Multivariate Locally Weighted Least Squares Regression
Author(s): D. Ruppert and M. P. Wand
Source: *The Annals of Statistics*, Vol. 22, No. 3 (Sep., 1994), pp. 1346-1370
Published by: Institute of Mathematical Statistics
Stable URL: http://www.jstor.org/stable/2242229
Accessed: 10/07/2013 17:39

The Annals of Statistics
1994, Vol. 22, No. 3, 1346–1370

MULTIVARIATE LOCALLY WEIGHTED LEAST SQUARES REGRESSION

BY D. RUPPERT[1] AND M. P. WAND[2]

Cornell University and University of New South Wales

Nonparametric regression using locally weighted least squares was first discussed by Stone and by Cleveland. Recently, it was shown by Fan and by Fan and Gijbels that the local linear kernel-weighted least squares regression estimator has asymptotic properties making it superior, in certain senses, to the Nadaraya–Watson and Gasser–Müller kernel estimators. In this paper we extend their results on asymptotic bias and variance to the case of multivariate predictor variables. We are able to derive the leading bias and variance terms for general multivariate kernel weights using weighted least squares matrix theory. This approach is especially convenient when analyzing the asymptotic conditional bias and variance of the estimator at points near the boundary of the support of the predictors. We also investigate the asymptotic properties of the multivariate local quadratic least squares regression estimator discussed by Cleveland and Devlin and, in the univariate case, higher-order polynomial fits and derivative estimation.

1. Introduction. Nonparametric regression has become a rapidly developing field as researchers have realized that parametric regression is not suitable for adequately fitting curves to many data sets that arise in practice. There have been several recent monographs on the topic [Eubank (1988), Müller (1988), Härdle (1990), Hastie and Tibshirani (1990) and Wahba (1990)], where it is shown that nonparametric regression techniques have much to offer in applications. The multivariate case has proved to be very important in practice, and there have been a number of proposed estimators for multivariate predictors, for example, projection pursuit [Friedman and Stuetzle (1981)], ACE [Breiman and Friedman (1985)], generalized additive models [Hastie and Tibshirani (1986)], local regression [Cleveland and Devlin (1988)] and MARS [Friedman (1991)]. Current versions of the S-PLUS computing package include several of these.

In this paper we study the asymptotic bias and variance of multivariate local regression estimators. These estimators are known to have optimal rates of convergence [Stone (1980, 1982)] and have proved to be very useful in modeling real data [Cleveland and Devlin (1988)]. However, it appears that their asymptotic bias and variance have not been studied.

Let $(X_1, Y_1), \dots, (X_n, Y_n)$ be a set of independent and identically distributed $\mathbb{R}^{d+1}$-valued random vectors, where the Y_i are scalar response variables and

Received February 1992; revised December 1993.

[1] Research supported by NSF Grant DMS-90-02791.

[2] Research partially supported by ONR Grant N00014-90-J-1176.

AMS 1991 *subject classifications.* Primary 62G07; secondary 62J02.

Key words and phrases. Bandwidth matrix, boundary effects, derivative estimation, kernel estimator, local polynomial fitting, nonparametric regression, weighted least squares.

the X_i are $\mathbb{R}^d$-valued predictor variables having common density f having support $\operatorname{supp}(f) \subseteq \mathbb{R}^d$. The multivariate nonparametric regression problem is that of estimating

$$m(x) = E(Y \mid X = x)$$

at a vector $x \in \operatorname{supp}(f)$ without the imposition of m belonging to a parametric family of functions. We will assume the model

$$Y_i = m(X_i) + v^{1/2}(X_i)\varepsilon_i, \qquad i = 1, \dots, n,$$

where $v(x) = \operatorname{Var}(Y \mid X = x)$ is finite and the ε_i are mutually independent and identically distributed random variables with zero mean and unit variance and are independent of the X_i's.

Commonly used estimators for $m(x)$ are multivariate versions of the Nadaraya–Watson kernel estimator [Nadaraya (1964), Watson (1964)], the Gasser–Müller kernel estimator [Gasser and Müller (1984)] and the smoothing spline [Schoenberg (1964) and Wahba (1990)].

Kernel estimators have the advantage of being simple to understand intuitively, to analyze mathematically and to implement on a computer, and they are consistent for any smooth m, provided the density f of the X_i's satisfies certain assumptions. In contrast, ACE and generalized additive models are consistent only for rather special m, and the consistency and other properties of projection pursuit and MARS are difficult to assess. However, both the Nadaraya–Watson and Gasser–Müller estimators have certain disadvantages when the design is random. This issue is discussed in depth by Chu and Marron (1991).

In this paper we will study a class of kernel-type nonparametric regression estimators which are known to share the simplicity and consistency of the aforementioned kernel estimators but overcome the main problems of those estimators. The estimators we consider are based on local least squares fitting using kernel weights. Much of our attention will be devoted to the local linear least squares kernel estimator of m which is $\widehat{\alpha}$, the solution for α to the following problem:

$$\text{(1.1)} \qquad \text{Minimize} \quad \sum_{i=1}^{n} \{Y_i - \alpha - \beta^T(X_i - x)\}^2 K_H(X_i - x),$$

where H is a $d \times d$ symmetric positive definite matrix depending on n; K is a d-variate kernel such that $\int K(u)\,du = 1$; and $K_H(u) = |H|^{-1/2}K(H^{-1/2}u)$. We will call $H^{1/2}$ the *bandwidth matrix* since it is the multivariate extension of the usual bandwidth parameter. Problem (1.1) is a straightforward weighted least squares problem, and, assuming that $X_x^T W_x X_x$ is nonsingular, (1.1) has solution

$$\begin{bmatrix} \widehat{\alpha} \\ \widehat{\beta} \end{bmatrix} = \left(X_x^T W_x X_x\right)^{-1} X_x^T W_x Y \quad \text{where } X_x = \begin{bmatrix} 1 & (X_1 - x)^T \\ \vdots & \vdots \\ 1 & (X_n - x)^T \end{bmatrix},$$

$Y = [Y_1, \dots, Y_n]^T$ and $W_x = \operatorname{diag}\{K_H(X_1 - x), \dots, K_H(X_n - x)\}$. The local least squares estimator of $m(x)$ is then

$$(1.2) \qquad \widehat{m}(x;H) = e_1^T \left(X_x^T W_x X_x\right)^{-1} X_x^T W_x Y,$$

where e_1 is the $(d+1) \times 1$ vector having 1 in the first entry and all other entries 0.

Estimator (1.2) has had long use in time series analysis and was introduced as a regression estimator by Stone (1977), and it is a special case of the robust local regression estimators in Cleveland (1979). Stone (1980, 1982) uses (1.2) and its generalization to higher-order polynomials to show the achievability of his bounds on rates of convergence of estimators of m and its derivatives. Cleveland and Devlin (1988) discuss practical implementation of (1.2) for multivariate regression. Cleveland and Devlin also present several interesting case studies where local regression data analysis is considerably more insightful than classical linear regression analysis. The asymptotic properties of the one-dimensional case were studied by Müller (1987) when the X_i are nonrandom and follow a "regular" grid design. Müller shows that at interior points locally weighted regression is asymptotically equivalent to a kernel estimator.

A major advantage of (1.2) is that it is very simple to visualize how the estimator is using the data when estimating m at a point x. This is particularly the case when K is a d-variate probability density function such as the $N(0, I_d)$ density, possibly truncated for compact support. The estimate $\widehat{m}(x;H)$ is found by fitting a plane to the data using weighted least squares, and in the case of the Gaussian kernel the weight given to a point X_i is the value of the $N(X_i - x, H)$ density which has ellipsoidal contours of the form $(X_i - x)^T H^{-1} (X_i - x) = c$, $c > 0$. Clearly, the farther X_i is from x, the less weighting it will receive, but this also depends quite heavily on the value of H, which controls both the size and orientation of the ellipsoids at a given density level. Ellipsoidal contours will occur for any other spherically symmetric kernel. If the true m has a high amount of curvature near x, then it will be important to have more information from nearby observations so one would want H to be chosen so that more weight is given to these observations and reduce the bias of the estimator. However, if m is closer to being linear at x, then variance considerations dictate that one would want more data included in the fitting process and it would be better to have larger ellipsoidal contours. One goal of this article is to show precisely how the bias and variance of $\widehat{m}(x;H)$ depend upon H (see Theorem 2.1.)

Often H is taken to be of simpler form, such as $H = \operatorname{diag}(h_1^2, \dots, h_d^2)$. Having a diagonal bandwidth matrix means that the ellipsoids have their axes in the same direction as the coordinate axis, while for general H they will correspond to the eigenvectors of H and, depending on the shape of m, there are situations where having a full bandwidth matrix is advantageous. This issue is discussed by Wand and Jones (1993) in the density estimation context.

There is another important advantage of local linear least squares kernel estimators which has been demonstrated in the univariate case by Fan (1992, 1993). This is that the asymptotic bias and variance expressions are particularly appealing and appear to be superior to those of the Nadaraya–Watson or

Gasser–Müller kernel estimators. In particular, Fan shows that the local linear squares estimator has an important asymptotic minimax property. Moreover, unlike the Nadaraya–Watson and Gasser–Müller estimators, the bias and variance of (1.2) near the boundary of $\mathrm{supp}(f)$ are of the same order of magnitude as in the interior [Fan and Gijbels (1992)]. This is a very appealing property since, in applications, the boundary region can comprise a large proportion of the data.

Estimator (1.2) is just one member of a hierarchical class of local least squares kernel estimators since one may choose to fit locally polynomials of arbitrary order. This class includes the Nadaraya–Watson kernel estimator which corresponds to local constant fits. Cleveland and Devlin (1988) successfully used local quadratic fits in several of their examples. Because of its importance we also study the asymptotic properties of multivariate local quadratic least squares kernel estimators. In the one-dimensional case we also investigate general polynomial fits and derivative estimation.

Section 2 is devoted to the derivation of the conditional bias and variance of (1.2) both when x is an interior point and when x is near the boundary of the support of f. The same is done for the quadratic fits in Section 3. Section 4 contains the one-dimensional extensions to higher-degree polynomials and derivatives.

The major finding—or one might say major theme—of this paper is this: by analyzing local polynomial fitting directly as a weighted least squares estimator rather than as an approximate kernel estimator, asymptotic behavior is easily elucidated, even in complex settings such as multivariate x, higher polynomials or derivative estimation.

2. Conditional mean squared error properties.

In this section we investigate the asymptotic properties of the conditional bias and variance of $\hat{m}(x;H)$.

We will need the following assumptions:

(A1) The kernel K is a compactly supported, bounded kernel such that $\int uu^TK(u)\,du = \mu_2(K)I$, where $\mu_2(K) \neq 0$ is scalar and I is the $d \times d$ identity matrix. In addition, all odd-order moments of K vanish, that is, $\int u_1^{l_1} \cdots u_d^{l_d}K(u)\,du = 0$ for all nonnegative integers $l_1, \dots, l_d$ such that their sum is odd. (This last condition is satisfied by spherically symmetric kernels and product kernels based on symmetric univariate kernels.)

(A2) The point x is in $\mathrm{supp}(f)$. At x, v is continuous, f is continuously differentiable and all second-order derivatives of m are continuous. Also, $f(x) > 0$ and $v(x) > 0$.

(A3) The sequence of bandwidth matrices $H^{1/2}$ is such that $n^{-1}|H|$ and each entry of H tends to zero as $n \to \infty$ with H remaining symmetric and positive definite. Also, there is a fixed constant L such that the condition number of H (i.e., the ratio of its largest to its smallest eigenvalue) is at most L for all n.

Because the X_i's are random, unless H depends on the X_i's there is a positive probability that $W_x = 0$ and then $\hat{m}(x;H)$ is undefined. For this reason, the ordinary moments of $\widehat{m}(x;H)$ are undefined if H is fixed. If H depends on the X_i's, as is likely in practice, then the moments may exist but could be difficult to compute. Nonetheless, the conditional (given $X_1,\dots,X_n$) moments of $\widehat{m}(x;H)$ are defined with probability tending to 1. In this paper, we study the conditional bias and variance of $\widehat{m}(x;H)$. Working conditionally makes the calculations tractable and has other advantages. Most of our proofs have the following form. First we derive expressions for the exact conditional mean and variance matrix of $\widehat{m}(x;H)$ and then we find the limits of these expressions as $n \to \infty$. The expressions for the exact conditional mean and variance will also hold if the X_i are nonrandom or if they are random but dependent and/or not identically distributed, and their limits can be studied under other assumptions than that the X_i are i.i.d. Another advantage of working conditionally is that, even if H depends upon $X_1,\dots,X_n$, H can still be treated as if fixed.

The asymptotic properties of $\widehat{m}(x;H)$ are different for x lying in the interior of supp(f) than for x lying near the boundary. To make this more precise, let $\mathcal{E}_{x,H} = \{z\colon H^{-1/2}(x-z) \in \text{supp}(K)\}$ be the support of $K_H(x-\cdot)$. We will call x an interior point if $\mathcal{E}_{x,H} \subset \text{supp}(f)$. Otherwise, x will be called a boundary point. If x is a fixed point in the interior of supp(f), then x is an interior point for all large n. However, it is worthwhile to consider a sequence $x = x_n$ converging to a point x_∂ on the boundary of supp(f) sufficiently rapidly that x is a boundary point for all n, say, $x = x_\partial + H^{1/2}c$, for fixed c in supp(K). To avoid degeneracies, we assume the following.

(A4) There is a convex set $\mathcal{C}$ with nonnull interior and containing x_∂ such that

$$\inf_{x\in\mathcal{C}} f(x) > 0. \tag{2.1}$$

Assumption (A4) is a weak assumption and will hold, for example, if the boundary of supp(f) is smooth (has a tangent plane) at x_∂, f is continuous at x_∂ and $f(x_\partial) > 0$. However, (A4) will hold even without the assumption that the boundary is smooth at x_∂. For example, if $d = 3$ and supp(f) is a cube, then (A4) will hold on the edges and vertices of the cube as well as at the faces, provided that f is bounded above 0 on the cube. To get some insight into what (A4) excludes, note that if $d = 2$ and supp$(f) = \{(x_1,x_2)\colon 0 \le x_1 \le 1,\ 0 \le x_2 \le x_1^2\}$, then there will be no $\mathcal{C}$ with the required properties if $x_\partial = (0,0)$.

Our first theorem concerns the conditional mean squared error properties of $\widehat{m}(x;H)$ when x is an interior point while the second theorem looks at boundary points. Their proofs rely on two facts. First, the local linear estimator is a linear function of the Y_i's, so it can be analyzed separately at the linear Taylor approximation to m at x and at the remainder of this approximation. Second, the local linear estimator is *exactly* conditionally unbiased if applied to any linear function, in particular to the linear Taylor approximation to m. After observing these facts, the proof becomes quite simple and relies upon straightforward matrix algebra. Let $K^*(u;x) = e_1^T(X_x^TW_xX_x)^{-1}[1 \quad (u-x)^T]^TK_H(u-x)$.

Then $\hat{m}(x; H) = \sum_{i=1}^{n} K^*(X_i; x)Y_i$,

$$\sum_{i=1}^{n} K^*(X_i; x) = 1 \quad \text{and} \quad \sum_{i=1}^{n} K^*(X_i; x)(X_i - x) = 0. \tag{2.2}$$

It is property (2.2) that makes $\hat{m}(x;H)$ exactly conditionally unbiased for linear functions. A kernel with property (2.2) will be called a *conditional second-order* kernel. A second-order kernel is one whose first moment is 0. Second-order kernels are useful for equally spaced designs if x is an interior point, but otherwise a conditional second-order kernel is much more desirable.

Let $R(K) = \int K(u)^2\,du$, let $D_g(x)$ denote the $d \times 1$ vector of first-order partial derivatives and $\mathcal{H}_g(x)$ denote the $d \times d$ Hessian matrix of a sufficiently smooth d-variate function g at x. Also, let $\mathbf{1}$ denote a generic matrix having each entry equal to 1, the dimensions of which will be clear from the context. Finally, if U_n is a random matrix then $O_P(U_n)$ and $o_P(U_n)$ are to be taken componentwise. We have the following theorem.

THEOREM 2.1. *Let x be a fixed element in the interior of* supp(f). *Assume that* (A1)–(A3) *hold. Then*

$$\begin{aligned} &E\{\hat{m}(x;H) - m(x) \mid X_1, \dots, X_n\} \\ &\qquad = \tfrac{1}{2}\mu_2(K)\operatorname{tr}\{H\mathcal{H}_m(x)\} + o_P\{\operatorname{tr}(H)\} \end{aligned} \tag{2.3}$$

and

$$\begin{aligned} &\operatorname{Var}\{\hat{m}(x;H) \mid X_1, \dots, X_n\} \\ &\qquad = \{n^{-1}|H|^{-1/2}R(K)/f(x)\}v(x)\{1 + o_P(1)\}. \end{aligned} \tag{2.4}$$

REMARK 1. The leading terms in (2.3) and (2.4) do not depend on $X_1, \dots, X_n$, so they can be regarded as playing the role of unconditional bias and variance, respectively. In particular, under appropriate conditions, $\hat{m}(x; H)$ should be asymptotically normal with asymptotic bias and variance given by these expressions. However, the expected absolute values of the remainders are not $o(1)$, so as mentioned before the unconditional bias and variance do not exist.

REMARK 2. The leading conditional bias and variance terms have an intuitively simple interpretation. First of all, $\operatorname{tr}\{H\mathcal{H}_m(x)\}$ is simply the sum of the elementwise products of H and $\mathcal{H}_m(x)$. Each entry of $\mathcal{H}_m(x)$ is a measure of the curvature of m at x in a particular direction and the corresponding entry of H reflects the amount of smoothing being performed in that direction. Hence, the intuitive idea of the bias being increased when there is more curvature and more smoothing is very well described by this leading bias term. As shown by Fan (1992, 1993) for the case $d = 1$, the leading term for the conditional bias does not involve the derivative of f. Fan calls an estimator with this property design-adaptive. If K is the density of the uniform distribution so that all nonzero weights are constant, then the first expression in curly brackets in (2.4) is approximately the reciprocal of the sample size used in the local fit. Otherwise, this expression can be thought of as the reciprocal of the "effective local sample size." Thus, (2.4) reflects the fact that the variance will be penalized by larger conditional variance of Y given $X = x$ and sparser data near x.

PROOF OF THEOREM 2.1. First note that

$$E\{\widehat{m}(x; H) \mid X_1, \ldots, X_n\} = e_1^T (X_x^T W_x X_x)^{-1} X_x^T W_x M \tag{2.5}$$

where $M = [m(X_1), \ldots, m(X_n)]^T$. Let $Q_m(x)$ be the $d \times 1$ vector given by

$$Q_m(x) = \left[(X_1 - x)^T \mathcal{H}_m(x)(X_1 - x), \ldots, (X_n - x)^T \mathcal{H}_m(x)(X_n - x)\right]^T. \tag{2.6}$$

Then Taylor's theorem implies that

$$M = X_x \begin{bmatrix} m(x) \\ D_m(x) \end{bmatrix} + \tfrac{1}{2} Q_m(x) + R_m(x), \tag{2.7}$$

where $R_m(x)$ is a vector of Taylor series remainder terms. When $R_m(x)$ is premultiplied by $e_1^T(X_x^T W_x X_x)^{-1} X_x^T W_x$, the resulting scalar is of negligible order compared to the term arising from $Q_m(x)$, provided $Q_m(x)$ is nonzero, and in any case this scalar is $o_P\{\operatorname{tr}(H)\}$. Then, by (2.5) and (2.6),

$$\begin{aligned} &E\{\widehat{m}(x; H) - m(x) \mid X_1, \ldots, X_n\} \\ &\quad = \tfrac{1}{2} e_1^T (X_x^T W_x X_x)^{-1} X_x^T W_x \{Q_m(x) + R_m(x)\}. \end{aligned} \tag{2.8}$$

Notice that the $D_m(x)$ expression in (2.7) vanishes in (2.8) since

$$e_1^T (X_x^T W_x X_x)^{-1} X_x^T W_x X_x \begin{bmatrix} m(x) \\ D_m(x) \end{bmatrix} = e_1^T \begin{bmatrix} m(x) \\ D_m(x) \end{bmatrix} = m(x). \tag{2.9}$$

Now

$$\begin{aligned} &n^{-1} X_x^T W_x X_x \\ &= \begin{bmatrix} n^{-1} \sum_{i=1}^n K_H(X_i - x) & n^{-1} \sum_{i=1}^n K_H(X_i - x)(X_i - x)^T \\ n^{-1} \sum_{i=1}^n K_H(X_i - x)(X_i - x) & n^{-1} \sum_{i=1}^n K_H(X_i - x)(X_i - x)(X_i - x)^T \end{bmatrix} \end{aligned} \tag{2.10}$$

and using standard results from density estimation,

$$n^{-1} \sum_{i=1}^n K_H(X_i - x) = f(x) + o_P(1),$$

$$n^{-1} \sum_{i=1}^n K_H(X_i - x)(X_i - x) = \mu_2(K) H D_f(x) + o_P(H\mathbf{1})$$

and

$$n^{-1}\sum_{i=1}^{n} K_H(X_i - x)(X_i - x)(X_i - x)^T = \mu_2(K) f(x) H + o_P(H).$$

It follows from this that

$$\begin{aligned}(2.11)\qquad &\left(n^{-1}X_x^T W_x X_x\right)^{-1}\\ &\quad = \begin{bmatrix} f(x)^{-1} + o_P(1) & -D_f(x)^T f(x)^{-2} + o_P(\mathbf{1}) \\ -D_f(x) f(x)^{-2} + o_P(\mathbf{1}) & \left\{\mu_2(K) f(x) H\right\}^{-1} + o_P(H^{-1}) \end{bmatrix}.\end{aligned}$$

Also, it is straightforward to show that

$$\begin{aligned}(2.12)\qquad &n^{-1}X_x^T W_x Q_m(x)\\ &\quad = \begin{bmatrix} n^{-1}\sum_{i=1}^{n} K_H(X_i - x)(X_i - x)^T \mathcal{H}_m(x)(X_i - x) \\ n^{-1}\sum_{i=1}^{n} \left\{K_H(X_i - x)(X_i - x)^T \mathcal{H}_m(x)(X_i - x)\right\}(X_i - x) \end{bmatrix}\end{aligned}$$

and

$$\begin{aligned}(2.13)\qquad &n^{-1}\sum_{i=1}^{n} K_H(X_i - x)\left\{(X_i - x)^T \mathcal{H}_m(x)(X_i - x)\right\}(X_i - x)\\ &\quad = \int K(u)\left\{\left(H^{1/2}u\right)^T \mathcal{H}_m(x)\left(H^{1/2}u\right)\right\}\left(H^{1/2}u\right) f\left(x + H^{1/2}u\right)du\\ &\qquad + o_P\left(H^{3/2}\mathbf{1}\right)\\ &\quad = O_P\left(H^{3/2}\mathbf{1}\right).\end{aligned}$$

It follows from (2.9) and (2.11) that

$$\begin{aligned}&E\left\{\hat{m}(x;H)\,|\,X_1,\dots,X_n\right\} - m(x)\\ &\quad = \tfrac{1}{2} f(x)^{-1} E\left\{n^{-1}\sum_{i=1}^{n} K_H(X_i - x)(X_i - x)^T \mathcal{H}_m(x)(X_i - x)\right\}\\ &\qquad + o_P\left\{\operatorname{tr}(H)\right\}\\ (2.14)\qquad &\quad = \tfrac{1}{2} f(x)^{-1}\left\{\int K(u)\left(H^{1/2}u\right)^T \mathcal{H}_m(x)\left(H^{1/2}u\right) f\left(x + H^{1/2}u\right)du\right\}\\ &\qquad + o_P\left\{\operatorname{tr}(H)\right\}\\ &\quad = \tfrac{1}{2}\operatorname{tr}\left\{H^{1/2}\mathcal{H}_m(x) H^{1/2}\int K(u) u u^T\,du\right\} + o_P\left\{\operatorname{tr}(H)\right\}\\ &\quad = \tfrac{1}{2}\,\mu_2(K)\operatorname{tr}\left\{H\mathcal{H}_m(x)\right\} + o_P\left\{\operatorname{tr}(H)\right\},\end{aligned}$$

as required. For the variance, let $V = \operatorname{diag}\{v(X_1), \dots, v(X_n)\}$. Then

$$\begin{aligned}&\operatorname{Var}\{\widehat{m}(x;H) \,|\, X_1, \dots, X_n\}\\ &\quad = e_1^T (X_x^T W_x X_x)^{-1} X_x^T W_x V W_x X_x (X_x^T W_x X_x)^{-1} e_1.\end{aligned}$$

The upper-left entry of $n^{-1} X_x^T W_x V W_x X_x$ is

$$\begin{aligned}&n^{-1} \sum_{i=1}^{n} K_H(X_i - x)^2 v(X_i)\\ &\quad = |H|^{-1/2} \int K^2(u) v(x + H^{1/2}u) f(x + H^{1/2}u)\, du \{1 + o_P(1)\}\\ &\quad = |H|^{-1/2} R(K) v(x) f(x) \{1 + o_P(1)\},\end{aligned} \tag{2.15}$$

the upper-right block is

$$\begin{aligned}&n^{-1} \sum_{i=1}^{n} K_H(X_i - x)^2 (X_i - x)^T v(X_i)\\ &\quad = |H|^{-1/2} \int K^2(u) u^T H^{1/2} v(x + H^{1/2}u) f(x + H^{1/2}u)\, du \{1 + o_P(1)\}\\ &\quad = O_P(|H|^{1/2})\end{aligned} \tag{2.16}$$

and the lower-right block is

$$\begin{aligned}&n^{-1} \sum_{i=1}^{n} K_H(X_i - x)^2 (X_i - x)(X_i - x)^T v(X_i)\\ &\quad = |H|^{-1/2} H^{1/2} \left\{ \int K^2(u) u u^T\, du \right\} H^{1/2} v(x) f(x) + o_P(|H|^{-1/2} H),\end{aligned} \tag{2.17}$$

so, using (2.11) again and (2.15)–(2.17), we arrive at

$$\begin{aligned}&\operatorname{Var}\{\widehat{m}(x;H) \,|\, X_1, \dots, X_n\}\\ &\quad = n^{-1} |H|^{-1/2} \{R(K) v(x)/f(x)\} \{1 + o_P(1)\}.\end{aligned} \tag{2.18}$$

□

Our next task is to treat boundary points. The key point is that (2.2) holds for all x. Therefore, the analysis used to prove Theorem 2.1 extends nicely to boundary points because (2.5)–(2.10), (2.16) and (2.17) hold for all x and all n. However, (2.11), (2.13) and therefore (2.14) and (2.18) fail at boundary points. We will now develop approximations to $n^{-1} X_x^T W_x X_x$, $n^{-1} X_x^T W_x^2 X_x$ and $n^{-1} X_x^T W_x Q_m(x)$ that are valid when x is a boundary point. Let $\mathcal{D}_{x,H} = \{z\colon (x + H^{1/2} z) \in \operatorname{supp}(f)\} \cap \operatorname{supp}(K)$. Then $\mathcal{D}_{x,H} = \operatorname{supp}(K)$ if and only if x is an interior point. Also let

$$N_x = \begin{bmatrix} \nu_{x,11} & \nu_{x,12} \\ \nu_{x,21} & \nu_{x,22} \end{bmatrix} = \int_{\mathcal{D}_{x,H}} \begin{bmatrix} 1 \\ u \end{bmatrix} [1 \quad u] K(u)\, du,$$

$$T_x = \begin{bmatrix} \tau_{x,11} & \tau_{x,12} \\ \tau_{x,21} & \tau_{x,22} \end{bmatrix} = \int_{\mathcal{D}_{x,H}} \begin{bmatrix} 1 \\ u \end{bmatrix} [1 \quad u] K^2(u)\, du,$$

$$A = \begin{bmatrix} 1 & 0 \\ 0 & H^{1/2} \end{bmatrix}.$$

The entries of N_x and T_x depend on both H and K although this dependence will not be made explicit. It is easily shown that

$$n^{-1}X_x^T W_x X_x = f(x) A N_x A + o_P(A\mathbf{1}A),$$

and

$$n^{-1}X_x^T W_x^2 X_x = |H|^{-1/2} f(x) A T_x A + o_P(A\mathbf{1}A).$$

Using (2.12) one can show that, at any x (either boundary or interior point),

$$n^{-1}X_x^T W_x Q_m(x) = \begin{bmatrix} f(x)\mathrm{tr}\{H^{1/2}\mathcal{H}_m(x)H^{1/2}\nu_{x,22}\} + o_P\{\mathrm{tr}(H)\} \\ f(x)H^{1/2}\int_{\mathcal{D}_{x,H}} uK(u)\{u^T H^{1/2}\mathcal{H}_m(x)H^{1/2}u\}\, du + o_P\{H^{1/2}\mathbf{1}\,\mathrm{tr}(H)\} \end{bmatrix}.$$

By (2.1), N_x is nonsingular and therefore

$$N_x^{-1} = \begin{bmatrix} \nu_x^{11} & \nu_x^{12} \\ \nu_x^{21} & \nu_x^{22} \end{bmatrix},$$

where $\nu_x^{11} = (\nu_{x,11} - \nu_{x,12}\nu_{x,22}^{-1}\nu_{x,21})^{-1}$, $\nu_x^{12} = -(\nu_{x,12}/\nu_{x,11})\nu_x^{22}$ and $\nu_x^{22} = (\nu_{x,22} - \nu_{x,21}\nu_{x,12}/\nu_{x,11})^{-1}$. By (2.9), (2.12) and approximations similar to (2.13) and (2.14), the asymptotic conditional bias and variance of $\widehat{m}(x;H)$ for any point $x \in \mathrm{supp}(f)$ is provided by the following theorem.

THEOREM 2.2. *Suppose that* $x = x_\partial + H^{1/2}c$, *where* c *is a fixed element of* supp(K) *and* (2.1) *holds. Then, under conditions* (A2) *and* (A3) *of Theorem* 2.1,

$$E\{\widehat{m}(x;H) - m(x) \mid X_1,\dots,X_n\} = \frac{e_1^T N_x^{-1}}{2}\int_{\mathcal{D}_{x,H}} \begin{bmatrix} 1 \\ u \end{bmatrix} K(u)u^T H^{1/2}\mathcal{H}_m(x)H^{1/2}u\, du + o_P\{\mathrm{tr}(H)\} \tag{2.19}$$

and

$$\mathrm{Var}\{\widehat{m}(x;H) \mid X_1,\dots,X_n\} = \{n^{-1}|H|^{-1/2}e_1^T N_x^{-1}T_x N_x^{-1}e_1/f(x)\}v(x)\{1 + o_P(1)\}. \tag{2.20}$$

REMARK 3. Theorem 2.2 shows that the conditional bias is $O_P\{\mathrm{tr}(H)\}$ at the boundary as well as in the interior. This result was proved by Fan and Gijbels (1992) for $d = 1$. In fact, for $d = 1$ let $h = H^{1/2}$ and $s_{l,c} = \int_{-c}^{1} u^l K(u)\,du$, and assume that $\mathrm{supp}(K) = [-1, 1]$, $\mathrm{supp}(f) = [0, 1]$ and $x = ch$. Then $\mathcal{D}_{x,H} = [-c, 1]$, $\nu_{x,22} = s_{2,c}$, $\nu_x^{11} = s_{2,c}/(s_{2,c}s_{0,c} - s_{1,c}^2)$ and $\nu_x^{12} = 1/(s_{2,c}s_{0,c} - s_{1,c}^2)$. Therefore, the right-hand side of (2.19) is

$$\frac{1}{2}\left[\frac{s_{2,c}^2 - s_{1,c}s_{3,c}}{s_{2,c}s_{0,c} - s_{1,c}^2}\right] h^2 m''(x),$$

which agrees with the conditional bias term in (3.1) of Fan and Gijbels (1992). Also,

$$\begin{aligned} e_1^T N_x^{-1} T_x N_x^{-1} e_1 &= \frac{s_{2,c}^2 \tau_{x,11} - 2 s_{1,c} s_{2,c} \tau_{x,12} + s_{1,c}^2 \tau_{x,22}}{\left(s_{2,c}s_{0,c} - s_{1,c}^2\right)^2} \\ &= \frac{\int_{-c}^{1} (s_{2,c} - u s_{1,c})^2 K^2(u)\,du}{\left(s_{2,c}s_{0,c} - s_{1,c}^2\right)^2}, \end{aligned}$$

which when substituted into (2.20) agrees with the variance term in (3.1) of Fan and Gijbels (1992).

REMARK 4 (A word of caution). Fan and Gijbels (1992) note that the conditional variance of $\widehat{m}(x; h)$ near the boundary is considerably larger than in the interior. This is illustrated quantitatively in their Figures 1b, 2b and 3b for the normal, Epanechnikov and uniform kernels, respectively. They explain this phenomenon by noting that near the boundary "less observations contribute in computing the estimator." This is only part of the problem. Near the boundary the parameters α and β in (1.1) are no longer asymptotically orthogonal as in the interior. To appreciate the seriousness of this nonorthogonality, consider the Nadaraya–Watson estimator that eliminates the nonorthogonality problem [at the expense of an $O(b)$ bias] by fitting a constant, that is, minimizing (1.1) over α with $\beta^T(X_i - x)$ omitted. The variance of the Nadaraya–Watson estimator is given by (2.20) with $N_x = \nu_{x,11}$ and $T_x = \tau_{x,11}$. To obtain simple expressions, again let $d = 1$ and $\mathrm{supp}(f) = [0, 1]$ and consider the uniform $[-1, 1]$ kernel for which $\int_{-1}^{c} u^l K(u)^2\,du = \frac{1}{2} s_{l,c}$. If their bandwidths are equal, then the ratio of the asymptotic variance of the locally weighted regression estimator to variance of the Nadaraya–Watson estimator is

$$\frac{\left(s_{0,c}s_{2,c}^2 - s_{1,c}^2 s_{2,c}\right)s_{0,c}}{\left(s_{2,c}s_{0,c} - s_{1,c}^2\right)^2} = \frac{s_{2,c}s_{0,c}}{s_{0,c}s_{2,c} - s_{1,c}^2}.$$

At the left boundary ($c = 0$), this ratio is 4. At interior points ($c \geq 1$), $s_{1,c} = 0$ and the ratio is 1 as Fan (1992) has shown. Fan recommends that the locally weighted linear fit become the *benchmark* for nonparametric regression. We certainly share Fan's enthusiasm for locally weighted linear regression, and we appreciate the importance of reducing boundary bias. Also, as a referee has mentioned, variance is easier to model than bias. However, it should be emphasized that, for finite samples, if $f'(0)$ is small relative to $v(0)$, then for x near the boundary the Nadaraya–Watson estimate could be considerably more accurate than a locally weighted linear fit.

REMARK 5. Theorem 2.2 could be extended to include interior points where f is discontinuous, but we will not pursue this here. At interior points where f is continuous, $\nu_{x,12} = \tau_{x,12} = 0$, $\nu_{x,11} = 1$ and $\tau_{x,11} = \int K^2(u)\,du$ so that Theorem 2.2 agrees with (2.3) and (2.4).

Results (2.3) and (2.4) can be combined to give the asymptotic conditional mean squared error (MSE) for estimation at an interior point x:

$$\begin{aligned}&\mathrm{MSE}\{\hat{m}(x;H) \mid X_1,\dots,X_n\}\\ &\quad= n^{-1}|H|^{-1/2}R(K)v(x)/f(x) + \tfrac{1}{4}\mu_2(K)^2\,\mathrm{tr}^2\{H\mathcal{H}_m(x)\}\\ &\qquad+ o_P\{n^{-1}|H|^{-1/2} + \mathrm{tr}^2(H)\}.\end{aligned}$$

However, in practice one typically wants to estimate m over $\mathrm{supp}(f)$, in which case an appropriate error criterion is conditional mean integrated squared error (MISE) given by

$$\begin{aligned}&\mathrm{MISE}\{\hat{m}(\cdot;H) \mid X_1,\dots,X_n\}\\ &\quad= E\int \Big[\{\hat{m}(x;H) - m(x)\}^2 \,\Big|\, X_1,\dots,X_n\Big] w(x)\,dx,\end{aligned}$$

where $w(x)$ is a weight function chosen to ensure that the integral converges. If F is a $d \times d$ symmetric matrix, let $\mathrm{vech}(F)$ be the $\frac{1}{2}d(d+1)$ column vector created by stacking the columns of F, each below the previous, but with entries above the main diagonal omitted, and let $\mathrm{vech}^T(F)$ be the transpose of $\mathrm{vech}(F)$. Also, let $\mathrm{dg}F$ be the same as F but with all off-diagonal entries equal to zero. Then it may be established [see Wand (1992)] that

$$\begin{aligned}&\mathrm{MISE}\{\hat{m}(x;H) \mid X_1,\dots,X_n\}\\ &\quad= n^{-1}|H|^{-1/2}R(K)\int v(x)w(x)/f(x)\,dx \qquad (2.21)\\ &\qquad+ \tfrac{1}{4}\mu_2(K)^2(\mathrm{vech}^T H)\Psi_m(\mathrm{vech}\,H) + o_P\{n^{-1}|H|^{-1/2} + \mathrm{tr}^2(H)\}\end{aligned}$$

where

$$\Psi_m = \int \mathrm{vech}\{2\mathcal{H}_m(x) - \mathrm{dg}\,\mathcal{H}_m(x)\}\mathrm{vech}^T\{2\mathcal{H}_m(x) - \mathrm{dg}\,\mathcal{H}_m(x)\}w(x)\,dx.$$

Details about the numerical minimization of the main terms of (2.21) are given in Wand (1992).

Finally, we mention that Theorems 2.1 and 2.2 may be used to obtain asymptotic approximations for metrics other than those based on squared error loss. In particular, the results of Wand (1990) for the mean *absolute* error of the Gasser–Müller kernel estimator can be readily adapted to $\widehat{m}(x;H)$.

3. Local quadratic regression. In several of the examples of Cleveland and Devlin (1988), improved fits were obtained by local quadratic rather than local linear least squares estimation. In this section we examine the conditional bias and variance of multivariate local quadratic fits.

For local quadratic regression, $\widehat{m}(x;H)$ is again defined by (1.2) but now with

$$X_x = \begin{bmatrix} 1 & (X_1 - x)^T & \mathrm{vech}^T\{(X_1 - x)(X_1 - x)^T\} \\ \vdots & \vdots & \vdots \\ 1 & (X_n - x)^T & \mathrm{vech}^T\{(X_n - x)(X_n - x)^T\} \end{bmatrix}$$

and with e_1 a $\{1 + d + \frac{1}{2}d(d+1)\} \times 1$ vector. If g is any real-valued function of x with all kth-order partial derivatives existing, then its kth-order differential at x_0 is the function of x defined by

$$(d_{x_0}^k)g(x) = \sum_{k_1 + \cdots + k_d = k} \binom{k}{k_1 \cdots k_d} x_1^{k_1} \cdots x_d^{k_d} \left. \frac{\partial^k g(x)}{\partial x_1^{k_1} \cdots \partial x_d^{k_d}} \right|_{x = x_0}.$$

By the same reasoning leading to (2.9), we have

$$\begin{aligned} &E\{\widehat{m}(x;H) - m(x) \mid X_1, \dots, X_n\} \\ &\quad = e_1^T (X_x^T W_x X_x)^{-1} X_x^T W_x \\ (3.1)\qquad &\quad \times \left\{ \frac{1}{3!} \begin{bmatrix} (d_x^3 m)(X_1 - x) \\ \vdots \\ (d_x^3 m)(X_n - x) \end{bmatrix} + \frac{1}{4!} \begin{bmatrix} (d_x^4 m)(X_1 - x) \\ \vdots \\ (d_x^4 m)(X_n - x) \end{bmatrix} + R(x) \right\}, \end{aligned}$$

where $R(x)$ is a vector of Taylor series remainder terms. Now let

$$N_x = \begin{bmatrix} \nu_{x,11} & \nu_{x,12} & \nu_{x,13} \\ \nu_{x,21} & \nu_{x,22} & \nu_{x,23} \\ \nu_{x,31} & \nu_{x,32} & \nu_{x,33} \end{bmatrix} = \int_{\mathcal{D}_{x,H}} \begin{bmatrix} 1 \\ u \\ \mathrm{vech}(uu^T) \end{bmatrix} \left[1 \quad u \quad \mathrm{vech}(uu^T) \right] K(u)\, du.$$

As in Section 2, $\tau_{x,ij}$ and T_x are defined in the same way as $\nu_{x,ij}$ and N_x but with K replaced by K^2. Let $[\nu_x^{11}\ \nu_x^{12}\ \nu_x^{13}]$ be the first row of N_x^{-1}.

THEOREM 3.1. *Suppose that all fourth-order partial derivatives of m are continuous in a neighborhood of x_∂ on the boundary of* $\mathrm{supp}(f)$, $x = x_\partial + H^{1/2}c$, *where c is a fixed element of* supp(K), *that f is continuous at x and that* (2.1) *holds. Also, suppose that H satisfies condition* (A3) *of Theorem* 2.1. *Then*

$$E\{\hat{m}(x;H) - m(x) \mid X_1,\dots,X_n\}$$

$$(3.2) \qquad = \frac{e_1^T N_x^{-1}}{3!} \int_{\mathcal{D}_{x,H}} \begin{bmatrix} 1 \\ u \\ \operatorname{vech}(uu^T) \end{bmatrix} K(u)(d_x^3 m)(H^{1/2}u)\,du$$

$$+ o_P\{\operatorname{tr}(H)^{3/2}\},$$

and

$$(3.3) \qquad \begin{aligned} &\operatorname{Var}\{\hat{m}(x;H) \mid X_1,\dots,X_n\} \\ &\quad = \{n^{-1}|H|^{-1/2} e_1^T N_x^{-1} T_x N_x^{-1} e_1 / f(x)\} v(x) \{1 + o_P(1)\}. \end{aligned}$$

REMARK 6. From (3.2) we see that the conditional bias of $\hat{m}(x;H)$ is $O_P\{(\operatorname{tr}(H))^{3/2}\}$ rather than $O_P\{\operatorname{tr}(H)\}$ as for local linear regression—see (2.3). By (3.6), for local quadratic regression the bias is $O_P[\{\operatorname{tr}(H)\}^2]$ in the interior of supp(f).

PROOF OF THEOREM 3.1. Let C be the $\frac{1}{2}d(d+1) \times \frac{1}{2}d(d+1)$ matrix such that

$$\operatorname{vech}(H^{1/2}uu^T H^{1/2}) = C \operatorname{vech}(uu^T),$$

for all d-vectors u. Each element of C is the product of two elements of $H^{1/2}$ or twice such a product since

$$(H^{1/2}uu^T H^{1/2})_{ij} = \sum_{k=1}^{d} \sum_{l=1}^{d} (H^{1/2})_{ik} (H^{1/2})_{lj} (uu^T)_{kl}.$$

We have

$$\begin{aligned} &En^{-1}\sum_{i=1}^{n} K_H(X_i - x)\operatorname{vech}^T\{(X_i - x)(X_i - x)^T\} \\ &\quad = \int |H|^{-1/2} K\{H^{-1/2}(y-x)\}\operatorname{vech}^T\{(y-x)(y-x)^T\} f(y)\,dy \\ &\quad = \int_{\mathcal{D}_{x,H}} K(u)\operatorname{vech}^T(H^{1/2}uu^T H^{1/2}) f(x + H^{1/2}u)\,du \\ &\quad = \nu_{x,13} C^T f(x) + o_P(\mathbf{1}C^T), \\ &En^{-1}\sum_{i=1}^{n} K_H(X_i - x)(X_i - x)\operatorname{vech}^T[(X_i - x)(X_i - x)^T] \\ &\quad = H^{1/2}\nu_{x,23} C^T f(x) + o_p(H^{1/2}\mathbf{1}C^T), \end{aligned}$$

and

$$\begin{aligned} &En^{-1}\sum_{i=1}^{n} K_H(X_i - x)\operatorname{vech}\{(X_i - x)(X_i - x)^T\}\operatorname{vech}^T\{(X_i - x)(X_i - x)^T\} \\ &\quad = C\nu_{x,33} C^T f(x) + o_P(C\mathbf{1}C^T), \end{aligned}$$

so, using the analog of (2.11) for quadratic regression, we have

$$\begin{aligned}n^{-1}X_x^T W_x X_x &= \left\{\operatorname{diag}\left(1,H^{1/2},C\right)\right\}N_x\left\{\operatorname{diag}\left(1,H^{1/2},C\right)\right\}f(x)\\ &\quad + o_P\left[\left\{\operatorname{diag}\left(1,H^{1/2},C\right)\right\}\mathbf{1}\left\{\operatorname{diag}\left(1,H^{1/2},C\right)\right\}\right].\end{aligned} \tag{3.4}$$

Next,

$$\begin{aligned}&n^{-1}X_x^T W_x\begin{bmatrix}\left(d_x^3 m\right)(X_1 - x)\\ \vdots\\ \left(d_x^3 m\right)(X_n - x)\end{bmatrix}\\ &\quad = f(x)\int_{\mathcal{D}_{x,H}} K(u)\begin{bmatrix}1\\ H^{1/2}u\\ C\operatorname{vech}\left(uu^T\right)\end{bmatrix}\left(d_x^3 m\right)\left(H^{1/2}u\right)du\\ &\qquad + o_P\left\{\operatorname{tr}(H)^{3/2}\begin{bmatrix}1\\ H^{1/2}\mathbf{1}\\ C\mathbf{1}\end{bmatrix}\right\}.\end{aligned} \tag{3.5}$$

By (3.1), (3.4) and (3.5), equation (3.2) holds. Note that in (3.1) the terms involving the fourth-order differential are negligible relative to those involving the third-order differential—this will not be true at interior points. □

In the interior, the right-hand side of (3.2) is 0. We now develop an expression for the asymptotic bias of order $O_P\{\operatorname{tr}(H)^2\}$ that holds in the interior. In the interior, $\nu_{x,12}$, $\nu_{x,21}$ and $\nu_{x,32}$ are all 0, as are the corresponding τ's. Therefore, $\nu_x^{12} = 0$. Also, $\nu_{x,33}$ is diagonal with entries μ_4, followed by $(d-1)$ μ_2's, followed by μ_4 and $(d-2)$ μ_2's and so on.

THEOREM 3.2. *Suppose that x is in the interior of* supp(f), *all fourth-order derivatives of m are continuous derivatives at x, and f has one continuous derivative at x. Also, suppose that condition* (A3) *of Theorem* 2.1 *holds. Then*

$$\begin{aligned}&E\left\{\hat{m}(x;H) - m(x) \mid X_1,\dots,X_n\right\}\\ &\quad = \int\left\{\nu_x^{11} + \nu_x^{13}\operatorname{vech}\left(uu^T\right)\right\}K(u)\\ &\qquad \times\left\{\frac{\left\{D_f(x)^T\left(H^{1/2}u\right)\right\}\left\{\left(d_x^3 m\right)\left(H^{1/2}u\right)\right\}}{3!\,f(x)} + \frac{\left(d_x^4 m\right)\left(H^{1/2}u\right)}{4!}\right\}du\\ &\qquad + o_P\left\{\operatorname{tr}^2(H)\right\}\end{aligned} \tag{3.6}$$

and

$$\begin{aligned}&\operatorname{Var}\left\{\hat{m}(x;H) \mid X_1,\dots,X_n\right\}\\ &\quad = \left[n^{-1}|H|^{-1/2}\left\{\nu_x^{11}\tau_{x,11}\nu_x^{11} + 2\nu_x^{11}\tau_{x,13}\nu_x^{31} + \nu_x^{13}\tau_{x,33}\nu_x^{31}\right\}/f(x)\right]\\ &\qquad \times v(x)\left\{1 + o_P(1)\right\}.\end{aligned} \tag{3.7}$$

PROOF OF THEOREM 3.2. Instead of using (3.5) as in the previous proof, we need the following refinement:

$$\begin{aligned}
&n^{-1}X_x^T W_x \begin{bmatrix} (d_x^3 m)(X_1 - x) \\ \vdots \\ (d_x^3 m)(X_n - x) \end{bmatrix} \\
&\quad = f(x) \int K(u) \begin{bmatrix} 0 \\ H^{1/2}u \\ 0 \end{bmatrix} \left\{ (d_x^3 m)\left(H^{1/2}u\right) \right\} du \\
(3.8) \qquad &\quad + o_P \left\{ \operatorname{tr}(H)^{3/2} \begin{bmatrix} 1 \\ H^{1/2}\mathbf{1} \\ 0 \end{bmatrix} \right\} \\
&\quad + \int K(u) \begin{bmatrix} 1 \\ 0 \\ C \operatorname{vech}\left(uu^T\right) \end{bmatrix} \left\{ (d_x^3 m)\left(H^{1/2}u\right) \right\} \left\{ D_f^T(x) H^{1/2} u \right\} du \\
&\quad + o_P \left\{ \operatorname{tr}(H)^2 \begin{bmatrix} 1 \\ 0 \\ C\mathbf{1} \end{bmatrix} \right\}.
\end{aligned}$$

It is easy to see that

$$\begin{aligned}
&n^{-1}X_x^T W_x \begin{bmatrix} (d_x^4 m)(X_1 - x) \\ \vdots \\ (d_x^4 m)(X_n - x) \end{bmatrix} \\
(3.9) \qquad &\quad = f(x) \int K(u) \begin{bmatrix} 1 \\ 0 \\ C \operatorname{vech}\left(uu^T\right) \end{bmatrix} \left\{ (d_x^4 m)\left(H^{1/2}u\right) \right\} du \\
&\quad + o_P \left\{ \operatorname{tr}(H)^2 \begin{bmatrix} 1 \\ 0 \\ C\mathbf{1} \end{bmatrix} \right\}.
\end{aligned}$$

Also, at interior points some of the components of N_x are 0, so we need to investigate the next higher order term. Thus, instead of (3.4) we need to consider

$$\begin{aligned}
&n^{-1}X_x^T W_x X_x \\
&\quad = \left\{ \operatorname{diag}\left(1, H^{1/2}, C^T\right) \right\} \{ f(x) N_x + Q_x \} \left\{ \operatorname{diag}\left(1, H^{1/2}, C\right) \right\} \{1 + o_P(1)\},
\end{aligned}$$

where

$$Q_x = \int K(u) \begin{bmatrix} 0 & u^T & 0 \\ u & 0 & u \operatorname{vech}^T\left(uu^T\right) \\ 0 & \operatorname{vech}\left(uu^T\right)u^T & 0 \end{bmatrix} \left\{ D_f^T(x) H^{1/2} u \right\} du.$$

Since $Q_x = O\{\mathrm{tr}(H^{1/2})\}$,

$$\begin{aligned} &e_1^T\left(n^{-1}X_x^T W_x X_x\right)^{-1} \\ &\quad = f(x)^{-1}e_1^T\{N_x^{-1} - f(x)^{-1}N_x^{-1}Q_x N_x^{-1}\}\mathrm{diag}\left(1, H^{-1/2}, C^{-1}\right) \\ &\qquad + o_P\left\{\mathbf{1}\ \mathrm{diag}\left(1, H^{-1/2}, C^{-1}\right)\right\}. \end{aligned} \tag{3.10}$$

We will now show that $e_1^T N_x^{-1} Q_x N_x^{-1} = 0$. First, since $\nu_x^{12} = 0$, $e_1^T N_x^{-1}$ is orthogonal to the first and the last $\frac{1}{2}d(d+1)$ columns of Q_x. By the following lemma, each of the other columns of Q_x (columns 2 through $d+1$) is a linear combination of the last $\frac{1}{2}d(d+1)$ columns of N_x and therefore is also orthogonal to $e_1^T N_x^{-1}$. This completes the proof that $e_1^T N_x^{-1} Q_x N_x^{-1} = 0$. Therefore, by (3.8)–(3.10), (3.6) holds. The proof of (3.7) is straightforward and will be omitted. □

LEMMA 3.1. *Assume that the conditions of Theorem* 3.2 *hold. Then each of columns* 2 *through to* $d+1$ *of* Q_x *is a linear combination of the last* $\frac{1}{2}d(d+1)$ *columns of* N_x.

PROOF. The proof is simpler if we rearrange the components of $\mathrm{vech}(uu^T)$. Let $u^2 = (u_1^2, \dots, u_d^2)$ and let $q(u)$ be the vector of all distinct pairs $u_i u_{i'}$, $i \neq i'$. Replace $\mathrm{vech}(uu^T)$ by $[(u^2)^T \quad q(u)^T]^T$. With this reordering, columns 2 to $d+1$ of Q_x are

$$\begin{bmatrix} \mu_2(K)1_{1\times d}\mathrm{diag}\{H^{1/2}D_f(x)\} \\ 0_{d\times d} \\ \left[\{\mu_4(K) - \mu_2^2(K)\}I_d + \mu_2^2 1_{d\times d}\right]\mathrm{diag}\{H^{1/2}D_f(x)\} \\ Q_{x,42} \end{bmatrix}, \tag{3.11}$$

where $0_{r\times s}$ and $1_{r\times s}$ are $r \times s$ matrices having each entry to 0 and 1, respectively, I_d is the $d \times d$ identity matrix and $Q_{x,42}$ is a $\frac{1}{2}d(d+1) \times d$ matrix. The last $\frac{1}{2}d(d+1)$ columns of N_x are

$$\begin{bmatrix} \mu_2(K)1_{1\times d} & 0_{1\times d(d-1)/2} \\ 0_{d\times d} & 0_{d\times d(d-1)/2} \\ \{\mu_4(K) - \mu_2(K)^2\}I_d + \mu(K)_2^2 1_{d\times d} & 0_{d\times d(d-1)/2} \\ 0_{d(d-1)/2\times d} & \mu_2^2(K)I_{d(d-1)/2} \end{bmatrix}, \tag{3.12}$$

and (3.11) is (3.12) times $[\mathrm{diag}\{H^{1/2}D_f(x)\} \quad \mu_2^{-2}(K)Q_{x,42}^T]^T$, which completes the proof. □

4. Further extensions.

4.1. *Higher-degree polynomials.* Generalization to higher-degree local polynomial fits is straightforward, although careful notation is important in order to keep expressions simple. The order of the bias is what would be expected.

The general case, which includes the boundary and points where f is discontinuous, is as follows: the conditional bias for pth-order polynomials will be of order $O_P\{(\mathrm{tr}H)^{(p+1)/2}\}$. The reason is that a local pth-degree polynomial fit is exactly conditionally unbiased if m is a pth-degree polynomial. This also implies that the local pth-degree polynomial fit is a kernel estimate with a conditional $(p+1)$th-order kernel, using the obvious generalization of conditional second-order kernel given in Section 2; see Lejeune [(1985), Section 5] for the univariate case with equally spaced X_i's. Moreover, if p is even, f has a continuous derivative in a neighborhood of x and x is an interior point, then the bias will be of order $O_P\{(\mathrm{tr}H)^{(p/2+1)}\}$. However, the bias will depend on $D_f(x)$ so will not be design-adaptive in the sense of Fan (1992).

For local cubic regression under the conditions of Theorem 3.2,

$$E\{\hat{m}(x;H)-m(x)\mid X_1,\dots,X_n\}$$
$$=\frac{1}{4!}\int\left\{\nu_x^{11}+\nu_x^{13}\,\mathrm{vech}\left(uu^T\right)\right\}\left(d_x^4 m\right)\left(H^{1/2}u\right)du+o_P\{\mathrm{tr}(H)^2\},$$

where ν_x^{11} and ν_x^{13} contain the nonzero entries in the first row of N_x^{-1}. For the cubic case $N_x=\int\zeta(u)\zeta(u)^TK(u)\,du$ and $\zeta(u)$ is the $[1+d+\{d+\binom{d}{2}\}+\{d+2\binom{d}{2}+\binom{d}{3}\}]\times 1$ vector containing 1, u and all distinct two-component and three-component products of the entries of u. The conditional variance is the same as (3.7), but with $T_x=\int\zeta(u)\zeta(u)^TK^2(u)\,du$.

For the univariate case and general p, substantial simplification is possible, especially at interior points. Let

$$X_x=\begin{bmatrix}1 & X_1-x & \cdots & (X_1-x)^p\\ \vdots & \vdots & \ddots & \vdots\\ 1 & X_n-x & \cdots & (X_n-x)^p\end{bmatrix}, \tag{4.1}$$

$H=h^2$ and $\mu_j=\int u^jK(u)\,du$, let N_p be the $(p+1)\times(p+1)$ matrix having (i,j)th entry equal to μ_{i+j-2} and let $M_p(u)$ be the same as N_p, but with the first column replaced by $(1,u,\dots,u^p)^T$. Then Lejeune and Sarda (1992) showed that the kernel given by

$$K_{(p)}(u)=\{|M_p(u)|/|N_p|\}K(u)$$

is a kernel of order $p+1$ if p is odd and of order $p+2$ for even p. This higher-order extension of K is particularly relevant to local polynomial fitting as the following theorem reveals.

THEOREM 4.1. ($d=1$). *Suppose that x is an interior point of* supp(f), *that* A1 *and* A2 *hold, that $m^{(p+2)}$ is continuous in a neighborhood of x and that $h\to 0$, $nh\to\infty$ as $n\to\infty$. Let $\hat{m}(x;h)=e_1^T(X_x^TW_xX_x)^{-1}X_x^TW_xY$, where X_x is given by* (4.1). *Then, for p odd,*

$$E\{\hat{m}(x;h)-m(x)\mid X_1,\dots,X_n\}$$
$$=\left\{\int u^{p+1}K_{(p)}(u)\,du\right\}\left\{\frac{m^{(p+1)}(x)}{(p+1)!}\right\}h^{p+1}+o_P\left(h^{p+1}\right) \tag{4.2}$$

and, for p even,

$$E\{\widehat{m}(x;h) - m(x) \mid X_1,\dots,X_n\}$$
$$(4.3)\qquad = \left\{\int u^{p+2}K_{(p)}(u)\,du\right\}\left\{\frac{m^{(p+1)}(x)f'(x)}{f(x)(p+1)!} + \frac{m^{(p+2)}(x)}{(p+2)!}\right\}h^{p+2}$$
$$+\,o_P\big(h^{p+2}\big).$$

In either case

$$\operatorname{Var}\{\widehat{m}(x;h) \mid X_1,\dots,X_n\}$$
$$(4.4)\qquad = \left\{\int K_{(p)}(u)^2\,du\right\}\{n^{-1}h^{-1}v(x)/f(x)\}\{1 + o_P(1)\}.$$

PROOF. First note that

$$E\{\widehat{m}(x;h) - m(x) \mid X_1,\dots,X_n\} = e_1^T\big(n^{-1}X_x^TW_xX_x\big)^{-1}(S_x + R_x),$$

where

$$S_x = n^{-1}X_x^TW_x$$
$$\times\left\{\frac{m^{(p+1)}(x)}{(p+1)!}\begin{bmatrix}(X_1-x)^{p+1}\\ \vdots\\ (X_n-x)^{p+1}\end{bmatrix} + \frac{m^{(p+2)}(x)}{(p+2)!}\begin{bmatrix}(X_1-x)^{p+2}\\ \vdots\\ (X_n-x)^{p+2}\end{bmatrix}\right\}$$

and R_x is a vector of Taylor series remainder terms. We carry two terms in the definition of S_x since the term involving $m^{(p+1)}(x)$ vanishes in the leading term for the conditional bias when p is even. Let $A = \operatorname{diag}(1, h, \dots, h^p)$, and let Q_p be the $(p+1)\times(p+1)$ matrix having (i,j)th entry equal to μ_{i+j-1}. Then

$$n^{-1}X_x^TW_xX_x = A\{f(x)N_p + hf'(x)Q_p\}A + o_P(hA\mathbf{1}A),$$

which leads to

$$(4.5)\qquad e_1^T\big(n^{-1}X_x^TW_xX_x\big)^{-1}$$
$$= f(x)^{-1}\{e_1^TN_p^{-1} - hf'(x)f(x)^{-1}e_1^TN_p^{-1}Q_pN_p^{-1}\}A^{-1} + o_P\big(h\mathbf{1}A^{-1}\big).$$

For $k = 0, 1, \dots,$ standard results from kernel density estimation lead to

$$A^{-1}n^{-1}X_x^TW_x\begin{bmatrix}(X_1-x)^k\\ \vdots\\ (X_n-x)^k\end{bmatrix}$$
$$= \left\{h^kf(x)\begin{bmatrix}\mu_k\\ \mu_{k+1}\\ \vdots\\ \mu_{k+p}\end{bmatrix} + h^{k+1}f'(x)\begin{bmatrix}\mu_{k+1}\\ \mu_{k+2}\\ \vdots\\ \mu_{k+p+1}\end{bmatrix} + o_P\big(h^{k+1}\big)\right\}.$$

Combining this result with (4.5), we obtain

$$\begin{aligned}
&E\{\widehat{m}(x;h) - m(x) \mid X_1,\ldots,X_n\} \\
&\quad = \left\{\sum_{j=1}^{p+1} \left(N_p^{-1}\right)_{1j}\mu_{p+j+1}\right\}\frac{m^{(p+1)}(x)}{(p+1)!}h^{p+1} \\
&\qquad + \left[\left\{\sum_{j=1}^{p+1} \left(N_p^{-1}\right)_{1j}\mu_{p+j+2}\right\}\frac{m^{(p+2)}(x)}{(p+2)!}\right. \\
&\qquad\quad + \left\{\sum_{j=1}^{p+1} \left(N_p^{-1}\right)_{1j}\mu_{p+j+2} - e_1^T N_p^{-1} Q_p N_p^{-1}(\mu_{p+1},\ldots,\mu_{2p+1})^T\right\} \\
&\qquad\qquad \left.\times \frac{m^{(p+1)}(x)f'(x)}{f(x)(p+1)!}\right] h^{p+2} \\
&\qquad + o_P\left(h^{p+2}\right).
\end{aligned} \tag{4.6}$$

To simplify (4.6) for certain cases, note that (a) $\mu_j = 0$ for j odd, (b) $(N_p)_{ij} = (N_p^{-1})_{ij} = 0$ for $i + j$ odd and (c) $(Q_p)_{ij} = 0$ for $i + j$ even. First consider p even. Combining (a) and (c), it is easily shown that the first term of (4.6) vanishes. Also, the first p columns of Q_p are identical to the last p columns of N_p. Combining this with (b) and (c) leads to $e_1^T N_p^{-1} Q_p = 0$, so the last term in (4.6) is also zero. Consequently, for p even,

$$\begin{aligned}
&E\{\widehat{m}(x;h) - m(x) \mid X_1,\ldots,X_n\} \\
&\quad = \left\{\sum_{j=1}^{p+1} \left(N_p^{-1}\right)_{1p}\mu_{p+j+2}\right\}\left\{\frac{m^{(p+1)}(x)f'(x)}{f(x)(p+1)!} + \frac{m^{(p+2)}(x)}{(p+2)!}\right\}h^{p+2} \\
&\qquad + o_P\left(h^{p+2}\right).
\end{aligned}$$

When p is odd the first term in (4.6) does not vanish, so this is the leading conditional bias term.

For the conditional variance we have $n^{-1}X_x^T W_x^2 X_x = h^{-1}f(x)AT_pA + o_P(h^{-1}A\mathbf{1}A)$, where T_p is the $(p+1) \times (p+1)$ matrix having (i,j)th entry equal to $\int u^{i+j-2}K(u)^2\,du$. Combining this with (2.16) and (4.5), we obtain

$$\begin{aligned}
&\operatorname{Var}\{\widehat{m}(x;h) \mid X_1,\ldots,X_n\} \\
&\quad = \left(e_1^T N_p^{-1} T_p N_p^{-1} e_1\right)\left\{n^{-1}h^{-1}v(x)/f(x)\right\}\left\{1 + o_P(1)\right\}.
\end{aligned}$$

It remains to show that the kernel dependent constants match those of (4.2), (4.3) and (4.4). Let c_{ij} denote the cofactor of $(N_p)_{ij}$. Then from the symmetry of N_p and a standard result concerning cofactors we have

$$\left(N_p^{-1}\right)_{1j} = c_{1j}/|N_p|, \qquad j = 1,\ldots,p+1. \tag{4.7}$$

Notice that $\int u^{p+k}K_{(p)}(u)\,du = |G_p|/|N_p|$, where G_p is the same as N_p but with the first column replaced by $(\mu_{p+k},\mu_{p+k+1},\ldots,\mu_{2p+k})^T$. Expanding $|G_p|$ along its first column we obtain

$$\int u^{p+k} K_{(p)}(u)\, du = |N_p|^{-1} \sum_{j=1}^{p+1} \mu_{p+j+k} c_{1j},$$

which, in view of (4.7), gives the first required result. For the second, using (4.7) again, we have

$$\begin{aligned}\int K_{(p)}(u)^2\, du &= |N_p|^{-2} \sum_{i=1}^{p+1} \sum_{j=1}^{p+1} c_{1i} c_{1j} (T_p)_{ij}\\ &= e_1^T N_p^{-1} T_p N_p^{-1} e_1.\end{aligned}$$

□

REMARK 7. Expressions for the bias and variance for boundary points can be easily obtained by reworking the arguments of the proof of Theorem 4.1 with the moments of K replaced by appropriate truncated moments. For example, in the case where $\operatorname{supp}(f) = [0,1]$, $\operatorname{supp}(K) = [-1,1]$ and $x = ch$, $0 < c \leq 1$, define $s_{l,c} = \int_{-c}^{1} u^l K(u)\, du$, $N_p(c)$ to be the $(p+1) \times (p+1)$ matrix having (i,j)th entry equal to $s_{i+j-2,c}$ and $M_p(u,c)$ be the same as $N_p(c)$, but with the first column replaced by $(1, u, \dots, u^p)^T$. The "left-hand boundary" version of the pth-degree Lejeune–Sarda kernel is

$$K_{(p)}(u,c) = \{|M_p(u,c)| / |N_p(c)|\} K(u)$$

and has moment properties analogous to boundary kernels of the type proposed by Gasser, Müller and Mammitzsch (1985). For odd p, precisely the same arguments as those used in the proof of Theorem 4.1 lead to

$$\begin{aligned}&E\{\hat{m}(x;h) - m(x) \mid X_1, \dots, X_n\}\\ &\quad = \left\{ \int_{-c}^{1} u^{p+1} K_{(p)}(u,c)\, du \right\} \left\{ \frac{m^{(p+1)}(x)}{(p+1)!} \right\} h^{p+1} + o_P(h^{p+1})\end{aligned}$$

and

$$\begin{aligned}&\operatorname{Var}\{\hat{m}(x;h) \mid X_1, \dots, X_n\}\\ &\quad = \left\{ \int_{-c}^{1} K_{(p)}(u,c)^2\, du \right\} \{n^{-1} h^{-1} v(x)/f(x)\} \{1 + o_P(1)\},\end{aligned}$$

for $x = ch$. (For $p = 1$ the results given in Remark 3 can be shown to agree with those given here.) Roughly speaking, these results suggest that local polynomial kernel estimators automatically induce a boundary kernel-type bias correction when p is odd.

4.2. *Estimating derivatives of m.* The are many reasons why derivatives of m are of interest. For example, in the study of human growth curves, the first two derivatives of height as a function of age have important biological interpretations [Müller (1988)]. Also, the optimal bandwidth for estimating m

depends upon higher derivatives of m which when estimated lead to "plug-in" rules for bandwidth selection.

Consider the case $d = 1$. For $r \leq p$, the estimate of $m^{(r)}(x)$ will be

$$\widehat{m}_r(x;h) = r!\, e_{r+1}^T \left(X_x^T W_x X_x\right)^{-1} X_x^T W_x Y,$$

where e_{r+1} is the $(p+1) \times 1$ vector with a 1 in its $r+1$ coordinate and 0's elsewhere. Stone (1980, 1982) uses the same method of estimating derivatives, considers the multivariate case and shows optimality in terms of rate of convergence, but he does not study asymptotic bias and variance. Notice that $\widehat{m}_r(x;h)$ is, in general, not equal to $\widehat{m}^{(r)}(x;h)$, the rth derivative of $\widehat{m}(x;h)$—in fact, $\widehat{m}(x;h)$ will not even be differentiable if, say, K is a uniform kernel. One criticism of the Nadaraya–Watson estimator is that it is difficult to analyze the behavior of its derivatives, which makes it somewhat unsuitable for estimating derivatives. The same criticism could be made of local polynomial fitting if we attempted to estimate $m^{(r)}(x)$ by $\widehat{m}^{(r)}(x;h)$. However, the behavior of $\widehat{m}_r(x;h)$ is straightforward to analyze, as we shall now see.

Let N_p be as in Section 4.1, let $M_{r,p}(u)$ be the same as N_p, but with the $(r+1)$th column replaced by $(1, u, \dots, u^p)^T$; and define the kernel

$$K_{(r,p)}(u) = \left\{r!|M_{r,p}(u)|/|N_p|\right\}K(u).$$

It is easily established that $K_{(r,p)}$ satisfies

$$\int u^j K_{(r,p)}(u)\,du = \begin{cases} 0, & 0 \leq j \leq s-1,\ j \neq r, \\ r!, & j = r, \\ \beta_{r,p}, & j = s, \end{cases}$$

where $s = p + 1$ when $p - r$ is odd and $s = p + 2$ when $p - r$ is even and $\beta_{r,p}$ is some non-zero constant. Therefore, $(-1)^r K_{(r,p)}$ is an order (r,s) kernel as defined by Gasser, Müller and Mammitzsch (1985). Such kernels are tailored for estimating rth derivatives of functions such as regression and density functions. For example, $\widehat{f}_r(x;h) = n^{-1}h^{-r-1}\sum_{i=1}^n K_{(r,p)}\,\{(X_i - x)/h\}$ is a consistent estimator for $f^{(r)}(x)$. The proof of Theorem 4.1 can be generalized to give the following theorem.

THEOREM 4.2. ($d = 1$). *Suppose that x is an interior point of* supp(f), *that* A1–A2 *hold, that $m^{(p+2)}$ is continuous in a neighborhood of x and that $h \to 0$, $nh^{2r+1} \to \infty$ as $n \to \infty$. Let $\widehat{m}_r(x;h) = r! e_{r+1}^T (X_x^T W_x X_x)^{-1} X_x^T W_x Y$, where X_x is given by* (4.1). *Then, for $p - r$ odd,*

$$E\{\widehat{m}_r(x;h) - m^{(r)}(x) \mid X_1, \dots, X_n\} = \left\{\int u^{p+1} K_{(r,p)}(u)\,du\right\}\left\{\frac{m^{(p+1)}(x)}{(p+1)!}\right\} h^{p-r+1} + o_P\left(h^{p-r+1}\right)$$

and, for $p - r$ even,

$$E\{\widehat{m}_r(x;h) - m^{(r)}(x) \mid X_1,\dots,X_n\}$$
$$= \left[\left\{\int u^{p+2}K_{(r,p)}(u)\,du\right\}\left\{\frac{m^{(p+2)}(x)}{(p+2)!}\right\} + \left\{\int u^{p+2}K_{(r,p)}(u)\,du - r\int u^{p+1}K_{(r-1,p)}(u)\,du\right\} \times \left\{\frac{m^{(p+1)}(x)f'(x)}{f(x)(p+1)!}\right\}\right]h^{p-r+2} + o_P\left(h^{p-r+2}\right). \tag{4.8}$$

In either case,

$$\operatorname{Var}\{\widehat{m}_r(x;h) \mid X_1,\dots,X_n\} = \left\{\int K_{(r,p)}(u)^2\,du\right\}\left\{n^{-1}h^{-2r-1}v(x)/f(x)\right\}\{1 + o_P(1)\}.$$

PROOF. Since the proof uses essentially the same ideas as those used in the proof of Theorem 4.1 we will restrict attention to the main differences. The extension of (4.5) to general $r \geq 0$ is

$$e_{r+1}^T\left(n^{-1}X_x^TW_xX_x\right)^{-1} = h^{-r}f(x)^{-1}e_{r+1}^TN_p^{-1}A^{-1} - h^{1-r}f'(x)f(x)^{-2}e_{r+1}^T N_p^{-1}Q_pN_p^{-1}A^{-1} + o_P\left(h^{1-r}\mathbf{1}A^{-1}\right). \tag{4.9}$$

When $r = 0$, it was shown in the proof of Theorem 4.1 that the second term vanishes when p is even. However, for $r > 0$ and $p - r$ even,

$$e_{r+1}^TN_p^{-1}Q_pN_p^{-1} = e_r^TN_p^{-1}.$$

This leads to the second kernel dependent term in the right-hand side of (4.8) by applying the same arguments as those given in the proof of Theorem 4.1. When $p - r$ is odd and $r > 0$, $e_{r+1}^TN_p^{-1}Q_pN_p^{-1}$ is also nonzero; however, the resulting corresponding bias term is $O_P(h^{p-r+3})$.

The leading conditional variance term is easily shown to depend only on the first term of (4.9). □

Note that the bias does not involve $f'(x)/f(x)$ if one fits polynomials whose degree p exceeds r by a positive, odd integer. It is also clear that, for such p, $m_r(x;h)$ has a bias of order $O_P(h^{p+1})$ at the boundary as well as in the interior. Thus, the nice boundary behavior of local polynomial fitting extends to the estimation of derivatives. An ordinary kernel estimator with kernel $K_{(r,p)}$ will not behave properly at the boundaries or other locations where f is discontinuous.

Acknowledgments. We thank J. Fan and M. C. Jones for their helpful comments. We are also grateful to a referee for a careful reading of the manuscript and comments which lead to an improved version of the paper. This research was commenced while the second author was in the Department of Statistics, Rice University, Houston, Texas.

REFERENCES

BREIMAN, L. and FRIEDMAN, J. H. (1985). Estimating optimal transformations for multiple regression and correlation (with discussion). *J. Amer. Statist. Assoc.* **80** 580–619.

CHU, C. K. and MARRON, J. S. (1991). Choosing a kernel regression estimator (with discussion). *Statist. Sci.* **6** 404–436.

CLEVELAND, W. (1979). Robust locally weighted regression and smoothing scatterplots. *J. Amer. Statist. Assoc.* **74** 829–836.

CLEVELAND, W. and DEVLIN, S. (1988). Locally weighted regression: an approach to regression analysis by local fitting. *J. Amer. Statist. Assoc.* **83** 596–610.

EUBANK, R. (1988). *Spline Smoothing and Nonparametric Regression.* Dekker, New York.

FAN, J. (1992). Design-adaptive nonparametric regression. *J. Amer. Statist. Assoc.* **87** 998–1004.

FAN, J. (1993). Local linear regression smoothers and their minimax efficiencies. *Ann. Statist.* **21** 196–216.

FAN, J. and GIJBELS, I. (1992). Variable bandwidth and local linear regression smoothers. *Ann. Statist.* **20** 2008–2036.

FRIEDMAN, J. (1991). Multivariate adaptive regression splines (with discussion). *Ann. Statist.* **19** 1–141.

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

GASSER, T. and MÜLLER, H.-G. (1984). Estimating regression functions and their derivatives by the kernel method. *Scand. J. Statist.* **11** 171–185.

GASSER, T., MÜLLER, H.-G. and MAMMITZSCH, V. (1985). Kernels for nonparametric curve estimation. *J. Roy. Statist. Soc. Ser. B* **47** 238–252.

HÄRDLE, W. (1990). *Applied Nonparametric Regression.* Cambridge Univ. Press.

HASTIE, T. and TIBSHIRANI, R. (1986). Generalized additive models (with discussion). *Statist. Sci.* **1** 297–318.

HASTIE, T. and TIBSHIRANI, R. (1990). *Generalized Additive Models.* Chapman and Hall, New York.

LEJEUNE, M. (1985). Estimation non-paramétrique par noyaux: régression polynomial mobile. *Rev. Statist. Appl.* **33** 43–67.

LEJEUNE, M. and SARDA, P. (1992). Smooth estimators of distribution and density functions. *Comput. Statist. Data Anal.* **14** 457–471.

MÜLLER, H.-G. (1987). Weighted local regression and kernel methods for nonparametric curve fitting. *J. Amer. Statist. Assoc.* **82** 231–238.

MÜLLER, H.-G. (1988). *Nonparametric Analysis of Longitudinal Data.* Springer, Berlin.

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

SCHOENBERG, I. (1964). Spline functions and the problem of graduation. *Proc. Nat. Acad. Sci. U.S.A.* **52** 947–950.

STONE, C. J. (1977). Consistent nonparametric regression. *Ann. Statist.* **5** 595–620.

STONE, C. J. (1980). Optimal rates of convergence for nonparametric estimators. *Ann. Statist.* **8** 1348–1360.

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

WAHBA, G. (1990). *Spline Models for Observational Data.* SIAM, Philadelphia.

WAND, M. P. (1990). On exact L_1 rates of convergence in non-parametric kernel regression. *Scand. J. Statist.* **17** 251–256.

WAND, M. P. (1992). Error analyses for general multivariate kernel estimators. *Journal of Nonparametric Statistics* **2** 1–15.

WAND, M. P. and JONES, M. C. (1993). Comparison of smoothing parameterizations in bivariate kernel density estimation. *J. Amer. Statist. Assoc.* **88** 520–528.

WATSON, G. S. (1964). Smooth regression analysis. *Sankhyā Ser. A* **26** 101–116.

SCHOOL OF OPERATIONS RESEARCH
AND INDUSTRIAL ENGINEERING
CORNELL UNIVERSITY
ITHACA, NEW YORK 14853

AUSTRALIAN GRADUATE SCHOOL
OF MANAGEMENT
UNIVERSITY OF NEW SOUTH WALES
KENSINGTON
NEW SOUTH WALES 2033
AUSTRALIA