The Annals of Statistics
1997, Vol. 25, No. 1, 186–211

FITTING A BIVARIATE ADDITIVE MODEL BY LOCAL POLYNOMIAL REGRESSION

BY JEAN D. OPSOMER[1] AND DAVID RUPPERT[2]

Iowa State University and Cornell University

While the additive model is a popular nonparametric regression method, many of its theoretical properties are not well understood, especially when the backfitting algorithm is used for computation of the estimators. This article explores those properties when the additive model is fitted by local polynomial regression. Sufficient conditions guaranteeing the asymptotic existence of unique estimators for the bivariate additive model are given. Asymptotic approximations to the bias and the variance of a homoscedastic bivariate additive model with local polynomial terms of odd and even degree are computed. This model is shown to have the same rate of convergence as that of univariate local polynomial regression.

1. Introduction. Nonparametric regression methods are a flexible and growing class of models in the statistician's toolbox. They allow researchers to evaluate data without having to postulate a shape for the relationship between the response variable and the covariate(s). Unfortunately, nonparametric regression methods become more cumbersome to implement when the number of covariates increases, and the ability to visually inspect estimated relationships is often lost when there are more than two covariates. An elegant solution to these problems is provided by the *additive model*, originally suggested by Friedman and Stuetzle (1981) and popularized by Hastie and Tibshirani (1990). The additive model assumes that the conditional expectation function of the dependent variable Y can be written as a sum of smooth terms in the covariates $X_1, \ldots, X_D$:

$$E(Y|X = (x_1, \ldots, x_D)) = m(x_1, \ldots, x_D) = m_1(x_1) + \cdots + m_D(x_D). \tag{1}$$

The backfitting algorithm proposed by Buja, Hastie and Tibshirani (1989) and the related fitting procedure in S-PLUS [see Chambers and Hastie (1992)] have made the additive model a popular choice for multivariate nonparametric fitting.

Compared to the development of practical applications, the understanding of the theoretical properties of the additive model has lagged. Stone (1985) showed in the case of additive splines that the optimal rate of convergence achievable by additive model estimators is independent of the number of covariates, and Burman (1990) proposed a cross-validation method for select-

Received November 1995; revised May 1996.
[1]Research supported by NIEHS Training Grant ESO7261 while at Cornell University.
[2]Research supported by NSA Grant MDA904-95-H-1025 and NSF Grant DMS-93-06196.
AMS 1991 *subject classifications*. Primary 62G07; secondary 62H99.
Key words and phrases. Additive model, local polynomial regression, optimal rates, existence, backfitting.

ing the number of knots. Other authors have proven existence results in the context of smoothing splines [Wahba (1986), Gu, Bates, Chen and Wahba (1989), and Chen (1993)] and interpolated regressograms [Härdle and Hall (1993)]. More recently, a paper by Linton and Nielsen (1995) describes a fitting procedure for bivariate additive models based on local linear regression and marginal integration, and they showed that their procedure also achieves the same $O_p(n^{-2/5})$ rate of convergence for the additive model as for the univariate local linear estimator. In the case of additive modeling through backfitting, the theoretical investigations are greatly complicated by the fact that the estimators are defined as the solution of an iterative algorithm. Only when two covariates are present are explicit expressions for the estimators available. As Linton and Nielsen (1995) note, however, "these expressions appear quite intractable."

Buja, Hastie and Tibshirani (1989) provide sufficient conditions that guarantee the convergence of the backfitting algorithm or, equivalently, the existence of the estimators. These conditions are only generally satisfied for regression splines and parametric terms, but not by kernel regression or local polynomial regression. This is unfortunate, because local polynomial regression has recently been shown to possess many desirable theoretical and practical properties [e.g., Cleveland and Devlin (1988), Fan, Gasser, Gijbels, Brockmann and Engel (1993), and Ruppert and Wand (1994)] and its combination with backfitting has proven to be very popular for fitting additive models in S-PLUS.

In this article, we will explore two important theoretical issues concerning the bivariate additive model, in the context of backfitted estimators using local polynomial regression:

1. What are sufficient conditions guaranteeing convergence of backfitting?
2. What are the asymptotic properties of the estimators?

In another paper [Opsomer and Ruppert (1995)], these results are used to develop a fully automated plug-in bandwidth selection method.

The rest of the article will proceed as follows. In Section 2, the bivariate additive model estimators are defined. In Section 3, sufficient conditions for the existence of unique estimators are discussed. Section 4 derives conditional asymptotic bias and variance expressions when the local polynomials are of odd degree. Section 5 extends the results to local regression of even degree.

2. Definition of the estimators. Let $(X_1, Z_1, Y_1), \ldots, (X_n, Z_n, Y_n)$ be a set of independent and identically distributed $\mathbb{R}^3$-valued random variables. We assume the following model:

$$Y_i = \alpha + m_1(X_i) + m_2(Z_i) + \varepsilon_i,$$

where the ε_i are independent and identically distributed with mean 0 and variance σ^2. To ensure identifiability of the functions m_1 and m_2, we include the intercept α and assume $E(m_1(X_i)) = E(m_2(Z_i)) = 0$.

We introduce some notation. Let $\boldsymbol{Y} = (Y_1, \ldots, Y_n)^T$ and similarly for $\boldsymbol{X}$ and $\boldsymbol{Z}$, and we write the vectors of additive functions at the observation points as $\boldsymbol{m}_1 = (m_1(X_1), \ldots, m_1(X_n))^T$, $\boldsymbol{m}_2 = (m_2(Z_1), \ldots, m_2(Z_n))^T$. For any constant d, $\boldsymbol{d}$ is the n-valued vector $(d, \ldots, d)^T$. Let $\boldsymbol{s}_{1,x}^T, \boldsymbol{s}_{2,z}^T$ represent the *equivalent kernels* for the local polynomial regression at x and z. In the case of x, this equivalent kernel can be written as

$$\boldsymbol{s}_{1,x}^T = \boldsymbol{e}_1^T(\boldsymbol{X}_x^T\boldsymbol{W}_x\boldsymbol{X}_x)^{-1}\boldsymbol{X}_x^T\boldsymbol{W}_x,$$

where $\boldsymbol{e}_1^T = (1, 0)$,

$$W_x = \operatorname{diag}\left\{\frac{1}{h_1}K\left(\frac{X_1 - x}{h_1}\right), \ldots, \frac{1}{h_1}K\left(\frac{X_n - x}{h_1}\right)\right\}$$

for some kernel function K and bandwidth h_1 and

$$\boldsymbol{X}_x = \begin{bmatrix} 1 & (X_1 - x) & \cdots & (X_1 - x)^{p_1} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & (X_n - x) & \cdots & (X_n - x)^{p_1} \end{bmatrix},$$

where p_1 is the order of the local polynomials for fitting m_1 [see Ruppert and Wand (1994)]. A similar expression holds for $\boldsymbol{s}_{2,z}^T$. Let $\boldsymbol{S}_1$ and $\boldsymbol{S}_2$ represent the smoother matrices whose rows are the equivalent kernels at the observations $\boldsymbol{X}$ and $\boldsymbol{Z}$, respectively:

$$\boldsymbol{S}_1 = \begin{bmatrix} \boldsymbol{s}_{1,X_1}^T \\ \vdots \\ \boldsymbol{s}_{1,X_n}^T \end{bmatrix}, \qquad \boldsymbol{S}_2 = \begin{bmatrix} \boldsymbol{s}_{2,Z_1}^T \\ \vdots \\ \boldsymbol{s}_{2,Z_n}^T \end{bmatrix}.$$

We define the vector of fitted values at the observation points as

$$\hat{\boldsymbol{m}} = \hat{\boldsymbol{\alpha}} + \hat{\boldsymbol{m}}_1 + \hat{\boldsymbol{m}}_2,$$

in which $\hat{\alpha} = \bar{Y}$, and $\hat{\boldsymbol{m}}_1$ and $\hat{\boldsymbol{m}}_2$ are the solutions to the set of estimating equations

$$\begin{bmatrix} \boldsymbol{I} & \boldsymbol{S}_1^* \\ \boldsymbol{S}_2^* & \boldsymbol{I} \end{bmatrix}\begin{bmatrix} \hat{\boldsymbol{m}}_1 \\ \hat{\boldsymbol{m}}_2 \end{bmatrix} = \begin{bmatrix} \boldsymbol{S}_1^* \\ \boldsymbol{S}_2^* \end{bmatrix}\boldsymbol{Y},$$

where $\boldsymbol{S}_1^* = (\boldsymbol{I} - \boldsymbol{1}\boldsymbol{1}^T/n)\boldsymbol{S}_1$ and similarly for $\boldsymbol{S}_2^*$. As discussed in Hastie and Tibshirani (1990), this adjustment of the smoothers, which they refer to as *centering*, is necessary to ensure uniqueness of the solutions to the estimating equations (if they exist), by requiring $\sum_{i=1}^n m_1(X_i) = \sum_{i=1}^n m_2(Z_i) = 0$. In practice, the estimating equations are solved using the backfitting algorithm, but in the bivariate case they also have the explicit solution

$$\begin{aligned} \hat{\boldsymbol{m}}_1 &= \{\boldsymbol{I} - (\boldsymbol{I} - \boldsymbol{S}_1^*\boldsymbol{S}_2^*)^{-1}(\boldsymbol{I} - \boldsymbol{S}_1^*)\}\boldsymbol{Y} \equiv \boldsymbol{W}_1\boldsymbol{Y}, \\ \hat{\boldsymbol{m}}_2 &= \{\boldsymbol{I} - (\boldsymbol{I} - \boldsymbol{S}_2^*\boldsymbol{S}_1^*)^{-1}(\boldsymbol{I} - \boldsymbol{S}_2^*)\}\boldsymbol{Y} \equiv \boldsymbol{W}_2\boldsymbol{Y}, \end{aligned} \tag{2}$$

provided the inverses exist. For $\hat{\boldsymbol{m}}$, the explicit estimator expression is

$$\hat{\boldsymbol{m}} = \{\mathbf{1}\mathbf{1}^T/n + 2\boldsymbol{I} - (\boldsymbol{I} - \boldsymbol{S}_1^*\boldsymbol{S}_2^*)^{-1}(\boldsymbol{I} - \boldsymbol{S}_1^*) - (\boldsymbol{I} - \boldsymbol{S}_2^*\boldsymbol{S}_1^*)^{-1}(\boldsymbol{I} - \boldsymbol{S}_2^*)\}\boldsymbol{Y} \equiv \boldsymbol{W}\boldsymbol{Y}. \tag{3}$$

3. Existence. We will specify a set of circumstances under which the estimators in (2) and (3) are guaranteed to exist. Let $f(x, z)$ represent the joint density of X_i and Z_i, with $f_X(x)$ and $f_Z(z)$ the corresponding marginal densities. For the kernel function K, we write the moments of K as $\mu_j(K) = \int u^j K(u)\,du$ for any j and let $R(K) = \int K(u)^2\,du$.

One of the important issues in the theoretical derivations in this and following sections is whether or not an observation (X_i, Z_i) is close to the boundary of its domain. Using the notation of Ruppert and Wand (1994), we can formalize this by defining

$$D_{x,h_1} = \{t\colon (x + h_1 t) \in \operatorname{supp}(f_X)\} \cap \operatorname{supp}(K).$$

We then say that x is an *interior point* if and only if $D_{x,h_1} = \operatorname{supp}(K)$. Otherwise, x is a *boundary point*. Another way to understand this distinction is depicted in Figure 1: for a given bandwidth value h_1, x_1 is a boundary point, because there are values of x for which $K((x - x_1)/h_1) \neq 0$ outside of $\operatorname{supp}(f_X) = [a, b]$, while x_2 is an interior point, because all values for which $K((x - x_2)/h_2) \neq 0$ are inside that support. Analogous definitions hold for D_{z,h_2} and z. We define the *boundary moments* of K with respect to x as

$$\mu_j(K, x) = \int_{D_{x,h_1}} u^j K(u)\,du,$$

where the dependency on h_1 will be suppressed for notational simplicity, and similarly for $R(K, x)$. Clearly, if x is an interior point, $\mu_j(K, x) = \mu_j(K)$ and $R(K, x) = R(K)$. Let $\boldsymbol{N}_p$ represent the $(p+1) \times (p+1)$ matrix whose (i, j)th element is equal to $\mu_{i+j-2}(K)$ and $\boldsymbol{M}_p(u)$ be the same as $\boldsymbol{N}_p$, but with the first column replaced by $(1, u, \ldots, u^p)^T$. As in Ruppert and Wand (1994), define the kernel

$$K_{(p)}(u) = \{|\boldsymbol{M}_p(u)|/|\boldsymbol{N}_p|\}K(u).$$

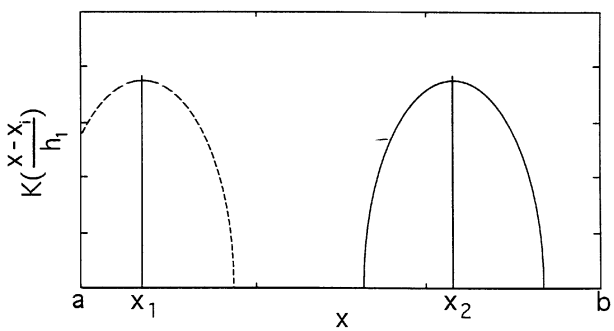

FIG. 1. *Graphical representation of a boundary point* (x_1) *and an interior point* (x_2) *for a kernel function* K *and* $\operatorname{supp}(f_X) = [a, b]$.

For the boundary moments, we define the matrices $\boldsymbol{N}_p(x)$ and $\boldsymbol{M}_p(u, x)$ exactly as above, but with the $\mu_j(K)$ replaced by $\mu_j(K, x)$. We also define the boundary kernel

$$K_{(p)}(u, x) = \{|\boldsymbol{M}_p(u, x)|/|\boldsymbol{N}_p(x)|\}K(u)I_{(u\in D_{x,h_1})}.$$

If x is not at the boundary of $\mathrm{supp}(f)$, $K_{(p)}(u, x) = K_{(p)}(u)$. Analogous definitions hold for the other covariate z.

We make the following assumptions.

ASSUMPTION 1. The kernel K is bounded and continuous, it has compact support and its first derivative has a finite number of sign changes over its support. Also, $\mu_j(K) = 0$ for all odd j and $\mu_{p_1+1}(K_{(p_1)}), \mu_{p_2+1}(K_{(p_2)}) \neq 0$.

ASSUMPTION 2. The densities f, f_X and f_Z are bounded and continuous, have compact support and their first derivatives have a finite number of sign changes over their supports. Also, $f_X(x) > 0$, $f_Z(z) > 0$ for all $(x, z) \in \mathrm{supp}(f)$ and

$$\sup_{x,z}\left|\frac{f(x, z)}{f_X(x)f_Z(z)} - 1\right| < 1. \tag{4}$$

ASSUMPTION 3. As $n \to \infty$, $h_1, h_2 \to 0$ and $nh_1/\log n, nh_2/\log n \to \infty$.

The following two lemmas show that, under these assumptions, the matrix inverses in the estimators (2) and (3) are well defined for local polynomials of any degree p_1 and p_2 "for sufficiently large n." Strictly speaking, the lemmas only prove the existence of the estimator $\hat{\boldsymbol{m}}_1$, but it is clear that the results also hold for $\hat{\boldsymbol{m}}_2$ and $\hat{\boldsymbol{m}}$. The proofs are given in Appendix A.

LEMMA 3.1. *Under Assumptions* 1–3, *the following asymptotic approximations hold uniformly over all elements of the matrices*:

$$\boldsymbol{S}_1^* = \boldsymbol{S}_1 - \mathbf{1}\mathbf{1}^T/n + o(\mathbf{1}\mathbf{1}^T/n) \quad a.s.,$$
$$\boldsymbol{S}_1^*\boldsymbol{S}_2^* = \boldsymbol{T}_{12}^* + o(\mathbf{1}\mathbf{1}^T/n) \quad a.s.,$$

where $\boldsymbol{T}_{12}^*$ *is a matrix whose* (i, j)*th element is*

$$[\boldsymbol{T}_{12}^*]_{ij} = \frac{1}{n}\frac{f(X_i, Z_j)}{f_X(X_i)f_Z(Z_j)} - \frac{1}{n}.$$

LEMMA 3.2. *If Assumptions* 1–3 *hold, then* $(\boldsymbol{I} - \boldsymbol{T}_{12}^*)$ *is invertible for all* n *and*

$$\mathrm{P}\{\textit{there exists } N \textit{ such that } (\boldsymbol{I} - \boldsymbol{S}_1^*\boldsymbol{S}_2^*) \textit{ is invertible for all } n \geq N\} = 1.$$

When $(\boldsymbol{I} - \boldsymbol{S}_1^*\boldsymbol{S}_2^*)^{-1}$ *exists*,

$$\begin{aligned}(\boldsymbol{I} - \boldsymbol{S}_1^*\boldsymbol{S}_2^*)^{-1} &= (\boldsymbol{I} - \boldsymbol{T}_{12}^*)^{-1} + o(\mathbf{1}\mathbf{1}^T/n) \quad a.s.\\ &= \boldsymbol{I} + O(\mathbf{1}\mathbf{1}^T/n) \quad a.s.\end{aligned}$$

uniformly over all elements of the matrices.

REMARK 3.1. The restriction (4) in Assumption 2 ensures that $\|\boldsymbol{T}_{12}^*\|_r < 1$, where $\|\boldsymbol{A}\|_r$ denotes the *maximum row sum matrix norm* of the square matrix $\boldsymbol{A}$: $\|\boldsymbol{A}\|_r = \max_{1\le i\le n} \sum_{j=1}^n |A_{ij}|$ [Horn and Johnson (1985)]. The main reason for selecting this norm was convenience, and it is clear that any other norm $\|\cdot\|$ which ensures that $(\boldsymbol{I} - \boldsymbol{T}_{12}^*)$ is invertible whenever $\|\boldsymbol{T}_{12}^*\| < 1$ would be equally appropriate. To assess the restrictiveness of (4), let us evaluate its effect on the bivariate normal distribution with censored support. For simplicity, assume that the mean of the distribution is at the center of the range for both covariates and that the standardized ranges (i.e., range divided by standard deviation) are the same in both dimensions. Let r represent this standardized range for both covariates. We can then rewrite (4) as

$$\sup_{-r/2\le x,\, z\le r/2} \left| \frac{1}{\sqrt{1-\rho^2}} \exp\left(-\frac{\rho}{2(1-\rho^2)} (\rho x^2 - 2xz + \rho z^2) \right) - 1 \right| < 1. \tag{5}$$

Figure 2 displays the values of the correlation coefficient ρ for which (5) is satisfied as a function of the standardized range r. Clearly, there is a tradeoff between the ranges of X_i and Z_i (in units of their standard deviations)

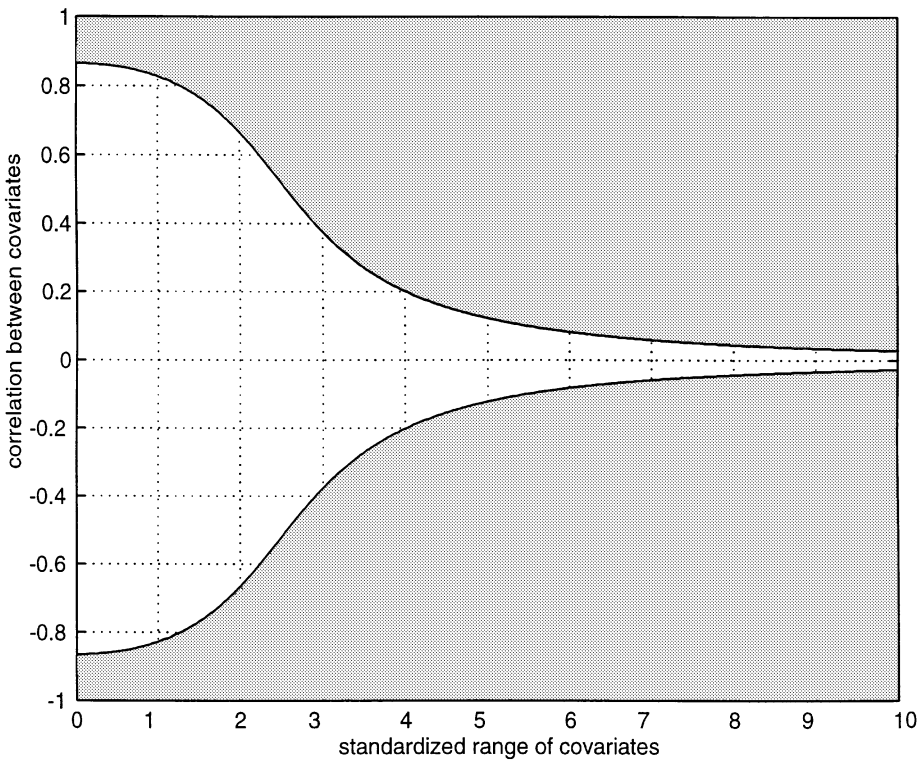

FIG. 2. *Allowable values for the correlation between covariates following a censored normal distribution* (*unshaded region*).

and the correlation. For instance, for a range of 3 standard deviation units, the correlation can take on values $-0.37 \leq \rho \leq 0.37$. While this may seem quite restrictive, it is important to realize that the constraint (4) is a *sufficient* condition for the existence of the estimators, not a necessary one. Note also that when the ranges of X_i and Z_i are very small relative to their standard deviation, that is, when the distribution is almost uniform, the amount of allowable correlation approaches $(-0.9, 0.9)$.

REMARK 3.2. Assumptions 1–3 are also somewhat stronger than ones usually made for local polynomial regression. Specifically, the kernel function and the density of the covariates have additional "smoothness" restrictions, and the maximum rate at which the bandwidths h_1, h_2 approach 0 is slightly slower. The purpose behind these restrictions is to allow us to use the uniform convergence results of Pollard (1984) in the proof of Lemma 3.1. They will not significantly affect the applicability of the results, since many commonly used kernels (including the Epanečnikov kernel) easily satisfy these conditions, and since the optimal rates of h_1, h_2 are fractional powers of n and therefore unaffected by the presence of the $\log n$ term in Assumption 3.

4. Conditional mean-squared error properties. As shown in Theorem 4.1 of Ruppert and Wand (1994), if the degree p of the local polynomial is even, the estimator has asymptotic bias of order $O_p(h^{p+2})$ in the interior, which is the same order as the estimator computed by local polynomial regression of (odd) degree $p+1$. The asymptotic bias of the former estimator also contains an additional term. For these reasons, several authors [e.g., Fan and Gijbels (1994)] have argued that odd-degree local polynomials are preferable to even-degree ones. In this section, we will therefore restrict our attention to the case where p_1 and p_2 are odd. We will briefly discuss the situation where p_1 and p_2 are even in Section 5.

Let

$$D^p m_1 = \begin{bmatrix} \dfrac{d^p m_1(X_1)}{dx^p} \\ \vdots \\ \dfrac{d^p m_1(X_n)}{dx^p} \end{bmatrix}$$

and

$$E(m_1^{(p)}(X_i)|\boldsymbol{Z}) = \begin{bmatrix} E(m_1^{(p)}(X_1)|Z_1) \\ \vdots \\ E(m_1^{(p)}(X_n)|Z_n) \end{bmatrix}$$

and analogously for $D^p \boldsymbol{m}_2$ and $E(m_2^{(p)}(Z_i)|\boldsymbol{X})$. Also, let $\boldsymbol{t}_i^T, \boldsymbol{v}_j$ represent the ith row and jth column of $(\boldsymbol{I} - \boldsymbol{T}_{12}^*)^{-1}$, respectively, and let $\boldsymbol{e}_i^T$ represent the ith unit vector.

In addition to Assumptions 1–3 from Section 3, we also need the following assumption.

ASSUMPTION 4. The (p_1+1)th derivative of m_1 and the (p_2+1)th derivative of m_2 exist and are continuous and bounded.

In the theorem and corollaries that follow, we will only show the results for $\hat{\alpha}$, $\hat{m}_1$ and $\hat{m}$. It is clear that the results for $\hat{m}_2$ can be found by interchanging X_i and Z_i and the subscripts 1 and 2 in those for $\hat{m}_1$.

THEOREM 4.1. *Suppose that Assumptions* 1–4 *hold. For the observation points* (X_i, Z_i), $i = 1, \ldots, n$, *the conditional bias and variance of* $\hat{\alpha}$ *and* $\hat{m}_1(X_i)$ *can be approximated by*

$$E(\hat{\alpha} - \alpha | \boldsymbol{X}, \boldsymbol{Z}) = \alpha + O_p\left(\frac{1}{\sqrt{n}}\right),$$

$$\begin{aligned} &E(\hat{m}_1(X_i) - m_1(X_i) | \boldsymbol{X}, \boldsymbol{Z}) \\ &\quad = \frac{1}{(p_1+1)!} h_1^{p_1+1} \mu_{p_1+1}(K_{(p_1)}, X_i) m_1^{(p_1+1)}(X_i) \\ &\qquad + \frac{1}{(p_1+1)!} h_1^{p_1+1} \mu_{p_1+1}(K_{(p_1)}) \left((\boldsymbol{t}_i^T - \boldsymbol{e}_i^T) D^{p_1+1} \boldsymbol{m}_1 - E(m_1^{(p_1+1)}(X_i)) \right) \\ &\qquad - \frac{1}{(p_2+1)!} h_2^{p_2+1} \mu_{p_2+1}(K_{(p_2)}) \left(\boldsymbol{t}_i^T E(m_2^{(p_2+1)}(Z_i) | \boldsymbol{X}) - E(m_2^{(p_2+1)}(Z_i)) \right) \\ &\qquad + O_p\left(\frac{1}{\sqrt{n}}\right) + o_p(h_1^{p_1+1} + h_2^{p_2+1}) \end{aligned}$$

and

$$\begin{aligned} \mathrm{Var}(\hat{\alpha}) &= \frac{\sigma^2}{n}, \\ \mathrm{Var}(\hat{m}_1(X_i) | \boldsymbol{X}, \boldsymbol{Z}) &= \sigma^2 \frac{R(K_{(p_1)}, X_i)}{n h_1} f_X(X_i)^{-1} + o_p\left(\frac{1}{n h_1}\right). \end{aligned}$$

The conditional bias and variance of $\hat{m}(X_i, Z_i)$ *are*

$$\begin{aligned} &E(\hat{m}(X_i, Z_i) - m(X_i, Z_i) | \boldsymbol{X}, \boldsymbol{Z}) \\ &\quad = \frac{1}{(p_1+1)!} h_1^{p_1+1} \Big(\mu_{p_1+1}(K_{(p_1)}, X_i) m_1^{(p_1+1)}(X_i) + \mu_{p_1+1}(K_{(p_1)}) \\ &\qquad\qquad \times \left((\boldsymbol{t}_i^T - \boldsymbol{e}_i^T) D^{p_1+1} \boldsymbol{m}_1 - \boldsymbol{v}_i^T E(m_1^{(p_1+1)}(X_i) | \boldsymbol{Z}) \right) \Big) \\ &\qquad + \frac{1}{(p_2+1)!} h_2^{p_2+1} \Big(\mu_{p_2+1}(K_{(p_2)}, Z_i) m_2^{(p_2+1)}(Z_i) + \mu_{p_2+1}(K_{(p_2)}) \\ &\qquad\qquad \times \left((\boldsymbol{t}_i^T - \boldsymbol{e}_i^T) D^{p_2+1} \boldsymbol{m}_2 - \boldsymbol{v}_i^T E(m_2^{(p_2+1)}(Z_i) | \boldsymbol{X}) \right) \Big) \\ &\qquad + o_p(h_1^{p_1+1} + h_2^{p_2+1}) \end{aligned}$$

and

$$\begin{aligned}\operatorname{Var}(\hat{m}(X_i, Z_i)|\boldsymbol{X}, \boldsymbol{Z}) &= \sigma^2\left(\frac{R(K_{(p_1)}, X_i)}{nh_1} f_X(X_i)^{-1} + \frac{R(K_{(p_2)}, Z_i)}{nh_2} f_Z(Z_i)^{-1}\right) \\ &\quad + o_p\left(\frac{1}{nh_1} + \frac{1}{nh_2}\right).\end{aligned}$$

COROLLARY 4.1. *If the observation point* (X_i, Z_i) *lies in the interior of* $\operatorname{supp}(f)$, *the conditional bias and variance of* $\hat{m}_1(X_i)$ *are approximated by*

$$\begin{aligned}E(\hat{m}_1(X_i) - m_1(X_i)|\boldsymbol{X}, \boldsymbol{Z}) &= \frac{1}{(p_1+1)!} h_1^{p_1+1} \mu_{p_1+1}(K_{(p_1)})\left(\boldsymbol{t}_i^T D^{p_1+1} \boldsymbol{m}_1 - E(m_1^{(p_1+1)}(X_i))\right) \\ &\quad - \frac{1}{(p_2+1)!} h_2^{p_2+1} \mu_{p_2+1}(K_{(p_2)})\left(\boldsymbol{t}_i^T E(m_2^{(p_2+1)}(Z_i)|\boldsymbol{X}) - E(m_2^{(p_2+1)}(Z_i))\right) \\ &\quad + O_p\left(\frac{1}{\sqrt{n}}\right) + o_p(h_1^{p_1+1} + h_2^{p_2+1})\end{aligned}$$

and

$$\operatorname{Var}(\hat{m}_1(X_i)|\boldsymbol{X}, \boldsymbol{Z}) = \sigma^2 \frac{R(K_{(p_1)})}{nh_1} f_X(X_i)^{-1} + o_p\left(\frac{1}{nh_1}\right).$$

The conditional bias and variance of $\hat{m}(X_i, Z_i)$ *are*

$$\begin{aligned}E(\hat{m}(X_i, Z_i) - m(X_i, Z_i)|\boldsymbol{X}, \boldsymbol{Z}) &= \frac{1}{(p_1+1)!} h_1^{p_1+1} \mu_{p_1+1}(K_{(p_1)})\left(\boldsymbol{t}_i^T D^{p_1+1} \boldsymbol{m}_1 - \boldsymbol{v}_i^T E(m_1^{(p_1+1)}(X_i)|\boldsymbol{Z})\right) \\ &\quad + \frac{1}{(p_2+1)!} h_2^{p_2+1} \mu_{p_2+1}(K_{(p_2)})\left(\boldsymbol{v}_i^T D^{p_2+1} \boldsymbol{m}_2 - \boldsymbol{t}_i^T E(m_2^{(p_2+1)}(Z_i)|\boldsymbol{X})\right) \\ &\quad + o_p(h_1^{p_1+1} + h_2^{p_2+1})\end{aligned}$$

and

$$\begin{aligned}\operatorname{Var}(\hat{m}(X_i, Z_i)|\boldsymbol{X}, \boldsymbol{Z}) &= \sigma^2\left(\frac{R(K_{(p_1)})}{nh_1} f_X(X_i)^{-1} + \frac{R(K_{(p_2)})}{nh_2} f_Z(Z_i)^{-1}\right) \\ &\quad + o_p\left(\frac{1}{nh_1} + \frac{1}{nh_2}\right).\end{aligned}$$

A convenient error criterion that only uses the fitted values at the observation points is provided by the conditional mean averaged squared error (MASE), discussed by Härdle, Hall and Marron (1988). The MASE of m can

be written as

$$\begin{aligned}\mathrm{MASE}(h_1, h_2|\boldsymbol{X}, \boldsymbol{Z}) &= \frac{1}{n}\sum_{i=1}^{n} E\{(\hat{m}(X_i, Z_i) - m(X_i, Z_i)|\boldsymbol{X}, \boldsymbol{Z})^2\} \\ &= \frac{1}{n}\sum_{i=1}^{n}\{E(\hat{m}(X_i, Z_i) - m(X_i, Z_i)|\boldsymbol{X}, \boldsymbol{Z})\}^2 \\ &\quad + \frac{1}{n}\sum_{i=1}^{n}\mathrm{Var}(\hat{m}(X_i, Z_i)|\boldsymbol{X}, \boldsymbol{Z})\end{aligned}$$

and its asymptotic approximation is easily constructed from the previous results. To simplify the notation, let

$$\begin{aligned}\theta_{11}(r) &= \frac{1}{n}\sum_{i=1}^{n}\big(\boldsymbol{t}_i^T D^r \boldsymbol{m}_1 - \boldsymbol{v}_i^T E(m_1^{(r)}(X_i)|\boldsymbol{Z})\big)^2, \\ \theta_{22}(r) &= \frac{1}{n}\sum_{i=1}^{n}\big(\boldsymbol{v}_i^T D^r \boldsymbol{m}_2 - \boldsymbol{t}_i^T E(m_2^{(r)}(Z_i)|\boldsymbol{X})\big)^2\end{aligned}$$

and

$$\theta_{12}(r, s) = \frac{1}{n}\sum_{i=1}^{n}\big(\boldsymbol{t}_i^T D^r \boldsymbol{m}_1 - \boldsymbol{v}_i^T E(m_1^{(r)}(X_i)|\boldsymbol{Z})\big)\big(\boldsymbol{v}_i^T D^s \boldsymbol{m}_2 - \boldsymbol{t}_i^T E(m_2^{(s)}(Z_i)|\boldsymbol{X})\big).$$

COROLLARY 4.2. *The conditional MASE for the bivariate additive model is approximated by*

$$\begin{aligned}&\mathrm{MASE}(h_1, h_2|\boldsymbol{X}, \boldsymbol{Z}) \\ &\quad = \left(\frac{\mu_{p_1+1}(K_{(p_1)})}{(p_1+1)!}\right)^2 h_1^{2p_1+2}\theta_{11}(p_1+1) + \left(\frac{\mu_{p_2+1}(K_{(p_2)})}{(p_2+1)!}\right)^2 h_2^{2p_2+2}\theta_{22}(p_2+1) \\ &\qquad + \frac{\mu_{p_1+1}(K_{(p_1)})}{(p_1+1)!}\frac{\mu_{p_2+1}(K_{(p_2)})}{(p_2+1)!}h_1^{p_1+1}h_2^{p_2+1}\theta_{12}(p_1+1, p_2+1) \\ &\qquad + \sigma^2\frac{1}{n}\sum_{i=1}^{n}\left(\frac{R(K_{(p_1)})}{nh_1}f_X(X_i)^{-1} + \frac{R(K_{(p_2)})}{nh_2}f_Z(Z_i)^{-1}\right) \\ &\qquad + o_p\left(h_1^{2p_1+2} + h_2^{2p_2+2} + \frac{1}{nh_1} + \frac{1}{nh_2}\right).\end{aligned}$$

If $\boldsymbol{X}$ and $\boldsymbol{Z}$ are independent, the preceding results can be simplified significantly. The expression for $\theta_{11}(r)$ becomes

$$\theta_{11}(r) = \frac{1}{n}\sum_{i=1}^{n}\big(m_1^{(r)}(X_i) - E(m_1^{(r)}(X_i))\big)^2$$

and similarly for $\theta_{22}(r)$. Since $\theta_{12}(r, s) = O_p(1/n)$ in this case, the term can be ignored.

COROLLARY 4.3. *If $\boldsymbol{X}$ and $\boldsymbol{Z}$ are independent, the conditional bias and variance of $\hat{\alpha}$ and $\hat{m}_1(X_i)$ in the interior of* $\mathrm{supp}(f)$ *can be approximated by*

$$E(\hat{\alpha} - \alpha | \boldsymbol{X}, \boldsymbol{Z}) = \alpha + O_p\left(\frac{1}{\sqrt{n}}\right),$$

$$\begin{aligned} &E(\hat{m}_1(X_i) - m_1(X_i) | \boldsymbol{X}, \boldsymbol{Z}) \\ &\quad = \frac{1}{(p_1+1)!} h_1^{p_1+1} \mu_{p_1+1}(K_{(p_1)}) \left(m_1^{(p_1+1)}(X_i) - E(m_1^{(p_1+1)}(X_i)) \right) \\ &\qquad + O_p\left(\frac{1}{\sqrt{n}}\right) + o_p(h_1^{p_1+1} + h_2^{p_2+1}) \end{aligned}$$

and

$$\mathrm{Var}(\hat{\alpha}) = \frac{\sigma^2}{n},$$

$$\mathrm{Var}(\hat{m}_1(X_i) | \boldsymbol{X}, \boldsymbol{Z}) = \sigma^2 \frac{R(K_{(p_1)})}{n h_1} f_X(X_i)^{-1} + o_p\left(\frac{1}{n h_1}\right).$$

For $\hat{m}(X_i, Z_i)$,

$$\begin{aligned} &E(\hat{m}(X_i, Z_i) - m(X_i, Z_i) | \boldsymbol{X}, \boldsymbol{Z}) \\ &\quad = \frac{1}{(p_1+1)!} h_1^{p_1+1} \mu_{p_1+1}(K_{(p_1)}) \left(m_1^{(p_1+1)}(X_i) - E(m_1^{(p_1+1)}(X_i)) \right) \\ &\qquad + \frac{1}{(p_2+1)!} h_2^{p_2+1} \mu_{p_2+1}(K_{(p_2)}) \left(m_2^{(p_2+1)}(Z_i) - E(m_2^{(p_2+1)}(Z_i)) \right) \\ &\qquad + o_p(h_1^{p_1+1} + h_2^{p_2+1}), \end{aligned}$$

$$\begin{aligned} \mathrm{Var}(\hat{m}(X_i, Z_i) | \boldsymbol{X}, \boldsymbol{Z}) = \sigma^2 &\left(\frac{R(K_{(p_1)})}{n h_1} f_X(X_i)^{-1} + \frac{R(K_{(p_2)})}{n h_2} f_Z(Z_i)^{-1} \right) \\ &+ o_p\left(\frac{1}{n h_1} + \frac{1}{n h_2} \right) \end{aligned}$$

and the conditional MASE is

$$\begin{aligned} &\mathrm{MASE}(h_1, h_2 | \boldsymbol{X}, \boldsymbol{Z}) \\ &\quad = \left(\frac{\mu_{p_1+1}(K_{(p_1)})}{(p_1+1)!} \right)^2 h_1^{2p_1+2} \theta_{11}(p_1+1) + \left(\frac{\mu_{p_2+1}(K_{(p_2)})}{(p_2+1)!} \right)^2 h_2^{2p_2+2} \theta_{22}(p_2+1) \\ &\qquad + \sigma^2 \frac{1}{n} \sum_{i=1}^{n} \left(\frac{R(K_{(p_1)})}{n h_1} f_X(X_i)^{-1} + \frac{R(K_{(p_2)})}{n h_2} f_Z(Z_i)^{-1} \right) \\ &\qquad + o_p\left(h_1^{2p_1+2} + h_2^{2p_2+2} + \frac{1}{n h_1} + \frac{1}{n h_2} \right). \end{aligned}$$

REMARK 4.1. Theorem 4.1 shows that the additive model fitted by local polynomial regression of degree $p_1 = p_2 = p$ and computed through backfitting has the same $O_p(n^{-(p+1)/(2p+3)})$ rate of convergence as univariate local polynomial regression. However, there are a number of interesting differences between the asymptotic bias of the terms of the additive model and that of the (nonadditive) local polynomial regression estimator. In the case of $\hat{\boldsymbol{m}}_1$, the bias contains a term based on the curvature of m_1,

$$-\frac{1}{(p_1+1)!}\mu_{p_1+1}(K_{(p_1)})h_1^{p_1+1}(\boldsymbol{t}_i^T E(m_1^{(p_1+1)}(X_i)|\boldsymbol{Z}) - E(m_1^{(p_1+1)}(X_i))),$$

another term based on the curvature of m_2,

$$-\frac{1}{(p_2+1)!}\mu_{p_2+1}(K_{(p_2)})h_2^{p_2+1}(\boldsymbol{t}_i^T E(m_2^{p_2+1}(Z_i)|\boldsymbol{X}) - E(m_2^{p_2+1}(Z_i)))$$

and a third term caused only by the centering adjustment of $\boldsymbol{S}_1^*$, $O_p(1/\sqrt{n})$ (but note that the first two terms are also centered around their means). Only the first of these terms has an equivalence in univariate local polynomial regression. As shown in Corollary 4.3, the m_2 curvature term disappears if $\boldsymbol{X}$ and $\boldsymbol{Z}$ are independent, so that the asymptotic bias of $\hat{m}_1$ no longer depends on m_2. The term $O_p(1/\sqrt{n})$ is not related to the dependence between $\boldsymbol{X}$ and $\boldsymbol{Z}$. As shown in the proof of Theorem 4.1 (below), this centering bias of $\hat{m}_1(X_i)$ and $\hat{m}_2(Z_i)$ cancels out with the bias of $\hat{\alpha}$, so that this term does not appear in the bias of $\hat{m}(X_i, Z_i)$. Another difference with local polynomial regression is that the asymptotic bias at a point (X_i, Z_i) not only depends on the curvature of m_1 and m_2 *at that point*, but is a weighted average of the curvature at *all* the observation points, with the weights determined by the matrix $(\boldsymbol{I}-\boldsymbol{T}_{12}^*)^{-1}$. This difference again disappears when $\boldsymbol{X}$ and $\boldsymbol{Z}$ are independent.

REMARK 4.2. If $\boldsymbol{X}$ and $\boldsymbol{Z}$ are independent, Corollary 4.3 shows that $\hat{m}$ has another interesting property. Suppose, for simplicity, that we are fitting the additive model by local polynomial regression of degree p. This additive model is unbiased as long as the unknown functions m_1 and m_2 are polynomials of degree less than or equal to $p+1$. This differs from nonadditive local polynomial regression of degree p, which is only unbiased when the unknown function itself is of degree less than or equal to p. This effect is due to the centering adjustment, which replaces $m_1^{(p+1)}(X_i)$ in the nonadditive local polynomial regression bias by $(m_1^{(p+1)}(X_i) - E(m_1^{(p+1)}(X_i)))$. It is easy to see that the "centered" derivatives are indeed 0 for polynomials up to degree $p+1$.

REMARK 4.3. Unlike the bias, the asymptotic variance terms are identical to those found in univariate local polynomial regression; that is, the asymptotic variance of the estimator of m_1 does *not* depend on the simultaneous estimation of m_2. This somewhat surprising result is reminescent of what happens in a two-way ANOVA design with an equal number of observations in each cell, where the effect for one factor can be estimated without adjusting for the other factor. The reason for this apparent independence is that any

product of the two smoothers $\boldsymbol{S}_1^*\boldsymbol{S}_2^*$ is of order $O_p(\mathbf{1}\mathbf{1}^T/n)$, as shown in Lemma 3.1, and hence is of smaller order than the leading variance terms. This result holds regardless of the dependence between $\boldsymbol{X}$ and $\boldsymbol{Z}$, as long as restriction (4) is satisfied.

REMARK 4.4. We compare these asymptotic bias and variance results with those of Linton and Nielsen (1995) for the case $p_1 = p_2 = 1$. If we use f_Z as the weighting function q in their result, the theorem of Linton and Nielsen (1995) can be rewritten as

$$E(\hat{m}_1(X_i) - m_1(X_i)|\boldsymbol{X}, \boldsymbol{Z}) = \tfrac{1}{2}\mu_2(K)h_1^2 m_1''(X_i) + \tfrac{1}{2}\mu_2(K)h_2^2 E(m_2''(Z_i)) + o_p(h_1^2 + h_2^2),$$

$$\mathrm{Var}(\hat{m}_1(X_i)|\boldsymbol{X}, \boldsymbol{Z}) = \sigma^2 R(K)\frac{1}{nh_1}E_Z(f_{X|Z}(X_i|Z)^{-1}) + o_p\left(\frac{1}{nh_1}\right).$$

Both expressions ignore the boundary effects, so that Corollary 4.1 provides the relevant comparison. The rates of convergence for the bias and the variance are the same. The major differences in the bias are due to the effect in our results of the matrix $(\boldsymbol{I} - \boldsymbol{T}_{12}^*)^{-1}$, which makes the bias at X_i dependent on the curvature at all other observation points, and of the centering adjustment, which they do not account for. An interesting difference between the estimators occurs when $\boldsymbol{X}$ and $\boldsymbol{Z}$ are independent. It seems quite natural to expect from additive model estimators that, when the covariates are independent, the asymptotic bias for estimating one of the component functions does not depend on the behavior of the other function. As explained in Remark 4.1, the backfitting estimator indeed has this property, while the Linton–Nielsen estimator does not, as can readily be seen from the above bias expression. Unless the bias effects of the component functions happen to offset each other, this is likely to result in increased bias relative to the backfitting estimator. The comparison for the asymptotic variances is more straightforward, and, interestingly, the asymptotic variance of the backfitted estimators can be shown to be smaller than that of the "marginal integration" estimators, unless $\boldsymbol{X}$ and $\boldsymbol{Z}$ are independent. This is easily proven by noting that, in general,

$$E_Z(f_{X|Z}(X_i|Z)^{-1}) \geq f_X(X_i)^{-1}$$

by Jensen's inequality for the function $h(x) = 1/x$. Since h is strictly convex, we get strict inequality unless $f_{Z|X}(X_i|z) = f_X(X_i)$ for almost all z. Thus, we get strict inequality for all X_i values in a set of positive probability unless $\boldsymbol{X}$ and $\boldsymbol{Z}$ are independent.

PROOF OF THEOREM 4.1. We first prove the theorem for the case $p_1 = p_2 = 1$. To simplify the notation, we suppress the fact that the bias and the variance we are approximating are conditional on $\boldsymbol{X}$ and $\boldsymbol{Z}$.

For $\hat{\alpha}$, it is easy to see that

$$E(\hat{\alpha}) = \alpha + \bar{m}_1 + \bar{m}_2 = \alpha + O_p\left(\frac{1}{\sqrt{n}}\right).$$

In the case of $\hat{\boldsymbol{m}}_1$, we have

$$E(\hat{\boldsymbol{m}}_1) = \left(\boldsymbol{I} - (\boldsymbol{I} - \boldsymbol{S}_1^* \boldsymbol{S}_2^*)^{-1}(\boldsymbol{I} - \boldsymbol{S}_1^*)\right)(\boldsymbol{\alpha} + \boldsymbol{m}_1 + \boldsymbol{m}_2), \tag{6}$$

and we will apply the same Taylor expansion approximations as in Theorem 2.1 of Ruppert and Wand (1994). Let

$$\boldsymbol{Q}_{m_1}(x) = \begin{bmatrix} (X_1 - x)^2 \\ \vdots \\ (X_n - x)^2 \end{bmatrix} \frac{\partial^2 m_1(x)}{\partial x^2}$$

and

$$\boldsymbol{Q}_1 = \begin{bmatrix} \boldsymbol{s}_{1, X_1}^T \boldsymbol{Q}_{m_1}(X_1) \\ \vdots \\ \boldsymbol{s}_{1, X_n}^T \boldsymbol{Q}_{m_1}(X_n) \end{bmatrix},$$

and similarly for $\boldsymbol{Q}_{m_2}(z)$ and $\boldsymbol{Q}_2$. Letting $\boldsymbol{h}_1^2 \equiv h_1^2 \mathbf{1}$, we can write

$$\boldsymbol{S}_2 \boldsymbol{m}_2 = \boldsymbol{m}_2 + \tfrac{1}{2}\boldsymbol{Q}_2 + o_p(\boldsymbol{h}_2^2)$$

and hence

$$(\boldsymbol{I} - \boldsymbol{S}_1^* \boldsymbol{S}_2^*)^{-1}(\boldsymbol{I} - \boldsymbol{S}_1^*)\boldsymbol{m}_2 = \boldsymbol{m}_2 + \tfrac{1}{2}(\boldsymbol{I} - \boldsymbol{S}_1^* \boldsymbol{S}_2^*)^{-1}\boldsymbol{S}_1^* \boldsymbol{Q}_2 + o_p(\boldsymbol{h}_2^2).$$

Similarly,

$$(\boldsymbol{I} - \boldsymbol{S}_1^* \boldsymbol{S}_2^*)^{-1}(\boldsymbol{I} - \boldsymbol{S}_1^*)\boldsymbol{m}_1 = \bar{\boldsymbol{m}}_1 - \tfrac{1}{2}(\boldsymbol{I} - \boldsymbol{S}_1^* \boldsymbol{S}_2^*)^{-1}\boldsymbol{Q}_1^* + o_p(\boldsymbol{h}_1^2),$$

where $\boldsymbol{Q}_1^* = (\boldsymbol{I} - \mathbf{1}\mathbf{1}^T/n)\boldsymbol{Q}_1$. Plugging these results into expression (6), we write the bias vector for $\hat{\boldsymbol{m}}_1$ as

$$E(\hat{\boldsymbol{m}}_1 - \boldsymbol{m}_1) = \tfrac{1}{2}(\boldsymbol{I} - \boldsymbol{S}_2^* \boldsymbol{S}_1^*)^{-1}(\boldsymbol{Q}_1^* - \boldsymbol{S}_1^* \boldsymbol{Q}_2) + O_p\left(\frac{\mathbf{1}}{\sqrt{n}}\right) + o_p(\boldsymbol{h}_1^2 + \boldsymbol{h}_2^2). \tag{7}$$

The bias of $\hat{\boldsymbol{m}}_2$ and $\hat{\boldsymbol{m}}$ can be computed analogously. Note that the terms $\bar{m}_1$ and $\bar{m}_2$ will cancel in the expectation of $\hat{\boldsymbol{m}}$, so that the approximation term $O_p(\mathbf{1}/\sqrt{n})$ will not appear in that case.

The asymptotic bias in Theorem 2.2 of Ruppert and Wand (1994) can be rewritten as

$$\boldsymbol{s}_{1, x}^T \boldsymbol{Q}_{m_1}(x) = h_1^2 \mu_2(K, x) m_1''(x) + o_p(h_1^2).$$

By letting $\boldsymbol{M}_1 = \operatorname{diag}\{\mu_2(K, X_1), \ldots, \mu_2(K, X_n)\}$, we have therefore

$$\boldsymbol{Q}_1 = h_1^2 \boldsymbol{M}_1 D^2 \boldsymbol{m}_1 + o_p(\boldsymbol{h}_1^2)$$

and

$$\boldsymbol{Q}_1^* = \boldsymbol{Q}_1 - \mu_2(K)h_1^2 E(\boldsymbol{m}_1''(X_i)) + o_p(\boldsymbol{h}_1^2). \tag{8}$$

Because

$$\int |(\mu_2(K,x) - \mu_2(K))m_1''(x)| f_X(x)dx = O_p(h_1),$$

the matrix of boundary moments $\boldsymbol{M}_1$ is replaced by $\mu_2(K)\boldsymbol{I}$ in the latter term. Similar expressions hold for $\boldsymbol{Q}_2$ and $\boldsymbol{Q}_2^*$. Next, using the fact that

$$[\boldsymbol{N}(x)^{-1}]_{11}\mu(K,x) + [\boldsymbol{N}(x)^{-1}]_{12}\mu_1(K,x) = 1,$$

one can compute

$$\boldsymbol{s}_{1,x}^T \boldsymbol{M}_2 D^2 \boldsymbol{m}_2 = \mu_2(K)E(m_2''(Z_i)|x) + o_p(1)$$

and hence

$$\boldsymbol{S}_1\boldsymbol{Q}_2 = \mu_2(K)h_2^2 E(m_2''(Z_i)|\boldsymbol{X}) + o_p(\boldsymbol{h}_2^2)$$

and

$$\boldsymbol{S}_1^*\boldsymbol{Q}_2 = \boldsymbol{S}_1\boldsymbol{Q}_2 - \mu_2(K)h_2^2 E(\boldsymbol{m}_2''(Z_i)) + o_p(\boldsymbol{h}_2^2). \tag{9}$$

Plugging in results (8) and (9), as well as Lemma 3.2 into the bias vector (7), we obtain

$$\begin{aligned}
&E(\hat{\boldsymbol{m}}_1 - \boldsymbol{m}_1)\\
&\quad= \frac{1}{2}h_1^2\big((\boldsymbol{I} - \boldsymbol{T}_{12}^*)^{-1}\boldsymbol{M}_1 D^2\boldsymbol{m}_1 - \mu_2(K)E(m_1''(X_i))\big)\\
&\qquad- \frac{1}{2}h_2^2\mu_2(K)\big((\boldsymbol{I} - \boldsymbol{T}_{12}^*)^{-1}E(m_2''(Z_i)|\boldsymbol{X}) - E(m_2''(Z_i))\big)\\
&\qquad+ O_p\left(\frac{1}{\sqrt{n}}\right) + o_p(\boldsymbol{h}_1^2 + \boldsymbol{h}_2^2)\\
&\quad= \frac{1}{2}h_1^2\big(\boldsymbol{M}_1 D^2\boldsymbol{m}_1 + \mu_2(K)((\boldsymbol{I} - \boldsymbol{T}_{12}^*)^{-1} - \boldsymbol{I})D^2\boldsymbol{m}_1 - \mu_2(K)E(m_1''(X_i))\big)\\
&\qquad- \frac{1}{2}h_2^2\mu_2(K)\big((\boldsymbol{I} - \boldsymbol{T}_{12}^*)^{-1}E(m_2''(Z_i)|\boldsymbol{X}) - E(m_2''(Z_i))\big)\\
&\qquad+ O_p\left(\frac{1}{\sqrt{n}}\right) + o_p(\boldsymbol{h}_1^2 + \boldsymbol{h}_2^2).
\end{aligned}$$

The computations for the bias of $\hat{\boldsymbol{m}}$ are entirely analogous.

With $\boldsymbol{W}_1$ defined as in (2), the variance of $\hat{m}_1(X_i)$ is

$$\begin{aligned}
\operatorname{Var}(\hat{m}_1(X_i)) &= \sigma^2\boldsymbol{e}_i^T\boldsymbol{W}_1\boldsymbol{W}_1^T\boldsymbol{e}_i\\
&= \sigma^2\big\{1 - 2\boldsymbol{e}_i^T(\boldsymbol{I} - \boldsymbol{S}_1^*\boldsymbol{S}_2^*)^{-1}(\boldsymbol{I} - \boldsymbol{S}_1^*)\boldsymbol{e}_i\\
&\qquad+ \boldsymbol{e}_i^T(\boldsymbol{I} - \boldsymbol{S}_1^*\boldsymbol{S}_2^*)^{-1}(\boldsymbol{I} - \boldsymbol{S}_1^*)(\boldsymbol{I} - \boldsymbol{S}_1^*)^T(\boldsymbol{I} - \boldsymbol{S}_1^*\boldsymbol{S}_2^*)^{-T}\boldsymbol{e}_i\big\}.
\end{aligned} \tag{10}$$

Similar computations as in the proof of Lemma 3.1 lead to

$$[\boldsymbol{S}_1^*]_{ii} = \frac{1}{nh_1} f_X(X_i)^{-1}[\boldsymbol{N}(X_i)^{-1}]_{11}K(0) - \frac{1}{n} + o_p\left(\frac{1}{n}\right)$$
$$= \frac{1}{nh_1} f_X(X_i)^{-1}[\boldsymbol{N}(X_i)^{-1}]_{11}K(0) + o_p\left(\frac{1}{nh_1}\right)$$

and show that, when we multiply $\boldsymbol{S}_1^*$ by a matrix of order $O(\mathbf{1}\mathbf{1}^T/n)$, the resulting product is $o(\mathbf{1}\mathbf{1}^T/nh_1)$.

For the matrix $\boldsymbol{S}_1^*\boldsymbol{S}_1^{*T}$, note first that

$$\boldsymbol{S}_1^*\boldsymbol{S}_1^{*T} = \boldsymbol{S}_1\boldsymbol{S}_1^T - \mathbf{1}\mathbf{1}^T/n + o_p(\mathbf{1}\mathbf{1}^T/n) = \boldsymbol{S}_1\boldsymbol{S}_1^T + o_p\left(\frac{\mathbf{1}\mathbf{1}^T}{nh_1}\right).$$

Letting $\approx$ denote equality up to order $(1+o_p(1))$, the (i, j)th element of $\boldsymbol{S}_1\boldsymbol{S}_1^T$ is

$$\begin{aligned}
[\boldsymbol{S}_1\boldsymbol{S}_1^T]_{ij} \approx \frac{1}{n^2}\sum_{k=1}^n & f_X(X_i)^{-1}f_X(X_j)^{-1}[\boldsymbol{N}(X_i)^{-1}]_{11}[\boldsymbol{N}(X_j)^{-1}]_{11} \\
& \times K_{h_1}(X_k - X_i)K_{h_1}(X_k - X_j) \\
+ \frac{1}{n^2}\sum_{k=1}^n \frac{1}{h_1} & f_X(X_i)^{-1}f_X(X_j)^{-1}[\boldsymbol{N}(X_i)^{-1}]_{12}[\boldsymbol{N}(X_j)^{-1}]_{11} \\
& \times K_{h_1}(X_k - X_i)(X_k - X_i)K_{h_1}(X_k - X_j) \\
+ \frac{1}{n^2}\sum_{k=1}^n \frac{1}{h_1} & f_X(X_i)^{-1}f_X(X_j)^{-1}[\boldsymbol{N}(X_i)^{-1}]_{11}[\boldsymbol{N}(X_j)^{-1}]_{12} \\
& \times K_{h_1}(X_k - X_i)K_{h_1}(X_k - X_j)(X_k - X_j) \\
+ \frac{1}{n^2}\sum_{k=1}^n \frac{1}{h_1^2} & f_X(X_i)^{-1}f_X(X_j)^{-1}[\boldsymbol{N}(X_i)^{-1}]_{12}[\boldsymbol{N}(X_j)^{-1}]_{12} \\
& \times K_{h_1}(X_k - X_i)(X_k - X_i)K_{h_1}(X_k - X_j)(X_k - X_j).
\end{aligned}$$

When $i = j$,

$$[\boldsymbol{S}_1^*\boldsymbol{S}_1^{*T}]_{ii} = \frac{1}{nh_1} f_X(X_i)^{-1}R(K, X_i) + o_p\left(\frac{1}{nh_1}\right).$$

Let us write $H(u)$ for the function $uK(u)$. The elements of $\boldsymbol{S}_1^*\boldsymbol{S}_1^{*T}$ not on the diagonal are equal to

$$\begin{aligned}
[\boldsymbol{S}_1^*\boldsymbol{S}_1^{*T}]_{ij} \approx & \frac{1}{nh_1} f_X(X_i)^{-1}[\boldsymbol{N}(X_i)^{-1}]_{11}[\boldsymbol{N}(X_j)^{-1}]_{11}K * K\left(\frac{X_j - X_i}{h_1}\right) \\
& + \frac{1}{nh_1} f_X(X_i)^{-1}[\boldsymbol{N}(X_i)^{-1}]_{11}[\boldsymbol{N}(X_j)^{-1}]_{12}K * H\left(\frac{X_j - X_i}{h_1}\right)
\end{aligned}$$

$$+\frac{1}{nh_1}f_X(X_i)^{-1}[\boldsymbol{N}(X_i)^{-1}]_{12}[\boldsymbol{N}(X_j)^{-1}]_{11}H*K\left(\frac{X_j-X_i}{h_1}\right)$$
$$+\frac{1}{nh_1}f_X(X_i)^{-1}[\boldsymbol{N}(X_i)^{-1}]_{12}[\boldsymbol{N}(X_j)^{-1}]_{12}H*H\left(\frac{X_j-X_i}{h_1}\right)-\frac{1}{n},$$

so that $O_p(1/n)\sum_{j=1}^n[\boldsymbol{S}_1^{*T}\boldsymbol{S}_1^*]_{ij} = o_p(1/nh_1)$. Plugging these results and Lemma 3.2 into the terms of (10), we find

$$\operatorname{Var}(\hat{m}_1(X_i)) = \sigma^2\frac{R(K,X_i)}{nh_1}f_X(X_i)^{-1}+o_p\left(\frac{1}{nh_1}\right).$$

With $\boldsymbol{W}$ as defined in (3), we can write the variance of $\hat{m}(X_i,Z_i)$ as

$$\begin{aligned}\operatorname{Var}(\hat{m}(X_i,Z_i)) &= \sigma^2\boldsymbol{e}_i^T\boldsymbol{W}\boldsymbol{W}^T\boldsymbol{e}_i\\ &= \sigma^2\{5/n+4-4\boldsymbol{e}_i^T(\boldsymbol{I}-\boldsymbol{S}_1^*\boldsymbol{S}_2^*)^{-1}(\boldsymbol{I}-\boldsymbol{S}_1^*)\boldsymbol{e}_i\\ &\quad -4\boldsymbol{e}_i^T(\boldsymbol{I}-\boldsymbol{S}_2^*\boldsymbol{S}_1^*)^{-1}(\boldsymbol{I}-\boldsymbol{S}_2^*)\boldsymbol{e}_i\\ &\quad -2\boldsymbol{e}_i^T(\boldsymbol{I}-\boldsymbol{S}_1^*\boldsymbol{S}_2^*)^{-1}(\boldsymbol{I}-\boldsymbol{S}_1^*)\mathbf{1}\mathbf{1}^T/n\boldsymbol{e}_i\\ &\quad -2\boldsymbol{e}_i^T(\boldsymbol{I}-\boldsymbol{S}_2^*\boldsymbol{S}_1^*)^{-1}(\boldsymbol{I}-\boldsymbol{S}_2^*)\mathbf{1}\mathbf{1}^T/n\boldsymbol{e}_i\\ &\quad +2\boldsymbol{e}_i^T(\boldsymbol{I}-\boldsymbol{S}_1^*\boldsymbol{S}_2^*)^{-1}(\boldsymbol{I}-\boldsymbol{S}_1^*)(\boldsymbol{I}-\boldsymbol{S}_2^*)^T(\boldsymbol{I}-\boldsymbol{S}_2^*\boldsymbol{S}_1^*)^{-T}\boldsymbol{e}_i\\ &\quad +\boldsymbol{e}_i^T(\boldsymbol{I}-\boldsymbol{S}_1^*\boldsymbol{S}_2^*)^{-1}(\boldsymbol{I}-\boldsymbol{S}_1^*)(\boldsymbol{I}-\boldsymbol{S}_1^*)^T(\boldsymbol{I}-\boldsymbol{S}_1^*\boldsymbol{S}_2^*)^{-T}\boldsymbol{e}_i\\ &\quad +\boldsymbol{e}_i^T(\boldsymbol{I}-\boldsymbol{S}_2^*\boldsymbol{S}_1^*)^{-1}(\boldsymbol{I}-\boldsymbol{S}_2^*)(\boldsymbol{I}-\boldsymbol{S}_2^*)^T(\boldsymbol{I}-\boldsymbol{S}_2^*\boldsymbol{S}_1^*)^{-T}\boldsymbol{e}_i\}.\end{aligned}\tag{11}$$

After computations entirely analogous to those in Lemma 3.1, we find that the (i,j)th element of $\boldsymbol{S}_1^*\boldsymbol{S}_2^{*T}$ is

$$[\boldsymbol{S}_1^*\boldsymbol{S}_2^{*T}]_{ij} = [\boldsymbol{T}_{12}^*]_{ij}+o\left(\frac{1}{n}\right) = o\left(\frac{1}{nh_1}\right)\quad\text{a.s.},$$

so that

$$\operatorname{Var}(\hat{m}(X_i,Z_i)) = \sigma^2\left\{\frac{R(K,X_i)}{nh_1}f_X(X_i)^{-1}+\frac{R(K,Z_i)}{nh_2}f_Z(Z_i)^{-1}\right\}+o_p\left(\frac{1}{nh_1}+\frac{1}{nh_2}\right).$$

The generalization to arbitrary odd p_1, p_2 is straightforward. Let $\boldsymbol{Q}_1^{(p_1+1)}$ and $\boldsymbol{Q}_2^{(p_2+1)}$ represent the higher order generalizations corresponding to $\boldsymbol{Q}_1$ and $\boldsymbol{Q}_2$ for the local linear case, and let $\boldsymbol{M}_{1,p_1} = \operatorname{diag}\{\mu_{p_1+1}(K_{(p_1)},X_1),\ldots,\mu_{p_1+1}(K_{(p_1)},X_n)\}$. To compute the bias of $\hat{\boldsymbol{m}}_1$, we note that (6)–(7) still hold after replacing the $\frac{1}{2}\boldsymbol{Q}_1$ and $\frac{1}{2}\boldsymbol{Q}_2$ by $(1/(p_1+1)!)\boldsymbol{Q}_1^{(p_1+1)}$ and $(1/(p_2+1)!)\boldsymbol{Q}_1^{(p_2+1)}$, respectively, and adjusting the orders of the approximations from h_1^2 and h_2^2 to

$h_1^{p_1+1}$ and $h_2^{p_2+1}$. Equations (8) and (9) can be generalized analogously. Therefore,

$$\begin{aligned} E(\hat{\boldsymbol{m}}_1 - \boldsymbol{m}_1) = \frac{1}{(p_1+1)!} h_1^{p_1+1} \big(& \boldsymbol{M}_{1,\,p_1} D^{p_1+1} \boldsymbol{m}_1 \\ & + \mu_{p_1+1}(K_{(p_1)})((\boldsymbol{I} - \boldsymbol{T}_{12}^*)^{-1} - \boldsymbol{I}) D^{p_1+1} \boldsymbol{m}_1 \\ & - \mu_{p_1+1}(K_{(p_1)}) E(m_1^{(p_1+1)}(X_i)) \big) \\ - \frac{1}{(p_2+1)!} h_2^{p_2+1} \mu_{p_2+1}(K_{(p_2)}) \big(& (\boldsymbol{I} - \boldsymbol{T}_{12}^*)^{-1} E(m_2^{(p_2+1)}(Z_i) | \boldsymbol{X}) \\ & - E(m_2^{(p_2+1)}(Z_i)) \big) \\ + O_p\left(\frac{1}{\sqrt{n}}\right) + o_p(\boldsymbol{h}_1^{p_1+1} + \boldsymbol{h}_2^{p_2+1}). & \end{aligned}$$

The bias of $\hat{\boldsymbol{m}}$ is computed analogously.

For the variance, we note that (10) and (11) still hold. Using approximation (15), it is easy to compute that

$$[\boldsymbol{S}_1^*]_{ii} = \frac{1}{nh_1} f_X(X_i)^{-1} [\boldsymbol{N}_{p_1}(X_i)^{-1}]_{11} K(0) + o_p\left(\frac{1}{nh_1}\right),$$

$$[\boldsymbol{S}_1^* \boldsymbol{S}_1^{*T}]_{ii} = \frac{1}{nh_1} f_X(X_i)^{-1} R(K_{(p_1)}, X_i) + o_p\left(\frac{1}{nh_1}\right),$$

with all other terms in (10) of order $o_p(1/nh_1)$, so that

$$\operatorname{Var}(\hat{m}_1(X_i)) = \sigma^2 \frac{R(K_{(p_1)}, X_i)}{nh_1} f_X(X_i)^{-1} + o_p\left(\frac{1}{nh_1}\right).$$

A similar derivation leads to the desired expression for $\operatorname{Var}(\hat{m}(X_i, Z_i))$. □

5. Extension to local polynomials of even degree. We now consider the case when both of the degrees p_1 and p_2 are even. Let

$$C_{p_1}(z) = E(\mu_{p_1+1}(K_{(p_1)}, X_i) m_1^{(p_1+1)}(X_i) | z)$$

and similarly for $C_{p_2}(x)$. We define the matrices $\boldsymbol{F}_X = \operatorname{diag}\{f'_X(X_1)/f_X(X_1), \ldots, f'_X(X_n)/f_X(X_n)\}$, and similarly for $\boldsymbol{F}_Z$. $\boldsymbol{M}_{1,\,p_1}$ is defined in the proof of Theorem 4.1, and $\boldsymbol{M}_{2,\,p_2}$ is the corresponding matrix for the second covariate. We also replace Assumptions 3 and 4 by the following assumptions.

ASSUMPTION 3′. As $n \to \infty, h_1, h_2 \to \infty$ and $nh_1/\log n, nh_2/\log n, nh_1h_2/\log(n) \to \infty$.

ASSUMPTION 4′. The (p_1+2)th derivative of m_1 and the (p_2+2)th derivative of m_2 exist and are continuous and bounded.

We only state the asymptotic bias and variance of $\hat{m}_1(X_i)$. The proof of the following theorem can be found in Opsomer (1995).

THEOREM 5.1. *Assume that p_1 and p_2 are even and that Assumptions* 1, 2, 3′ *and* 4′ *hold. For the observation points (X_i, Z_i), $i = 1, \ldots, n$, the conditional bias and variance of $\hat{m}_1(X_i)$ can be approximated by*

$$
\begin{aligned}
&E(\hat{m}_1(X_i) - m_1(X_i)|\boldsymbol{X}, \boldsymbol{Z}) \\
&\quad = \frac{h_1^{p_1+1}}{(p_1+1)!}(\boldsymbol{t}_i^T \boldsymbol{M}_{1,\,p_1} D^{p_1+1}\boldsymbol{m}_1 - E(C_{p_1}(Z_i))) \\
&\qquad + h_1^{p_1+2}\Bigg(\mu_{p_1+2}(K_{(p_1)}, X_i)\Bigg(\frac{f'_X(X_i)}{f_X(X_i)}\frac{m_1^{(p_1+1)}(X_i)}{(p_1+1)!} + \frac{m_1^{(p_1+2)}(X_i)}{(p_1+2)!}\Bigg) \\
&\qquad\qquad + \mu_{p_1+2}(K_{(p_1)})(\boldsymbol{t}_i^T - \boldsymbol{e}_i^T)\Bigg(\boldsymbol{F}_X\frac{D^{p_1+1}\boldsymbol{m}_1}{(p_1+1)!} + \frac{D^{p_1+2}\boldsymbol{m}_1}{(p_1+2)!}\Bigg) \\
&\qquad\qquad - \mu_{p_1+2}(K_{(p_1)})E\Bigg(\frac{f'_X(X_i)}{f_X(X_i)}\frac{m_1^{(p_1+1)}(X_i)}{(p_1+1)!} + \frac{m_1^{(p_1+2)}(X_i)}{(p_1+2)!}\Bigg)\Bigg) \\
&\qquad - \frac{h_2^{p_2+1}}{(p_2+1)!}(\boldsymbol{t}_i^T C_{p_2}(\boldsymbol{X}) - E(C_{p_2}(X_i))) \\
&\qquad - h_2^{p_2+2}\mu_{p_2+2}(K_{(p_2)})\Bigg(\boldsymbol{t}_i^T E\Bigg(\frac{f'_Z(Z_i)}{f_Z(Z_i)}\frac{m_2^{(p_2+1)}(Z_i)}{(p_2+1)!} + \frac{m_2^{(p_2+2)}(Z_i)}{(p_2+2)!}\Bigg|\boldsymbol{X}\Bigg) \\
&\qquad\qquad - E\Bigg(\frac{f'_Z(Z_i)}{f_Z(Z_i)}\frac{m_2^{(p_2+1)}(Z_i)}{(p_2+1)!} + \frac{m_2^{(p_2+2)}(Z_i)}{(p_2+2)!}\Bigg)\Bigg) \\
&\qquad + O_p\Bigg(\frac{1}{\sqrt{n}}\Bigg) + o_p(h_1^{p_1+2} + h_2^{p_2+2})
\end{aligned}
$$

and

$$
\operatorname{Var}(\hat{m}_1(X_i)|\boldsymbol{X}, \boldsymbol{Z}) = \sigma^2\frac{R(K_{(p_1)})}{nh_1}f_X(X_i)^{-1} + o_p\Bigg(\frac{1}{nh_1}\Bigg).
$$

COROLLARY 5.1. *When $\boldsymbol{X}$ and $\boldsymbol{Z}$ are independent, the conditional bias of $\hat{m}_1(X_i)$ in the interior of* $\operatorname{supp}(f)$ *is*

$$
\begin{aligned}
&E(\hat{m}_1(X_i) - m_1(X_i)|\boldsymbol{X}, \boldsymbol{Z}) \\
&\quad = -\frac{h_1^{p_1+1}}{(p_1+1)!}E(C_{p_1}(Z_i)) \\
&\qquad + h_1^{p_1+2}\mu_{p_1+2}(K_{(p_1)})\Bigg(\Bigg(\frac{f'_X(X_i)}{f_X(X_i)}\frac{m_1^{(p_1+1)}(X_i)}{(p_1+1)!} + \frac{m_1^{(p_1+2)}(X_i)}{(p_1+2)!}\Bigg) \\
&\qquad\qquad - E\Bigg(\frac{f'_X(X_i)}{f_X(X_i)}\frac{m_1^{(p_1+1)}(X_i)}{(p_1+1)!} + \frac{m_1^{(p_1+2)}(X_i)}{(p_1+2)!}\Bigg)\Bigg) \\
&\qquad + O_p\Bigg(\frac{1}{\sqrt{n}}\Bigg) + o_p(h_1^{p_1+2} + h_2^{p_2+2}).
\end{aligned}
$$

REMARK 5.1. If X_i is in the interior, the bias of $\hat{m}_1(X_i)$ is of order $O_p(h_1^{p_1+2} + h_2^{p_2+2})$. To see this, note that $\mu_{p_1+1}(K_{(p_1)}) = 0$ for even p_1, so that $\boldsymbol{t}_i^T \boldsymbol{M}_{1,p_1} D^{p_1+1} \boldsymbol{m}_1 = O_p(h_1)$ and $C_{p_1}(Z_i) = O_p(h_1)$ for interior X_i. This order is the same as that found in Theorem 4.1 of Ruppert and Wand (1994) for nonadditive local polynomial regression, but the bias in Theorem 5.1 contains several additional terms. If X_i is on the boundary, $\boldsymbol{t}_i^T \boldsymbol{M}_{1,p_1} D^{p_1+1} \boldsymbol{m}_1$ is the leading term, so that the bias is of order $O_p(h_1^{p_1+1} + h_2^{p_2+2})$. If $\boldsymbol{X}$ and $\boldsymbol{Z}$ are independent, most of the additional terms disappear, with the exception of those due to centering, as shown in Corollary 5.1.

APPENDIX A

Proofs of lemmas.

PROOF OF LEMMA 3.1. For simplicity, we prove the result for $p_1 = p_2 = 1$, and then show how it can be extended to arbitrary p_1 and p_2. We first prove that the approximations in the lemma hold in probability. Let $\boldsymbol{A} \approx \boldsymbol{B}$ denote $\boldsymbol{A} = \boldsymbol{B}(1 + o_p(1))$ componentwise for any matrices $\boldsymbol{A}, \boldsymbol{B}$ of the same dimension. Theorem 2.2 of Ruppert and Wand (1994) shows that, for any x,

$$\left(\frac{1}{n}\boldsymbol{X}_x^T \boldsymbol{W}_x \boldsymbol{X}_x\right)^{-1} \approx f_X(x)^{-1} \boldsymbol{A}_1^{-1} \boldsymbol{N}(x)^{-1} \boldsymbol{A}_1^{-1}, \tag{12}$$

where $\boldsymbol{A}_1 = \operatorname{diag}\{1, h_1\}$. We can therefore write the (i, j)th element of $\boldsymbol{S}_1$ as

$$\begin{aligned}[\boldsymbol{S}_1]_{ij} \approx{}& \frac{1}{n} f_X(X_i)^{-1} h_1^{-1} [\boldsymbol{N}(X_i)^{-1}]_{11} K\left(\frac{X_j - X_i}{h_1}\right) \\ &+ \frac{1}{n} f_X(X_i)^{-1} h_1^{-1} [\boldsymbol{N}(X_i)^{-1}]_{12} K\left(\frac{X_j - X_i}{h_1}\right)\left(\frac{X_j - X_i}{h_1}\right).\end{aligned}$$

Using standard results from density estimation, it is straightforward to compute that

$$[\boldsymbol{S}_1^*]_{ij} = [\boldsymbol{S}_1]_{ij} - \frac{1}{n} + o_p\left(\frac{1}{n}\right). \tag{13}$$

In the case of $\boldsymbol{S}_1 \boldsymbol{S}_2$, similar computations show that

$$\begin{aligned}&[\boldsymbol{S}_1 \boldsymbol{S}_2]_{ij} \\ &\quad\approx \frac{1}{n^2} f_X(X_i)^{-1} [\boldsymbol{N}(X_i)^{-1}]_{11} \sum_{k=1}^{n} f_Z(Z_k)^{-1} [\boldsymbol{N}(Z_k)^{-1}]_{11} \\ &\qquad\qquad \times K_{h_1}(X_k - X_i) K_{h_2}(Z_j - Z_k) \\ &\qquad + \frac{1}{n^2} f_X(X_i)^{-1} [\boldsymbol{N}(X_i)^{-1}]_{11} \sum_{k=1}^{n} f_Z(Z_k)^{-1} [\boldsymbol{N}(Z_k)^{-1}]_{12} K_{h_1}(X_k - X_i) \\ &\qquad\qquad \times \frac{1}{h_2} K_{h_2}(Z_j - Z_k)(Z_j - Z_k)\end{aligned}$$

$$
\begin{aligned}
&+ \frac{1}{n^2} f_X(X_i)^{-1} [\boldsymbol{N}(X_i)^{-1}]_{12} \sum_{k=1}^{n} f_Z(Z_k)^{-1} [\boldsymbol{N}(Z_k)^{-1}]_{11} \frac{1}{h_1} K_{h_1}(X_k - X_i) \\
&\qquad \times (X_k - X_i) K_{h_2}(Z_j - Z_k) \\
&+ \frac{1}{n^2} f_X(X_i)^{-1} [\boldsymbol{N}(X_i)^{-1}]_{12} \sum_{k=1}^{n} f_Z(Z_k)^{-1} [\boldsymbol{N}(Z_k)^{-1}]_{12} \frac{1}{h_1} (X_k - X_i) \\
&\qquad \times \frac{1}{h_2} K_{h_2}(Z_j - Z_k)(Z_j - Z_k) \\
&= \frac{1}{n} \frac{f(X_i, Z_j)}{f_X(X_i) f_Z(Z_j)} + o_p\left(\frac{1}{n}\right).
\end{aligned}
$$

Using (13) and this result, we immediately find

$$
[\boldsymbol{S}_1^* \boldsymbol{S}_2^*]_{ij} = [\boldsymbol{T}_{12}^*]_{ij} + o_p\left(\frac{1}{n}\right). \tag{14}
$$

We now prove that (13) and (14) hold uniformly for all i, j and with $o_p(1)$ replaced by $o(1)$, using the theory from Chapter 2 of Pollard (1984) and two technical results which are proven in Appendix B. Expression (12) involves estimates of the moments of K up to order 2. By Assumptions 1 and 2 and Lemma B.1, the classes of graphs of the translation classes

$$
\left\{ (\cdot - x)^t K\left(\frac{\cdot - x}{h_1}\right) : x \in \operatorname{supp}(f_X) \right\} \quad \text{for } t = 0, 1, 2
$$

have polynomial discrimination (see Appendix B for definitions of these terms). Theorem II.37 of Pollard (1994) therefore guarantees that the approximation (12) holds uniformly for all $x \in \operatorname{supp}(f_X)$. For (13), the translation classes are

$$
\left\{ f_X(\cdot)^{-1} [\boldsymbol{N}(\cdot)^{-1}]_{1t} K\left(\frac{\cdot - x}{h_1}\right) \left(\frac{\cdot - x}{h_1}\right)^{t-1} : x \in \operatorname{supp}(f_X) \right\}
$$

for $t = 1, 2$, and the same reasoning as above, combined with the fact that (12) holds uniformly, can be used to show that the approximation (13) holds uniformly over all i, j, proving the first part of the lemma.

Similarly, we consider (14). In this case, the translation classes are

$$
\left\{ f_Z(\cdot - z) [\boldsymbol{N}(\cdot)^{-1}]_{1u} K\left(\frac{\cdot - z}{h_2}\right) \left(\frac{\cdot - z}{h_2}\right)^{u-1} K\left(\frac{\cdot - x}{h_1}\right) \left(\frac{\cdot - x}{h_1}\right)^{t-1} \right\}
$$

for $t = 1, 2$, $u = 1, 2$. Since these translation classes are generated by the product of functions whose graphs have polynomial discrimination, Lemma 2.25 of Pollard (1984) and Lemma B.2 ensure that the assumptions of Theorem 2.37 of Pollard (1984) are satisfied. The approximation (14) therefore also holds uniformly over all i, j, completing the proof for the local linear case.

The generalization to arbitrary degree local polynomials is straightforward. The (i, j)th element of $\boldsymbol{S}_1$ is now

$$
\begin{aligned}
[\boldsymbol{S}_1]_{ij} &\approx \frac{1}{n} f_X(X_i)^{-1} \sum_{k=1}^{p_1+1} [\boldsymbol{N}_{p_1}(X_i)^{-1}]_{1k} K_{h_1}(X_j - X_i)\left(\frac{X_j - X_i}{h_1}\right)^{k-1} \\
&\approx \frac{1}{n h_1} f_X(x)^{-1} K_{(p_1)}\left(\frac{X_j - x}{h_1}, x\right),
\end{aligned} \tag{15}
$$

so that again

$$
[\boldsymbol{S}_1^*]_{ij} = [\boldsymbol{S}_1]_{ij} - \frac{1}{n} + o_p\left(\frac{1}{n}\right),
$$

since $\mu(K_{(p)}, x) = 1$ for any x. As before, we can show that the classes of graphs

$$
\{K_{(p_1)}\left(\frac{\cdot - x}{h_1}\right): x \in \operatorname{supp}(f_X)\}
$$

have polynomial discrimination for any p_1 using Pollard (1984) and the results in Appendix B, so that this approximation also holds uniformly over all i, j. The reasoning for $\boldsymbol{S}_1^*\boldsymbol{S}_2^*$ is completely analogous. □

PROOF OF LEMMA 3.2. For any matrix $\boldsymbol{A}$, write $\rho(\boldsymbol{A})$ for the spectral radius of $\boldsymbol{A}$. Because of Assumption 2, we know that $\rho(\boldsymbol{T}_{12}^*) \le \|\boldsymbol{T}_{12}^*\|_r < 1$ so that $(\boldsymbol{I} - \boldsymbol{T}_{12}^*)$ is invertible. By Lemma 3.1, it immediately follows that

$$
P(\exists\ N\colon \rho(\boldsymbol{S}_1^*\boldsymbol{S}_2^*) < 1 \text{ if } n \ge N) = 1,
$$

establishing the first part of the lemma.

Assume now that $(\boldsymbol{I} - \boldsymbol{S}_1^*\boldsymbol{S}_2^*)^{-1}$ exists. If we can show that $o(\mathbf{1}\mathbf{1}^T/n)(\boldsymbol{I} - \boldsymbol{T}_{12}^*)^{-1} = o(\mathbf{1}\mathbf{1}^T/n)$, then

$$
(\boldsymbol{I} - \boldsymbol{S}_1^*\boldsymbol{S}_2^*)^{-1} = (\boldsymbol{I} - \boldsymbol{T}_{12}^*)^{-1} + o(\mathbf{1}\mathbf{1}^T/n) \quad \text{a.s.}
$$

by the formula for the inverse of a sum of matrices [Horn and Johnson (1985), page 19]. Since $\rho(\boldsymbol{T}_{12}^*) < 1$, we can write

$$
(\boldsymbol{I} - \boldsymbol{T}_{12}^*)^{-1} = \boldsymbol{I} + \sum_{p=1}^{\infty} \boldsymbol{T}_{12}^{*p}.
$$

Using Assumption 2 again, it is easy to show by induction that, for all p, $\max_{i,j} |[\boldsymbol{T}_{12}^{*p}]_{ij}| \le (1 - \varepsilon)^p/n$ for some $\varepsilon > 0$. Hence,

$$
\max_{i,j} \left|(\boldsymbol{I} - \boldsymbol{T}_{12}^*)^{-1} - \boldsymbol{I}\right| \le \frac{K}{n}
$$

for all n and $o(\mathbf{1}\mathbf{1}^T/n)(\boldsymbol{I} - \boldsymbol{T}_{12}^*)^{-1} = o(\mathbf{1}\mathbf{1}^T/n)$ as desired. □

APPENDIX B

Results from empirical process theory. We briefly state some of the definitions used in Pollard (1984). For a probability measure Q on S, a class of functions $\mathscr{F}$ and $\varepsilon > 0$, define the *covering number* $N(\varepsilon, Q, \mathscr{F})$ as the smallest value of m for which there exist functions $g_1, \ldots, g_m$ (not necessarily in $\mathscr{F}$) such that $\min_j P|f - g_j| \leq \varepsilon$ for each f in $\mathscr{F}$. The *graph* of a real-valued function f on a set S, written $\mathrm{gr}(f)$, is defined as the subset

$$\mathrm{gr}(f) = \{(s, t): 0 \leq t \leq f(s) \text{ or } f(s) \leq t < 0\}$$

of $S \times \mathbb{R}$. A class $\mathscr{D}$ of subsets of some space H has *polynomial discrimination* if there exists a polynomial $\rho(\cdot)$ such that, from every set of N points in H, the class picks out at most $\rho(N)$ distinct subsets.

We define a *translation class* of functions on g as the class $\{g(\cdot - x)\}$ for any function g on a set $S \subseteq \mathbb{R}$. It will be convenient to set $g(t) = 0$ for $t \notin S$, so that the domain of $g(t - x)$ does not depend on x. The set of graphs generated by the translation class on g will be written as $\mathscr{G}_g$. We also define a *monotonicity change* for a function g as (1) any point t_0 for which $g(t)$ changes from monotone increasing to monotone decreasing (or vice versa) in an interval $(t_0 - \varepsilon, t_0 + \varepsilon)$ for some $\varepsilon > 0$, and (2) any set of points (t_1, t_2), $t_1 \neq t_2$, for which $g(t) = c$ for all $t \in (t_1, t_2)$ for some c and $g(t)$ changes from monotone increasing to monotone decreasing (or vice versa) in an interval $(t_1 - \varepsilon, t_2 + \varepsilon)$ for some $\varepsilon > 0$. We prove two lemmas.

LEMMA B.1. (i) *Suppose that the function $g(u)$ on $S \subseteq \mathbb{R}$ has a finite number of monotonicity changes. Then the set of graphs $\mathscr{G}_g$ has polynomial discrimination in $S \times \mathbb{R}$.*

(ii) *If the function $h(u)$ has the same properties as g, then the following sets also have polynomial discrimination in $S \times \mathbb{R}$: $\mathscr{G}_{gh}$ generated by the functions $\{g(\cdot - x)h(\cdot)\}$ and $\mathscr{G}'_{gh}$ generated by $\{g(\cdot - x)h(\cdot - x)\}$.*

PROOF. Let us first look at the simplest possible case, where g does not change sign and has no monotonicity changes. We assume, without loss of generality, that $g \geq 0$ and is increasing. From a set containing only two points, $\mathscr{G}_g$ can never pick out both singletons, since, for any $x_1 < x_2$, we have $g(s - x_1) \geq g(s - x_2)$ for all s, so that $\mathrm{gr}(g(\cdot - x_2)) \subseteq \mathrm{gr}(g(\cdot - x_1))$. By Lemma 2.17 of Pollard (1984), $\mathscr{G}_g$ therefore has linear discrimination.

Suppose now that g is monotone. We can write

$$g = g^+ + g^- \equiv gI_{\{g>0\}} + gI_{\{g\leq o\}}$$

so that

$$\mathrm{gr}(g(\cdot - x)) = \mathrm{gr}(g^+(\cdot - x)) \cup \mathrm{gr}(g^-(\cdot - x)).$$

By Lemma 2.15 of Pollard (1984), we conclude that $\mathscr{G}_g$ is also a polynomial class in this case.

Next, suppose that g has no sign changes but has exactly one monotonicity change. Without loss of generality, let $g \geq 0$. If the monotonicity changes from decreasing to increasing, then the graphs of g can be written as the union of graphs of two monotone functions. If the change is from increasing to decreasing, we can write it as the intersection of two graphs of monotone functions. Hence, Lemma 2.15 of Pollard (1984) can also be applied in this case.

Finally, consider any function g with a finite number of monotonicity changes. Begin by writing it as the sum of a positive and a nonnegative function, say g^+ and g^-. The graphs of each of these can be written as the finite union of graphs of functions with one monotonicity change as well as at most two graphs of monotone functions. Hence, for g^+,

$$\operatorname{gr}(g^+(\cdot - x)) = \bigcup_{i=1}^{M_+} (\operatorname{gr}(g_{i1}^+(\cdot - x)) \cap \operatorname{gr}(g_{i2}^+(\cdot - x))),$$

where some of the sets $\operatorname{gr}(g_{ij}^+(\cdot - x))$ can be empty sets for $i = 1$ and $i = M_+$. A similar expression holds for g^-. Using Lemma 2.15 of Pollard (1984) again proves result (i).

Result (ii) follows immediately from the fact that if the derivative of a function g has a finite number of sign changes, then so does the function itself. □

LEMMA B.2. *Let $\mathscr{G}$, $\mathscr{R}$ represent two classes of bounded, real-valued functions on S and T, respectively. Suppose there exists constant $A_g, A_r, w_g, w_r > 0$ and the covering numbers of $\mathscr{G}$, $\mathscr{R}$ satisfy*

$$N(\varepsilon, P, \mathscr{G}) \leq A_g \varepsilon^{-w_g} \quad \text{for } 0 < \varepsilon < 1,$$
$$N(\varepsilon, P, \mathscr{R}) \leq A_r \varepsilon^{-w_r} \quad \text{for } 0 < \varepsilon < 1$$

for any probability measure P. Let $f(u, t) = g(u)r(t)$, the product function on $S \times T$. The covering numbers of the class $\mathscr{F} = \{f\colon f = g \times r,\ g \in \mathscr{G},\ r \in \mathscr{R}\}$ satisfy

$$N(\varepsilon, P, \mathscr{F}) \leq A_f \varepsilon^{-w_f} \quad \text{for } 0 < \varepsilon < 1$$

for some A_f, w_f.

PROOF. Let G represent a function for which $|g| \leq G$ for all $g \in \mathscr{G}$, and R a similar function for $\mathscr{R}$. Let $M_g = \max |G|$, $M_r = \max |R|$. We need to find how many functions f_i are required for each $\varepsilon > 0$, so that

$$\min_{f_i} P|f - f_i| < \varepsilon$$

for all f, or, equivalently, how many pairs of functions (g_k, r_l) are required, so that

$$\min_{(g_k, r_l)} P|g \times r - g_k \times r_l| < \varepsilon$$

for all (g, r). Since

$$P|g \times r - g_k \times r_l| \leq M_r P|g - g_k| + M_g P|r - r_l|,$$

if we require that

$$\min_{g_k} P|g - g_k| < \frac{\varepsilon}{2M_r},$$
$$\min_{r_l} P|r - r_l| < \frac{\varepsilon}{2M_g},$$

the number of functions g_k, r_l required to achieve this are $A_g(\varepsilon/2M_r)^{-w_g}$ and $A_r(\varepsilon/2M_g)^{-w_r}$, respectively. Hence,

$$N(\varepsilon, P, \mathscr{F}) \leq A_f \varepsilon^{-w_f},$$

with $A_f = A_g A_r (2M_r)^{w_g} (2M_g)^{w_r}$ and $w_f = w_g + w_r$. □

REFERENCES

BUJA, A., HASTIE, T. J. and TIBSHIRANI, R. J. (1989). Linear smoothers and additive models. *Ann. Statist.* **17** 453–555.

BURMAN, P. (1990). Estimation of generalized additive models. *J. Multivariate Anal.* **32** 230–255.

CHAMBERS, J. M. and HASTIE, T. J., eds. (1992). *Statistical models in S*. Wadsworth and Brooks/Cole, Pacific Grove, CA.

CHEN, Z. (1993). Fitting multivariate regression functions by interaction spline models. *J. Roy. Statist. Soc. Ser. B* **55** 473–491.

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

FAN, J., GASSER, T., GIJBELS, I., BROCKMANN, M. and ENGEL, J. (1993). Local polynomial fitting: a standard for nonparametric regression. Mimeo Series 2302, Institute of Statistics, Univ. North Carolina, Chapel Hill.

FAN, J. and GIJBELS, I. (1994). Adaptive order polynomial fitting: bandwidth robustification and bias reduction. Unpublished manuscript.

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

GU, C., BATES, D. M., CHEN, Z. and WAHBA, G. (1989). The computation of gcv functions through householder tridiagonalization with application to the fitting of interaction spline models. *SIAM J. Matrix Anal. Appl.* **10** 457–480.

HÄRDLE, W., HALL, P. and MARRON, J. S. (1988). How far are automatically chosen regression smoothing parameters from their optimum? *J. Amer. Statist. Assoc.* **83** 86–95.

HÄRDLE, W. and HALL, P. (1993). On the backfitting algorithm for additive regression models. *Statist. Neerlandica* **47** 43–57.

HASTIE, T. J. and TIBSHIRANI, R. J. (1990). *Generalized Additive Models*. Chapman and Hall, Washington, DC.

HORN, R. A. and JOHNSON, C. A. (1985). *Matrix Analysis*. Cambridge Univ. Press.

LINTON, O. and NIELSEN, J. P. (1995). A kernel method of estimating structured nonparametric regression based on marginal integration. *Biometrika* **82** 93–100.

OPSOMER, J.-D. (1995). Optimal bandwidth selection for fitting an additive model by local polynomial regression. Ph.D. dissertation, Cornell Univ.

OPSOMER, J.-D. and RUPPERT, D. (1995). A fully automated bandwidth selection method for fitting additive models. Preprint 95-32, Dept. Statistics, Iowa State Univ.

POLLARD, D. (1984). *Convergence of Stochastic Processes*. Springer, New York.

RUPPERT, D. and WAND, M. P. (1994). Multivariate locally weighted least squares regression. *Ann. Statist.* **22** 1346–1370.

STONE, C. J. (1985). Additive regression and other nonparametric models. *Ann. Statist.* **13** 689–705.

WAHBA, G. (1986). Partial and interaction splines for the semiparametric estimation of functions of several variables. In *Computer Science and Statistics: Proceedings of the 18th Symposium on the Interface* (T. J. Boardman, ed.) 75–80. Amer. Statist. Assoc., Alexandria, VA.

DEPARTMENT OF STATISTICS
SNEDECOR HALL
IOWA STATE UNIVERSITY
AMES, IOWA 50011-1210

COLLEGE OF ENGINEERING
RHODES HALL
CORNELL UNIVERSITY
ITHACA, NEW YORK 14853-3801