline{h}_n, \bar{h}_n]$

$$B_{1n} = o_p(n^{-1}h^{-1}) = o_p(MISE) ,$$

$$B_{2n} = o_p(n^{-1}h^{-1}) = o_p(MISE) .$$

It now follows from (4.2), (4.8) and (4.9) that, uniformly over $h \in [\underline{h}_n, \bar{h}_n]$

$$\hat{RSS} = RSS + \hat{MISE} + o_p(MISE) ,$$

which completes the proofs of Lemma 2 and Theorem 1.

5. PROOF OF THEOREM 2

This proof is very similar to the proof of Theorem 1 and only parts that are quite different will be given in detail here. Define

$$\hat{MISE}^* = n^{-1} \sum_{j\in J} [m^*(X_j)-m(X_j)]^2 .$$

Lemma 3: Under the assumptions of Theorem 2,

$$\hat{MISE}^* = MISE^* + o_p(MISE^*) ,$$

uniformly over $h \in [\underline{h}_n, \bar{h}_n]$.

Lemma 3 is not proved here because this is Theorem 2 of Härdle (1983). Theorem 2 of this paper follows immediately from Lemma 3 and

Lemma 4: Under the assumptions of Theorem 2,

$$\hat{RSS}^* = RSS + \hat{MISE}^* + o_p(MISE^*) ,$$

uniformly over $h \in [\underline{h}_n, \bar{h}_n]$.

Proof of Lemma 4

As in section 4 write

$$\hat{RSS}^* = RSS + A_n^* + B_n^* \quad , \tag{5.1}$$

where

$$A_n^* = n^{-1} \sum_{j \in J} 2\epsilon_j [m(X_j) - m_j^*(X_j)] \quad ,$$

$$B_n^* = n^{-1} \sum_{j \in J} [m(X_j) - m_j^*(X_j)]^2 \quad .$$

Now by the leave-one-out analogs of (2.4) and (2.6) write (on the event U_n)

$$-\tfrac{1}{2} A_n^* = A_{1n}^* + A_{2n}^* + o_p(A_{1n}^* + A_{2n}^*) \quad ,$$

uniformly over $h \in [\underline{h}_n, \bar{h}_n]$, where

$$A_{1n}^* = n^{-1} \sum_{j \in J} \epsilon_j f(X_j)^{-1} \sum_{i \neq j} \alpha_i(X_j) \epsilon_i \quad ,$$

$$A_{2n}^* = n^{-1} \sum_{j \in J} \epsilon_j f(X_j)^{-1} \sum_{i \neq j} \alpha_i(X_j) [m(X_i) - m(X_j)] \quad .$$

Note that by the methods used to approximate A_{1n} in section 4, uniformly over $h \in [\underline{h}_n, \bar{h}_n]$,

$$A_{1n}^* = o_p(MISE^*) \quad .$$

Again following section 4, write

$$A_{2n}^* = A_{2n}^{*v} + A_{2n}^{*c} + A_{2n}^{*b} \quad ,$$

where

$$A_{2n}^{*v} = n^{-1} \sum_{j \in J} \epsilon_j f(X_j)^{-1} [\sum_{i \neq j} \alpha_i(X_j)(m(X_i) - m(X_j)) - E[\sum_{i \neq j} \alpha_i(X_j)(m(X_i) - m(X_j) | J, X_j]]$$

$$A_{2n}^{*c} = n^{-1} \sum_{j\in J} \epsilon_j f(X_j)^{-1} [E[\sum_{i\neq j} \alpha_i(X_j)(m(X_i)-m(X_j))|J,X_j] - \int K(u)\,[m(X_j-hu)-m(X_i)]f(X_j-hu)du]$$

$$A_{2n}^{*b} = n^{-1} \sum_{j\in J} \epsilon_j f(X_j)^{-1} \int K(u)\,[m(X_j-hu)-m(X_j)]f(X_j-hu)du \ .$$

But now, by methods very similar to those used on A_{2n}^{v}, A_{2n}^{c} and A_{2n}^{b} in section 4, uniformly over $h \in [\underline{h}_n, \bar{h}_n]$,

$$A_{2n}^{*v} = o_p(n^{-1}h^{-1}) \ ,$$

$$A_{2n}^{*c} = o_p(n^{-1}h^{-1}) \ ,$$

$$A_{2n}^{*b} = o_p(n^{-\frac{1}{2}}h^{-\frac{1}{2}}s_2(h)) \ .$$

Thus by (2.7), uniformly over $h \in [\underline{h}_n, \bar{h}_n]$,

$$A_{2n}^{*} = o_p(MISE^*) \ ,$$

and so

$$A_{n}^{*} = o_p(MISE^*) \ . \tag{5.2}$$

To approximate the term B_n^*, write

$$B_n^* = \hat{MISE}^* + B_{1n}^* + B_{2n}^* \ ,$$

where

$$B_{1n}^* = 2n^{-1} \sum_{j\in J} [m(X_j)-\hat{m}(X_j)/\hat{f}(X_j)]\,[\hat{m}(X_j)-\hat{m}_j(X_j)]\hat{f}(X_j)^{-1} \ ,$$

$$B_{2n}^* = n^{-1} \sum_{j\in J} [\hat{m}(X_j)-\hat{m}_j(X_j)]^2\hat{f}(X_j)^{-2} \ .$$

Now approximations similar to (2.4) together with the methods used on B_{1n} and B_{2n} in section 4 yield, uniformly over $h \in [\underline{h}_n, \bar{h}_n]$,

$$B_n^* = \hat{MISE}^* + o_p(MISE^*) \ .$$

Lemma 4 is an easy consequence of this together with (5.1) and (5.2). This completes the proof of Theorem 2.

ACKNOWLEDGEMENT

The authors are grateful to Charles J. Stone for posing the problem solved in this paper, and to Raymond J. Carroll for several useful comments and suggestions.

REFERENCES

GASSER, T. and MÜLLER, H.G. (1979). Kernel estimation of regression functions. Smoothing techniques for curve estimation. Lecture Notes in Math. 757, 23-68.

HÄRDLE, W. (1983). Approximations to the mean integrated square error with applications to optimal bandwidth selection for nonparametric regression estimators. North Carolina Institute of Statistics, Mimeo Series #1529.

JOHNSTON, G.J. (1982). Probabilities of maximal deviations for nonparametric regression function estimation. J. Mult. Anal., 12, 402-414.

MACK, Y.P. and SILVERMAN, B.W. (1982). Weak and strong uniform consistency of kernel regression estimates. Z. Wahrsch. 61, 405-415.

NADARAYA, E.A. (1964). On estimating regression. Theor. Prob. Appl. 9, 141-142.

PARZEN, E. (1962). On the estimation of a probability density and mode. Ann. Math. Statist. 33, 1065-1076.

RICE, J. (1982). Bandwidth choice for nonparametric kernel regression. Unpublished manuscript.

RICE, J. and ROSENBLATT, M. (1983). Smoothing splines: Regression, derivatives and deconvolution. Ann. Statist. 11, 141-156.

ROSENBLATT, M. (1969). Conditional probability density and regression estimators. Multivariate Analysis Vol. 2 (ed. Krishnaiah) 25-31.

ROSENBLATT, M. (1971). Curve Estimates. Ann. Math. Statist. 42, 1815-1842.

SILVERMAN, B.W. (1978). Weak and strong uniform consistency of the kernel estimate of a density and its derivatives. Ann. Statist. 6, 177-184.

STONE, C.J. (1982). Optimal global rates of convergence for nonparametric regression. Ann. Statist. 10, 1040-1053.

WAHBA, G. and WOLD, S. (1975). A completely automatic french curve: fitting spline functions by cross-validation. Comm. Statist. 4, 1-17.

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

WEGMAN, E.J. (1972). Nonparametric probability density estimation II. A comparison of density estimation methods. J. Statist. Comp. Simul. 1, 225-245.