The Annals of Statistics
1985, Vol. 13, No. 4, 1465–1481

OPTIMAL BANDWIDTH SELECTION IN NONPARAMETRIC REGRESSION FUNCTION ESTIMATION

BY WOLFGANG HÄRDLE[1] AND JAMES STEPHEN MARRON[2]

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

Kernel estimators of an unknown multivariate regression function are investigated. A bandwidth-selection rule is considered, which can be formulated in terms of cross validation. Under mild assumptions on the kernel and the unknown regression function, it is seen that this rule is asymptotically optimal.

1. Introduction. Let $(X, Y), (X_1, Y_1), (X_2, Y_2) \dots$ be independent identically distributed $\mathbb{R}^{d+1}$ valued random vectors with Y real valued. Consider the problem of estimating the regression function,

$$m(x) = E[Y|X = x],$$

using $(X_1, Y_1), \dots, (X_n, Y_n)$. In this paper, kernel estimators with a data-driven bandwidth are investigated. Asymptotic optimality is established for a bandwidth-selection rule which can be interpreted in terms of cross validation. The results address two issues. First, they are important in exploratory data analysis, [see, for example, the Projection Pursuit Regression algorithm given in Friedman and Stuetzle (1981).] Second, they settle an open problem of Stone (1982).

Kernel estimators, as introduced by Nadaraya (1964) and Watson (1964), are a local weighted average of the Y_i given by

$$\hat{m}(x) = \hat{m}_h(x) = n^{-1} \sum_{i=1}^{n} h^{-d} K\left(\frac{x - X_i}{h}\right) Y_i / \hat{f}_h(x),$$

where $K\colon \mathbb{R}^d \to \mathbb{R}$ is a kernel (i.e., window) function, $h = h(n) \in \mathbb{R}^+$ is the bandwidth (i.e., smoothing parameter), and $\hat{f}_h(x)$ is the familiar Rosenblatt–Parzen kernel density estimator,

$$\hat{f}(x) = \hat{f}_h(x) = n^{-1} \sum_{i=1}^{n} h^{-d} K\left(\frac{x - X_i}{h}\right),$$

of the marginal density $f(x)$ of X. A slight generalization of this estimator may be obtained by allowing h to be a d-dimensional vector or even a $d \times d$ matrix. The results of this paper extend to that case in a straightforward fashion, although for simplicity of presentation, only scalar h is treated here.

Received May 1983; revised June 1984, December 1984, and May 1985.

[1] Research partially supported by Air Force Office of Scientific Research Contract AFOSR-F49620-82-C0009 and the Deutsche Forschungsgemeinschaft, SFB123.

[2] Research partially supported by Office of Naval Research Contract N00014-81-K-0373.

AMS 1980 *subject classifications.* Primary 62G05; secondary 62G20.

Key words and phrases. Nonparametric regression estimation, kernel estimators, optimal bandwidth, smoothing parameter, cross validation.

One of the crucial points in applying $\hat{m}_h$ is the choice of the bandwidth h. Suppose that h is in some set $H_n \subseteq \mathbb{R}_n^+$ of interest. A *bandwidth-selection rule* $\hat{h} = \hat{h}(n)$ is an H_n-valued function of $(X_1, Y_1), \dots, (X_n, Y_n)$. Let the distance $d(\hat{m}_h, m)$ denote a given measure of accuracy for the estimator $\hat{m}_h$. Following Shibata (1981), the bandwidth-selection rule $\hat{h}$ is said to be *asymptotically optimal with respect to d* when

$$\lim_{n \to \infty} \left[\frac{d(\hat{m}_{\hat{h}}, m)}{\inf\limits_{h \in H_n} d(\hat{m}_h, m)} \right] = 1,$$

with probability one.

In this paper, a bandwidth-selection rule is given, which is then shown to be asymptotically optimal with respect to the distances:

Averaged Squared Error:

$$d_A(\hat{m}, m) = n^{-1} \sum_{j=1}^{n} \left[\hat{m}(X_j) - m(X_j) \right]^2 w(X_j);$$

Integrated Squared Error:

$$d_I(\hat{m}, m) = \int [\hat{m}(x) - m(x)]^2 w(x) f(x)\, dx;$$

Conditional Mean Integrated Squared Error:

$$d_C(\hat{m}, m) = E\left[d_I(\hat{m}, m) | X_1 \dots, X_n \right],$$

where $w(x)$ is a nonnegative weight function.

A bandwidth-selection rule $\hat{h}$ will now be motivated. Write

$$d_I(\hat{m}_h, m) = \int \hat{m}_h^2 wf - 2 \int \hat{m}_h mwf + \int m^2 wf.$$

Since the last summand is independent of h, the goal of minimizing this loss is equivalent to that of minimizing

$$\int \hat{m}_h^2 wf - 2 \int \hat{m}_h mwf. \tag{1.1}$$

But this cannot be realized in practice because this quantity depends on the unknowns m and f. Observe, however, that the second term, for instance, may be written as

$$\int \hat{m}_h mwf = E_{(X, Y)}\left[\hat{m}_h(X) Y w(X) \right].$$

This motivates estimating the second term by

$$n^{-1} \sum_{j=1}^{n} \left[\hat{m}_j(X_j) Y_j w(X_j) \right],$$

where $\hat{m}_j$ is the "leave-one-out" estimator given by

$$\begin{aligned}\hat{m}_j(x) &= (n-1)^{-1}\sum_{i\neq j} h^{-d}K\left(\frac{x-X_i}{h}\right)Y_i/\hat{f}_j(x),\\ \hat{f}_j(x) &= (n-1)^{-1}\sum_{i\neq j} h^{-d}K\left(\frac{x-X_i}{h}\right).\end{aligned}\tag{1.2}$$

Similarly, the first term of (1.1) may be approximated by

$$n^{-1}\sum_{j=1}^{n}\left[\hat{m}_j^2(X_j)w(X_j)\right].$$

Thus, it seems reasonable to take h to minimize the sum of the estimates of the first two terms. Adding a term which is independent of h does not change the bandwidth-selection rule, which is then:

Choose $\hat{h}$ to minimize

$$CV(h) = n^{-1}\sum_{j=1}^{n}\left[Y_j\text{-}\hat{m}_j(X_j)\right]^2 w(X_j).$$

The above motivation is related to some ideas of Rudemo (1982) and Bowman (1984).

Note that the bandwidth-selection rule $\hat{h}$ may also be thought of in terms of choosing h to make $\hat{m}_j(X_j)$ an effective predictor of Y_j. This approach, based on the idea of cross validation, was taken by Clark (1975) and Wahba and Wold (1975) in the setting of spline estimation. See Rice (1984) and Härdle and Marron (1985) for a discussion of other asymptotically optimal bandwith selectors.

In Section 2, a theorem is stated which shows that this bandwidth-selection rule is asymptotically optimal with respect to the distances d_A, d_I, d_C. In Section 3 it is seen how the theorem of Section 2 provides an answer to Question 3 of Stone (1982). Section 4 demonstrates an application of these results. The rest of the paper consists of proofs.

2. Asymptotic optimality.

Assume the weight function w is bounded and supported on a compact set with nonempty interior. Assumptions to be made on the bandwidth, the kernel, and the probability distribution of (X, Y) are:

(A.1) For $n = 1, 2, \dots$ $H_n = [\underline{h}, \bar{h}]$ where

$$\underline{h} \geq C^{-1}n^{\delta-1/d}, \qquad \bar{h} \leq Cn^{-\delta},$$

for some constants C, $\delta > 0$.

(A.2) K is Hölder continuous, ie,

$$|K(x) - K(t)| \leq C\|x-t\|^{\xi},$$

where $\|\cdot\|$ denotes Euclidean norm on $\mathbb{R}^d$, and also

$$\int K(u)\,du = 1,$$

$$\int \|u\|^{\xi}|K(u)|\,du < \infty.$$

(A.3) The regression function m and the marginal density f are Hölder continuous.

(A.4) The conditional moments of Y given $X = x$ are bounded in the sense that there are positive constants $C_1, C_2, \ldots$ so that for $i = 1, 2, \ldots$

$$E\left[|Y|^i|X = x\right] \leq C_i \quad \text{for all } x.$$

(A.5) The marginal density $f(x)$ of X is bounded from below on the support of w.

(A.6) The marginal density $f(x)$ of X is compactly supported.

THEOREM 1. *Under the assumptions* (A.1)–(A.6), *the bandwidth-selection rule, "choose $\hat{h}$ to minimize $CV(h)$," is asymptotically optimal with respect to the distances d_A, d_I, and d_C.*

Condition (A.1) may appear somewhat restrictive because minimization is being performed over an interval whose length tends to zero. This is not a severe restriction because in order to obtain the consistency of $\hat{m}$, the bandwidth must satisfy some similar condition.

The condition (A.4) is substantially weaker than the boundedness conditions on Y that have been imposed by a number of authors, starting with Nadaraya (1964). This condition may be weakened to only a certain finite number of conditional moments being bounded.

Condition (A.5) allows handling of the random denominator of $\hat{m}(x)$. Also, since by (A.3), f and m are assumed to be continuous beyond the support of w, any concern about "boundary effects," such as those described by Gasser and Müller (1979), and Rice and Rosenblatt (1983) is eliminated.

The assumption (A.6) is added for convenience in the proof. It may be weakened to either the existence of any moment of X, or to the compact support of K.

The techniques of this paper may also be applied to estimators related to $\hat{m}$. For example, if the marginal density f is known, as in the "fixed-design" (ie, X not random) case, it makes sense to consider the estimator

$$n^{-1} \sum_{i=1}^{n} h^{-d} K\left(\frac{x - X_i}{h}\right) Y_i/f(X_i),$$

as studied by Johnston (1982).

3. Stone's Question 3.

Stone (1982) investigates the way in which the rate of convergence of nonparametric regression estimators depends on the smoothness of the regression functions. In particular, Stone defines smoothness classes Θ_r, indexed by $r \in \mathbb{R}^+$, and finds an estimator $\hat{m}$, depending on r, which "achieves the rate of convergence r" in the sense that there is a constant C so that

$$\lim_{n \to \infty} \sup_{m \in \Theta_r} P_m\left[d_I(\hat{m}, m) \geq C n^{-r}\right] = 0,$$

where the notation P_m is used to indicate parametrization by m. [See Stone (1982) for the details.] Stone then shows that the rate of convergence r is "optimal" by showing that no estimator of any type can have a faster rate of convergence uniformly over Θ_r. Stone's Question 3 may be expressed as: Is there an estimator $\hat{m}$, independent of r, which achieves the optimal rate uniformly over the smoothness classes?

Under an additional assumption on the smoothness of the marginal density of X, an estimator having this property can be obtained by using a kernel estimator with bandwidth selected as above:

THEOREM 2. *Given* $\eta \in (0, \frac{1}{2})$, *there is a kernel* K *and a constant* $C_r > 0$ *so that, under the assumptions* (A.1)–(A.6),

$$\lim_{n \to \infty} \sup_{r \in [\eta, 1-\eta]} \sup_{f, m \in \Theta_r} P_{f,m}\left[d_I(\hat{m}_{\hat{h}}, m) \geq C_r n^{-r}\right] = 0.$$

The proof of Theorem 2 is in Section 10.

4. An application. In this section it is seen how the proposed kernel regression estimator performs in a real life example. The data consist of 300 pairs of variables where Y denotes liver weight and X denotes age (note here $d = 1$), gathered by the Institute of Forensic Medicine, Universität Heidelberg. It is apparent from the scatter diagram (Figure 1) that the data are quite nonlinear and heteroscedastic, so that a nonparametric approach seems reasonable.

The above theorems make the choice of the smoothing parameter automatic, but there are several quantities that still must be chosen. It is well known [see Table 1 of Rosenblatt (1971)] that the choice of the kernel function, K, is of relatively small importance. We used the kernel of Epanechnikov (1969) given by

$$K(u) = 3(1 - u^2)1_{[-1,1]}(u)/4.$$

Of more concern is the choice of the weight function, w, and through w the choice of its support S. To study the effect on our estimators of different choices of S, we chose

$$w(x) = 1_{[\Delta x, 100 - \Delta x]}(x),$$

where several different values of Δx were considered. Figure 2 shows the graph of the cross-validation function for several choices of Δx. Note the minimum is roughly at $h = 22$ except in the extreme case $\Delta x = 10$ where about 20% of the data has been deleted.

Since this is a real data set, it is impossible to show that $h = 22$ optimizes any of d_A, d_I, or d_C, but Figure 3 allows some comparison. The bandwidths 14 and 30 give regression estimates $\hat{m}(x)$ which seem under (and over, respectively) smoothed. For a final comparison, Figure 1 shows how $\hat{m}(x)$ with $h = 22$ fits the data.

Thus, at least in this example, the techniques of this paper seem relatively independent of the choice of S.

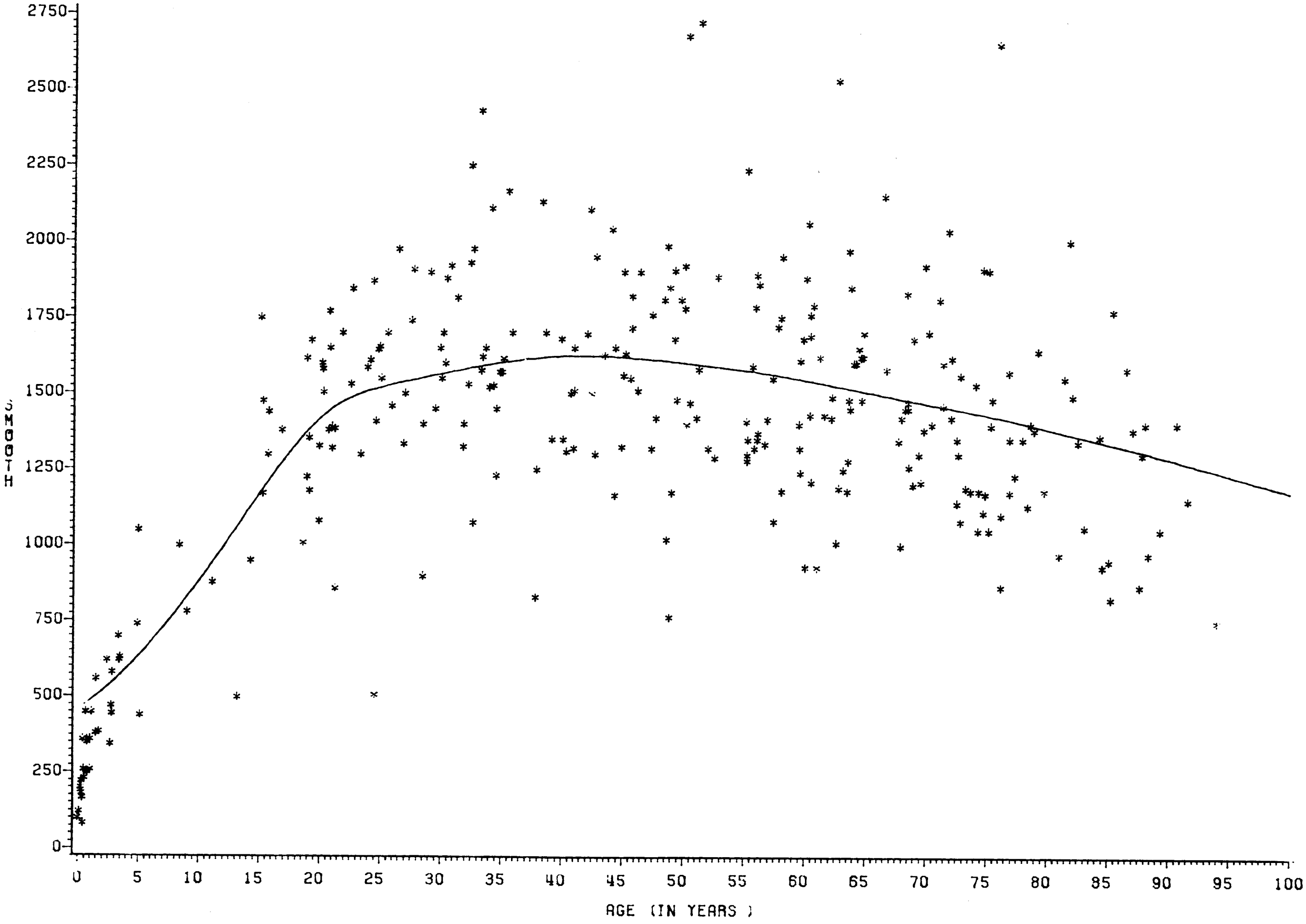

FIG. 1. *Liver weights * age of 300 female persons (smoothed with kernel estimate h = 22).*

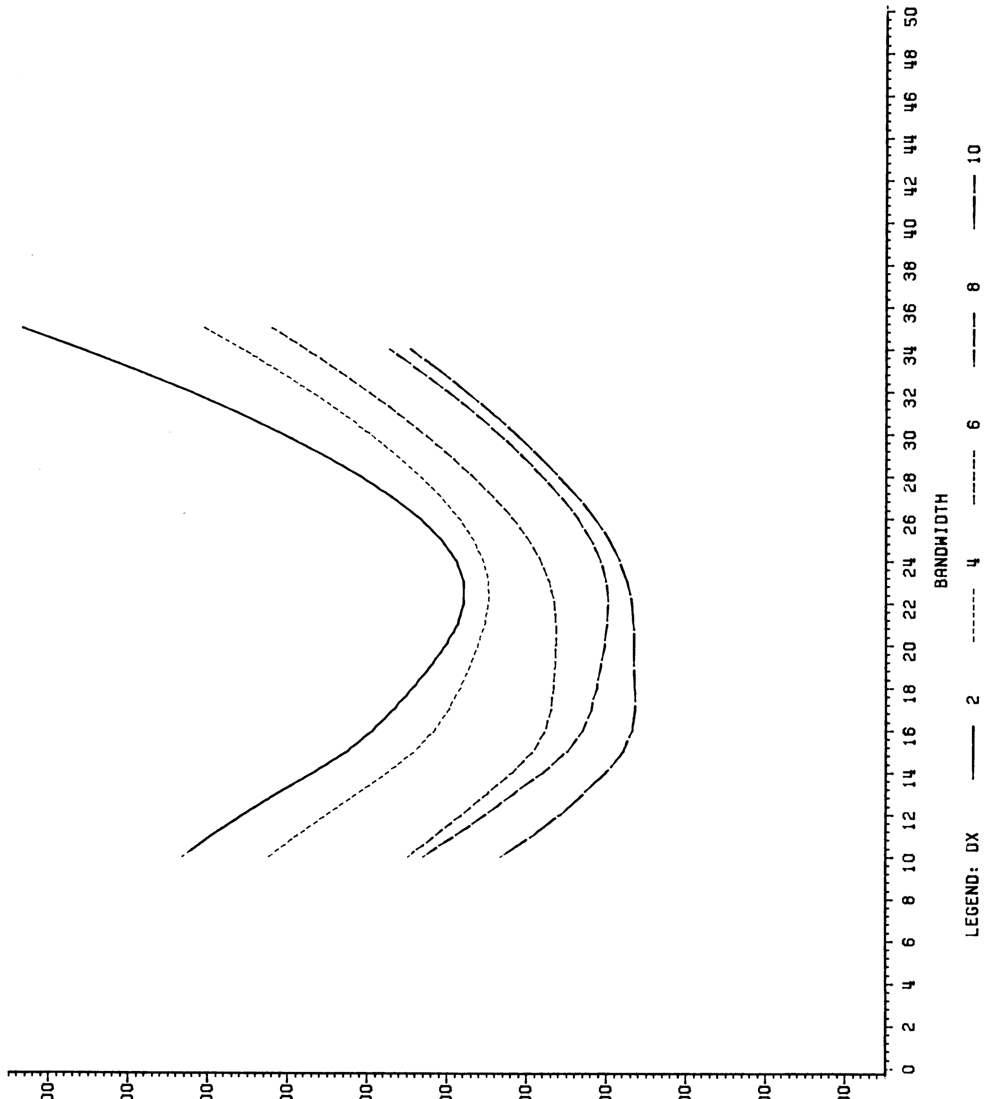

FIG. 2. *CV-graph for liver weight data* ($n = 300$, $\Delta x = 2, 4, 6, 8, 10$, *spacing* $= .01$).

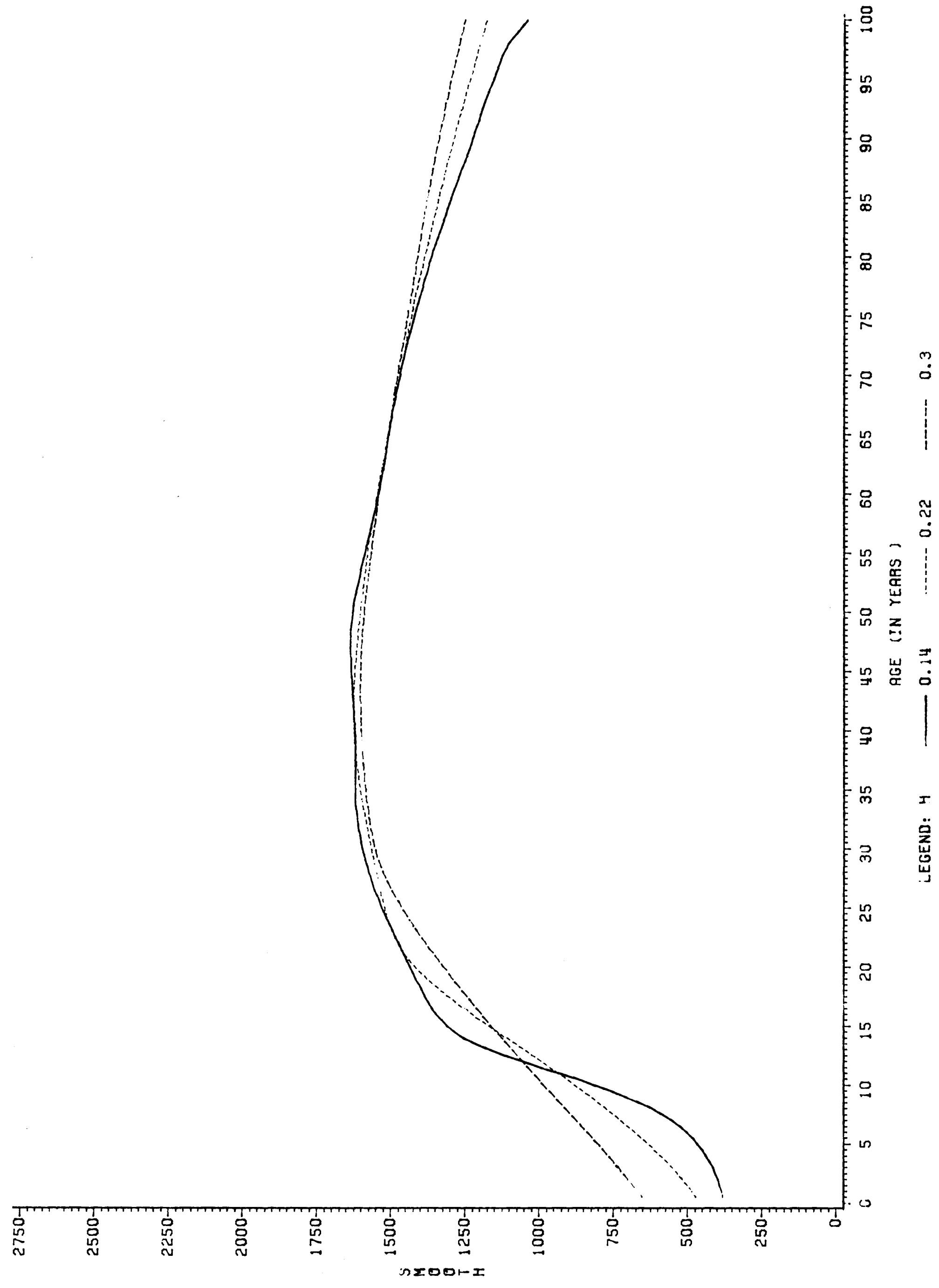

FIG. 3. *Liver weights * age of 300 female persons.*

5. Proof of Theorem 1.

A difficult feature, from an analytical point of view, of the estimator $\hat{m}$ is that it has a random denominator. This will be dealt with by the following device. For x in the compact support of w, write

$$\hat{m} - m = (\hat{m} - m)\hat{f}/f + (\hat{m} - m)(f - \hat{f})/f. \tag{5.1}$$

Note that by the uniform consistency of $\hat{f}$ to f (see Lemma 1 below), the second term in negligible compared to the first [in a sense that is made precise in (5.3) below]. Hence the following distances will be considered

$$d_A^*(\hat{m}, m) = d_A(\hat{m}\hat{f}/f, m\hat{f}/f),$$
$$d_I^*(\hat{m}, m) = d_I(\hat{m}\hat{f}/f, m\hat{f}/f),$$
$$d_C^*(\hat{m}, m) = E[d_I^*(\hat{m}, m)|X_1, \dots, X_n],$$

and also

$$d_M^*(\hat{m}, m) = E[d_I^*(\hat{m}, m)].$$

[The unstarred analogue of d_M^* is not considered here because it may fail to exist, see Härdle and Marron (1983).]

Marron and Härdle (1984) have shown that, under the assumption of Theorem 1,

$$\sup_h \left| \frac{d_A^*(\hat{m}_h, m) - d_M^*(\hat{m}_h, m)}{d_M^*(\hat{m}_h, m)} \right| \to 0 \quad \text{a.s.,}$$
$$\sup_h \left| \frac{d_I^*(\hat{m}_h, m) - d_M^*(\hat{m}_h, m)}{d_M^*(\hat{m}_h, m)} \right| \to 0 \quad \text{a.s.,} \tag{5.2}$$

where $\sup_h$ denotes supremum over H_n. (Actually, this is shown for h in a finite set H_n', whose cardinality grows only algebraically fast, but that can be easily extended to $H_n = [\underline{h}, \overline{h}]$ by a Hölder continuity argument like that used in the proof of the following Lemma 1.) In the rest of this paper, H_n' will denote a finite subset of H_n whose cardinality is bounded by n^ρ, for some $\rho > 0$. The fact that d_A, d_I, d_C, and d_C^* are also similar to d_M^* in the sense (5.2) is the key to the proof.

A substantial part of this is the verification of:

LEMMA 1. *If* (A.1), (A.2), (A.3), *and* (A.6) *hold, then for any compact set* $S \subset \mathbb{R}^d$

$$\sup_{x \in S} \sup_h |\hat{f}_h(x) - f(x)| \to 0 \quad \text{a.s.}$$

The proof of Lemma 1 is in Section 6.

It follows immediately from Lemma 1, (5.1), and (5.2) that

$$\sup_h \left| \frac{d_A(\hat{m}_h, m) - d_M^*(\hat{m}_h, m)}{d_M^*(\hat{m}_h, m)} \right| \to 0 \quad \text{a.s.}$$
$$\sup_h \left| \frac{d_I(\hat{m}_h, m) - d_M^*(\hat{m}_h, m)}{d_M^*(\hat{m}_h, m)} \right| \to 0 \quad \text{a.s.} \tag{5.3}$$

In a similar spirit,

$$\sup_h \left| \frac{d_C(\hat{m}_h, m) - d_M^*(\hat{m}_h, m)}{d_M^*(\hat{m}_h, m)} \right| \to 0 \quad \text{a.s.}$$

follows from:

Lemma 2. *Under the assumptions of Theorem* 1

$$\sup_h \left| \frac{d_C^*(\hat{m}_h, m) - d_M^*(\hat{m}_h, m)}{d_M^*(\hat{m}_h, m)} \right| \to 0 \quad a.s.$$

The proof of Lemma 2 is in Section 7.

Next, to bridge the gap between d_A and $CV(h)$, using the notation (1.2), define

$$\bar{d}_A(\hat{m}, m) = n^{-1} \sum_{j=1}^{n} \left[\hat{m}_j(X_j) - m(X_j)\right]^2 w(X_j),$$

$$\bar{d}_A^*(\hat{m}, m) = n^{-1} \sum_{j=1}^{n} \left[\hat{m}_j(X_j) - m(X_j)\right]^2 \hat{f}_j(X_j)^2 f(X_j)^{-2} w(X_j).$$

Note that, for $j = 1, \dots, n$,

$$\hat{f}_j(x) - \hat{f}(x) = (n-1)^{-1}\hat{f}(x) - (n-1)^{-1} h^{-d} K\left(\frac{x - X_j}{h}\right). \tag{5.4}$$

This relationship and (8.1) allow expressions containing the leave-one-out estimators to be approximated by the same expressions in terms of the ordinary estimators. Thus, by Lemma 1 and (A.1)

$$\sup_{j=1,\dots,n} \sup_x \sup_h |\hat{f}_j(x) - f(x)| \to 0 \quad \text{a.s.}, \tag{5.5}$$

where $\sup_x$ denotes supremum over the support of w. So, as above, with $\hat{m}$ and $\hat{f}$ replaced by $\hat{m}_j$ and $\hat{f}_j$ in (5.1),

$$\sup_h \left| \frac{\bar{d}_A(\hat{m}, m) - d_M^*(\hat{m}, m)}{d_M^*(\hat{m}, m)} \right| \to 0 \quad \text{a.s.}$$

follows from:

Lemma 3. *Under the assumption of Theorem* 1

$$\sup_h \left| \frac{\bar{d}_A^*(\hat{m}, m) - d_M^*(\hat{m}, m)}{d_M^*(\hat{m}, m)} \right| \to 0 \quad a.s.$$

The proof of Lemma 3 is in Section 8.

Let d denote any of d_A, d_I, d_C, d_A^*, d_I^*, d_C^*, d_M^*, $\bar{d}_A$, or $\bar{d}_A^*$. To show

$$\frac{d(\hat{m}_h, m)}{\inf_h d(\hat{m}_h, m)} \to 1 \quad \text{a.s.}, \tag{5.6}$$

it is enough to check that

$$\sup_{h,h'} \frac{|d(\hat{m}_h, m) - d(\hat{m}_{h'}, m) - (CV(h) - CV(h'))|}{d(\hat{m}_h, m) + d(\hat{m}_{h'}, m)} \to 0 \quad \text{a.s.}$$

But in view of the above equivalences, this may be done by showing

$$\sup_{h,h'} \left| \frac{\bar{d}_A(\hat{m}_h, m) - \bar{d}_A(\hat{m}_{h'}, m) - (CV(h) - CV(h'))}{d_M^*(\hat{m}_h, m) + d_M^*(\hat{m}_{h'}, m)} \right| \to 0 \quad \text{a.s.} \tag{5.7}$$

To check this write

$$\bar{d}_A(\hat{m}_h, m) - CV(h) = 2\,\mathrm{Cross}(h) + n^{-1} \sum_{j=1}^{n} \left[m(X_j) - Y_j\right]^2 w(X_j), \tag{5.8}$$

where

$$\mathrm{Cross}(h) = n^{-1} \sum_{j=1}^{n} \left(\hat{m}_j(X_j) - m(X_j)\right)\left(m(X_j) - Y_j\right) w(X_j).$$

Note that the last term on the right of (5.8) is independent of h. So the proof of (5.7) and hence of Theorem 1 will be finished when it is seen that:

LEMMA 4. *Under the assumptions of Theorem* 1

$$\sup_h |\mathrm{Cross}(h)/d_M^*(\hat{m}_h, m)| \to 0 \quad a.s.$$

The proof of Lemma 4 is in Section 9.

6. Proof of Lemma 1. Given $\eta > 0$, for $n = 1, 2, \dots$, find a set $H_n' \subset H_n$ and a set $C_n' \subset C$ so that for any $h \in H_n$ and any $x \in C$, there is $h' \in H_n'$ and $x' \in C_n'$ with

$$|h - h'| \le n^{-\eta} \quad \text{and} \quad |x - x'| \le n^{-\eta}.$$

Note that H_n' and C_n' can be chosen so that their cardinality increases algebraically fast in $n \to \infty$.

Given $\varepsilon > 0$,

$$P\left[\sup_{h \in H_n} \sup_{x \in C} |\hat{f}(x, h) - f(x)| > \varepsilon\right] \le I_n + II_n,$$

where

$$I_n = P\left[\sup_{h' \in H_n'} \sup_{x' \in C_n'} |\hat{f}(x', h') - f(x') > \frac{\varepsilon}{2}\right],$$

$$II_n = P\left[\sup_{h, h', x, x'} |\hat{f}(x, h) - f(x) - (\hat{f}(x', h') - f(x'))| > \frac{\varepsilon}{2}\right],$$

and where $\sup_{h, h', x, x'}$ denotes supremum over $h \in H_n$, $h' \in H_n'$, $x \in C$, and

$x' \in C_n'$. By the Borel–Cantelli Lemma, the proof of Lemma 1 is complete when it is seen that

$$\sum_{n=1}^{\infty} I_n < \infty, \tag{6.1}$$

$$\sum_{n=1}^{\infty} II_n < \infty. \tag{6.2}$$

An argument based on Bernstein's Inequality (Hoeffding, 1963), quite similar to the proof of Lemma 2 of Stone (1984), may be used to establish (6.1). The verification of (6.2) follows in a straightforward fashion from the Hölder continuity of f and K.

7. Proof of Lemma 2. Write

$$d_C^*(\hat{m}_h, m) = \int \left[n^{-1} \sum_{i=1}^{n} \delta_h(x, X_i) \right]^2 f(x)^{-2} w(x)\, dx,$$

where

$$\delta_h(x, X_i) = h^{-d} K\left(\frac{x - X_i}{h} \right) [m(X_i) - m(x)].$$

Under the assumptions of Theorem 1,

$$n^{-1} \sum_{i=1}^{n} \delta_h(x, X_i)$$

is a so called delta sequence estimator [of $g(x) \equiv 0$] which satisfies the conditions of Theorem 1 in Marron and Härdle (1984). Hence,

$$\sup_{h \in H_n'} \left| \frac{d_C^*(\hat{m}_h, m) - d_M^*(\hat{m}_h, m)}{d_M^*(\hat{m}_h, m)} \right| \to 0 \quad \text{a.s.}$$

The above supremum may be easily extended to $H_n = [\underline{h}, \overline{h}]$ by taking the points of H_n' to be sufficiently close together and then using a Hölder continuity argument.

8. Proof of Lemma 3. First note that, as in (5.4), for $j = 1, \dots, n$

$$\begin{aligned} &\hat{m}_j(x) \hat{f}_j(x) - \hat{m}(x) \hat{f}(x) \\ &\quad = (n-1)^{-1} \hat{m}(x) \hat{f}(x) - (n-1)^{-1} h^{-d} K\left(\frac{x - X_j}{h} \right) Y_j. \end{aligned} \tag{8.1}$$

In the following the functions m, $\hat{m}$, $\hat{m}_j$, f, $\hat{f}$, $\hat{f}_j$, and w will be always evaluated at X_j, so it is to be understood that "m" means "$m(X_j)$", and so on. Write

$$\bar{d}_A^*(\hat{m}_h, m) = n^{-1} \sum_{j=1}^{n} \left[A_j + (\hat{m}\hat{f} - m\hat{f}) \right]^2 f^{-2} w,$$

where

$$\begin{aligned} A_j &= \hat{m}_j\hat{f}_j - m\hat{f}_j - (\hat{m}\hat{f} - m\hat{f}) \\ &= (n-1)^{-1}\left[\hat{m}\hat{f} - m\hat{f} - h^{-d}K(0)(Y_j - m)\right]. \end{aligned}$$

Then

$$\begin{aligned} \bar{d}_A^*(\hat{m}_h, m) - d_A^*(\hat{m}_h, m) &= n^{-1}\sum_{j=1}^{n}\left(A_j^2 + 2A_j(\hat{m}\hat{f} - m\hat{f})\right)f^{-2}w \\ &= \left((n-1)^{-2} + 2(n-1)^{-1}\right)d_A^*(\hat{m}_h, m) \\ &\quad -2\left((n-1)^{-2} + (n-1)^{-1}\right)n^{-1} \\ &\quad \cdot \sum_{j=1}^{n}(\hat{m}\hat{f} - m\hat{f})h^{-d}K(0)(Y_j - m)f^{-2}w \\ &\quad +(n-1)^{-2}n^{-1}\sum_{j=1}^{n}h^{-2d}K(0)^2(Y_j - m)^2 f^{-2}w. \end{aligned} \tag{8.2}$$

But, by the Schwartz Inequality,

$$\begin{aligned} &\left|n^{-1}\sum_{j=1}^{n}(\hat{m}\hat{f} - mf)h^{-d}K(0)(Y_j - m)f^{-2}w\right| \\ &\quad \le \left(d_A^*(\hat{m}_h, m)\right)^{1/2}h^{-d}K(0)\left(n^{-1}\sum_{j=1}^{n}(Y_j - m)^2 f^{-2}w\right)^{1/2}, \end{aligned} \tag{8.3}$$

and by the Strong Law of Large Numbers,

$$n^{-1}\sum_{j=1}^{n}(Y_j - m)^2 f^{-2}w \to E\left((Y_j - m)^2 f^{-2}w\right) \quad \text{a.s.} \tag{8.4}$$

By a variance–bias^2 decomposition [see, for example, Parzen (1962), Rosenblatt (1969, 1971)], $d_M^*(\hat{m}_h, m)$ can be written

$$\begin{aligned} d_M^*(\hat{m}_h, m) &= n^{-1}h^{-d}\left[\int V(x)w(x)\,dx\right]\left[\int K(u)^2\,du\right] \\ &\quad + o(n^{-1}h^{-d}) + b^2(h), \end{aligned} \tag{8.5}$$

where the o is uniform over $h \in H_n$, where $V(x)$ denotes the conditional variance

$$V(x) = E\left[Y^2 - m(x)^2 | X = x\right],$$

and where the part analogous to squared bias is denoted

$$\begin{aligned} b^2(h) = \int\Bigg[\int K(u)&[m(x - hu) - m(x)] \\ &\cdot f(x - hu)\,du\Bigg]^2 f(x)^{-1}w(x)\,dx. \end{aligned} \tag{8.6}$$

It follows from (5.2), (8.2), (8.3), (8.4), and (8.5) that

$$\sup_h \frac{|\bar{d}_A^*(\hat{m}_h, m) - d_M^*(\hat{m}_h, m)|}{d_M^*(\hat{m}_h, m)} \to 0 \quad \text{a.s.}$$

This completes the proof of Lemma 3.

9. Proof of Lemma 4. By the expansion (5.1), with $\hat{m}$ and $\hat{f}$ replaced by $\hat{m}_j$ and $\hat{f}_j$, and by (5.5), the proof of Lemma 4 will be complete when it is shown that

$$(9.1) \qquad \sup_h \left| \frac{n^{-1} \sum_{j=1}^{n} (\hat{m}_j(X_j) - m(X_j)) \hat{f}_j(X_j)(Y_j - m(X_j)) f(X_j)^{-1} w(X_j)}{d_M^*(\hat{m}_h, m)} \right| \to 0 \quad \text{a.s.}$$

The numerator of (9.1) may be written as

$$n^{-2} \sum_{i \neq j} U_{i,j} + n^{-2} \sum_{i \neq j} V_{i,j},$$

where

$$U_{i,j} = \left(\frac{n}{n-1}\right) \frac{1}{h^d} K\left(\frac{X_j - X_i}{h}\right) (Y_i - m(X_i))(Y_j - m(X_j)) f(X_j)^{-1} w(X_j),$$

$$V_{i,j} = \left(\frac{n}{n-1}\right) \frac{1}{h^d} K\left(\frac{X_j - X_i}{h}\right) (m(X_i) - m(X_j))(Y_j - m(X_j)) f(X_j)^{-1} w(X_j).$$

Hence (9.1) and the Lemma 4 will be established when it is shown that

$$(9.2) \qquad \sup_h \left| \frac{n^{-2} \sum_{i \neq j} U_{i,j}}{d_M^*(\hat{m}_h, m)} \right| \to 0 \quad \text{a.s.},$$

$$(9.3) \qquad \sup_h \left| \frac{n^{-2} \sum_{i \neq j} V_{i,j}}{d_M^*(\hat{m}_h, m)} \right| \to 0 \quad \text{a.s.}$$

To verify (9.2), note that by Hölder-continuity considerations, it is enough to show that, for H_n' as above,

$$\sup_{h \in H_n'} \left| \frac{n^{-2} \sum_{i \neq j} U_{i,j}}{d_M^*(\hat{m}_h, m)} \right| \to 0 \quad \text{a.s.}$$

For this, note that given $\varepsilon > 0$, $k = 1, 2, \ldots$

$$P\left[\sup_{h \in H_n'} \left| \frac{n^{-2} \sum_{i \neq j} U_{i,j}}{d_M^*(\hat{m}_h, m)} \right| > \varepsilon \right] \leq \varepsilon^{-2k} \#(H_n') \sup_{h \in H_n'} E \left[\frac{n^{-2} \sum_{i \neq j} U_{i,j}}{d_M^*(m_h, m)} \right]^{2k},$$

so that the proof of (9.2) will be complete when it is seen that there is a constant $\tau > 0$, so that for $k = 1, 2, \dots,$ there are constants C_k so that

$$\sup_h E\left[\frac{n^{-2}\sum_{i\neq j} U_{i,j}}{d_M^*(\hat{m}_h, m)}\right]^{2k} \le C_k n^{-\tau k}, \tag{9.4}$$

Similarly (9.3) will be verified by showing that

$$\sup_h E\left[\frac{n^{-2}\sum_{i\neq j} V_{i,j}}{d_M^*(\hat{m}_h, m)}\right]^{2k} \le C_k n^{-\tau k}. \tag{9.5}$$

To check (9.4), for $i, j = 1, \dots, n$ define

$$\begin{aligned} Z_i &= Y_i - m(X_i), \\ a_{ij} &= (n-1)^{-1} h^{-d} K\left(\frac{X_j - X_i}{h}\right) f^{-1}(X_j) w(X_j) 1_{(i \neq j)}. \end{aligned} \tag{9.6}$$

In the following, C will denote a generic constant which may depend on k and may take on different values even in the same formula. From Theorem 2 of Whittle (1960) and (A.4), it follows that

$$\begin{aligned} E\left[\left(n^{-1}\sum_{i\neq j} U_{i,j}\right)^{2k} | X_1, \dots, X_n\right] &= E\left[\left(\sum_{i,j} a_{ij} Z_i Z_j\right)^{2k} | X_1, \dots, X_n\right] \\ &\le C\left(\sum_{i,j} a_{ij}^2\right)^k. \end{aligned}$$

Thus, by (A.5) and integration by substitution,

$$\begin{aligned} E\left[n^{-1}\sum_{i\neq j} U_{i,j}\right]^{2k} &\le CE\left[(n-1)^{-2}\sum_{i\neq j} h^{-2d} K\left(\frac{X_i - X_j}{h}\right)^2\right]^k \\ &\le Cn^{-2k} h^{-2dk} \sum_{l=2}^{2k} n^l h^{dl/2} \le Ch^{-dk}. \end{aligned}$$

The inequality (9.4) follows easily from this and (8.5).

To check (9.5), in addition to the notation (9.6), define

$$b_j = (n-1)^{-1} \sum_{i=1}^n h^{-d} K\left(\frac{X_j - X_i}{h}\right)\left(m(X_i) - m(X_j)\right) f(X_j)^{-1} w(X_j) 1_{(i \neq j)}.$$

Again using Theorem 2 of Whittle (1960) and (A.4),

$$\begin{aligned} E\left[\left(n^{-1}\sum_{i\neq j} V_{i,j}\right)^{2k} | X_1, \dots, X_n\right] &= E\left[\left(\sum_{j=1}^n b_j Z_j\right)^{2k} | X_1, \dots, X_n\right] \\ &\le C\left(\sum_{j=1}^n b_j^2\right)^k. \end{aligned}$$

Using the notation (8.2), it follows that

$$E\left[n^{-1}\sum_{i\neq j} V_{i,j}\right]^{2k} \leq CE\left[(n-1)^{-2}\sum_{j=1}^{n}\left(\sum_{i=1,\,i\neq j}^{n} h^{-d}K\left(\frac{X_j - X_i}{h}\right)\right.\right.$$

$$\left.\left.\times\left(m(X_i) - m(X_j)\right)\right)^{2}\right]^{k}$$

$$\leq C\sum_{l=0}^{k} h^{-dl}\left[nb^2(h)\right]^{l}.$$

The inequality (9.5) is a consequence of this and (8.5).

This completes the proof of Lemma 4.

10. Proof of Theorem 2. Note that in the proof of Theorem 1, all computations are valid uniformly over the sets Θ_r. In particular, letting $\sup_{f,m}$ denote supremum over $m \in \Theta_r$, $r \in [\eta, 1-\eta]$ (5.3) implies that, for $\varepsilon > 0$,

$$\lim_{n\to\infty}\sup_{f,m} P_{f,m}\left[\sup_{h}\left|\frac{d_I(\hat{m}_h, m) - d_M^*(\hat{m}_h, m)}{d_M^*(\hat{m}_h, m)}\right| > \varepsilon\right] = 0, \tag{10.1}$$

and (8.5) may be written as:

$$\sup_{f,m}\left|\frac{d_M^*(\hat{m}_h, m) - n^{-1}h^{-d}\left[\int V(x)w(x)\,dx\right]\left[\int K(u)^2\,du\right] - b^2(h)}{d_M^*(\hat{m}_h, m)}\right| \to 0, \tag{10.2}$$

as $n \to \infty$.

Now given a positive integer l, assume the kernel function K has the property that for nonnegative integers $j_1, \dots, j_d$, with $0 < j_1 + \cdots + j_d \leq l$,

$$\int x_1^{j_1} \cdots x_d^{j_d} K(x)\,dx = 0,$$

where

$$x = (x_1, \dots, x_d).$$

It follows from (A.3), (A.5), and Taylor's Theorem that if m satisfies the condition (1.2) of Stone (1982), with $p = k + \beta$ (Stone's notation) then, for $h \in H_n$,

$$b^2(h) \leq Ch^{2p}.$$

Thus, by (10.2), taking l sufficiently large, if $r = 2p/(2p + d) \in (\eta, 1 - \eta)$

$$\inf_{h} d_M^*(\hat{m}_h, m) \leq Cn^{-r}. \tag{10.3}$$

Theorem 2 follows from (10.1) and (10.3).

Acknowledgment. The authors are grateful to Charles J. Stone for posing the problem solved in this paper as well as suggesting numerous improvements over an earlier draft of this paper, and to Raymond J. Carroll for several useful comments and suggestions.

REFERENCES

BOWMAN, A. W. (1984). An alternative method of cross-validation for the smoothing of density estimates. *Biometrika* **71** 353–361.

CLARK, R. M. (1975). A calibration curve for radio carbon dates. *Antiquity* **49** 251–266.

EPANECHNIKOV, V. (1969). Nonparametric estimates of a multivariate probability density. *Theory Probab. Appl.* **14** 153–158.

FRIEDMAN, J. H. and STUETZLE, W. (1981). Projection pursuit regression. *J. Amer. Statist. Assoc.* **76** 817–823.

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

HÄRDLE, W. and MARRON, J. S. (1983). The nonexistence of moments of some kernel regression estimators. North Carolina Institute of Statistics, Mimeo Series No. 1537.

HÄRDLE, W. and MARRON, J. S. (1985). Asymptotic nonequivalence of some bandwidth selectors in nonparametric regression. *Biometrika* **72**, 481–484.

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

JOHNSTON, G. J. (1982). Probabilities of maximal deviations for nonparametric regression function estimates. *J. Multivariate Anal.* **12** 402–414.

MARRON, J. S. AND HÄRDLE, W. (1984). Random approximations to some measures of accuracy in nonparametric curve estimation. N.C. Inst. of Statist. Mimeo Series No. 1566.

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

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

RICE, J. (1984). Bandwidth choice for nonparametric regression. *Ann. Statist.* **12**, 1215–1230.

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* **2** (P. R. Krishnaiah, ed.) 25–31, Amsterdam: North-Holland Pub. Co.

ROSENBLATT, M. (1971). Curve estimates. *Ann. Math. Statist.* **42** 1815–1842.

RUDEMO, M. (1982). Empirical choice of histograms and kernel density estimators. *Scand. J. Statist.* **9** 65–78.

SHIBATA, R. (1981). An optimal selection of regression variables. *Biometrika* **68** 45–54.

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

STONE, C. J. (1984). An asymptotically optimal window selection rule for kernel density estimates. *Ann. Statist.* **12** 1285–1297.

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. *Sankhya Ser. A* **26** 359–372.

WHITTLE, P. (1960). Bounds for the moments of linear and quadratic forms in independent variables. *Theory Probab. Appl.* **5** 302–305.

DEPARTMENT OF STATISTICS
UNIVERSITY OF NORTH CAROLINA
CHAPEL HILL, NORTH CAROLINA 27514