A Fully Automated Bandwidth Selection Method for Fitting Additive Models^{1 2 3}

Jean D. Opsomer⁴
Iowa State University

David Ruppert⁵
Cornell University

February 14, 1996

¹AMS 1991 subject classifications: primary 62G07; secondary 62H99.

²Key words and phrases: additive model, local polynomial regression, automatic bandwidth selection, variance estimation, housing prices.

³Abbreviated title: Bandwidth Selection for Additive Models.

⁴Jean Opsomer was supported by NIEHS Training Grant ESO7261 while at Cornell University.

⁵David Ruppert was supported by NSA Grant MDA904-95-H-1025 and NSF Grant DMS-9306196.

Abstract

This article describes a fully automated bandwidth selection method for additive models that is applicable to the widely used backfitting algorithm of Buja, Hastie and Tibshirani (1989) and does not rely on cross-validation. The proposed plug-in estimator is an extension of the local linear regression estimator of Ruppert, Sheather and Wand (1996) and is shown to achieve the same $O_p(n^{-2/7})$ relative convergence rate for bivariate additive models. If more than two covariates are present, theoretical justification of the method requires independence of the covariates, but simulation experiments show that in practice the method is very robust to violations of this assumption. The behavior of the method is demonstrated on a real dataset.

1 Introduction

Additive models (Hastie and Tibshirani [16]) are a popular multivariate nonparametric fitting technique. The additive model assumes that the conditional expectation function of the dependent variable Y can be written as the sum of smooth terms in the covariates $X_1, \dots, X_D$:

$$E(Y|X = (x_1, \dots, x_D)) = m_1(x_1) + \dots + m_D(x_D). \quad (1)$$

The additive model's appeal is that the fitted models are free of restrictive parametric assumptions, as with any other nonparametric method, but unlike most of them, the effects of individual covariates on the dependent variable can still easily be interpreted, regardless of the number of covariates D . The availability of easy to use model estimation software in S-Plus (Chambers and Hastie [6]) has further contributed to its widespread use.

In a previous paper (Opsomer and Ruppert [19]), we explored the asymptotic bias and variance properties of the bivariate additive model, fit by local polynomial regression. While these results provide valuable insights in the theoretical behavior of the additive model, they can also be put to more practical use. The current paper will expand on these results to develop a fully automatic bandwidth selection method for the additive model with any number of covariates.

Most bandwidth selection methods proposed for the additive model rely on cross-validation or one of its approximations (Hastie and Tibshirani [16]). Despite its intuitive appeal and simplicity, this approach suffers from two drawbacks. The first concerns the properties of the bandwidth estimators. In the closely related regression smoothing context, cross-validation estimators have been shown to be limited to a $O_p(n^{-1/10})$ relative rate of convergence and to display large sample-to-sample variability (Härdle *et al.* [14]). Perhaps even more important from a practical standpoint, the second drawback is that these bandwidth selectors are very computation-intensive: for a model with D covariates, the search for the “optimal” bandwidth has to take place by numerical approximation over $\mathbb{R}_+^D$. While methods are available to make this search more efficient (e.g. Gu and Wahba [11]), it still necessitates the calculation of numerous additive model fits. In the current article, *plug-in* bandwidth estimators for the additive model will be developed and shown to address both these drawbacks. Plug-in bandwidth estimators are well-known in kernel smoothing, kernel regression and local polynomial regression, and several authors have developed estimators with good theoretical and practical properties. For an overview of the literature on this subject, see Wand and Jones [23].

The two goals of this article are: (1) develop plug-in estimators for the optimal bandwidth vector $(h_1, \dots, h_D)^T$ of the additive model (1) fitted by local polynomial regression, and (2) provide a fully automated bandwidth selection and fitting method that is compatible with the backfitting algorithm of Buja *et al.* [5]. In Section 2, the theoretical framework for the plug-in bandwidth estimators is developed. Section 3 describes the proposed bandwidth selection and model fitting algorithm. The practical behavior of the algorithm is demonstrated on simulated and real data in Sections 4 and 5, respectively.

2 Theoretical Properties

2.1 A Simple Case

We begin by describing the bandwidth selection problem in the context of a bivariate additive model fitted by local linear regression. Let $(X_1, Z_1, Y_1), \dots, (X_n, Z_n, Y_n)$ be a set of independent and identically distributed $\mathbb{R}^3$ -valued random variables. We consider the following model

$$Y_i = \alpha + m_1(X_i) + m_2(Z_i) + \varepsilon_i \quad (2)$$

where the ε_i are independent and identically distributed with mean 0 and variance σ^2 (to ensure identifiability of the estimators, we include the intercept α and assume $E(m_1(X_i)) = E(m_2(Z_i)) = 0$). Fitting an additive model requires choosing bandwidth parameters for m_1 and m_2 , denoted by h_1 and h_2 , respectively. The “optimal” values for h_1 and h_2 are taken to be those that minimize the conditional Mean Average Squared Error (MASE), as discussed by Härdle *et al.* [14]. The MASE of m is defined as

$$\text{MASE}(h_1, h_2 | \mathbf{X}, \mathbf{Z}) = \frac{1}{n} \sum_{i=1}^n E(\hat{m}(X_i, Z_i) - m(X_i, Z_i) | \mathbf{X}, \mathbf{Z})^2.$$

We review some of the notation in Opsomer and Ruppert [19]. Let $\mathbf{Y} = (Y_1, \dots, Y_n)^T$ and similarly for $\mathbf{X}$ and $\mathbf{Z}$, and write the vectors of additive functions at the observation points as $\mathbf{m}_1 = (m_1(X_1), \dots, m_1(X_n))^T$, $\mathbf{m}_2 = (m_2(Z_1), \dots, m_2(Z_n))^T$. Let $\mathbf{s}_{1,x}^T, \mathbf{s}_{2,z}^T$ represent the local polynomial smoothers at x and z . In the case of x , this smoother can be written as

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

where $\mathbf{e}_1^T = (1, 0)$, $\mathbf{W}_x = \text{diag}\{\frac{1}{h_1}K(\frac{X_1-x}{h_1}), \dots, \frac{1}{h_1}K(\frac{X_n-x}{h_1})\}$ for some kernel function K

and

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

with p the degree of the local polynomials (in this Subsection, $p = 1$). Let $\mathbf{S}_1, \mathbf{S}_2$ represent the matrices whose rows are the smoothers at the observations $\mathbf{X}$ and $\mathbf{Z}$, respectively:

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

f represents the joint density of X_i and Z_i , with f_X and f_Z as corresponding marginal densities, and $(a_x, b_x), (a_z, b_z)$ are the ranges of X_i and Z_i . We define $\mathbf{T}_{12}^*$ as the matrix whose ij th element is

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

and let $\mathbf{t}_i^T$ and $\mathbf{v}_i$ represent the i th row and column, respectively, of $(\mathbf{I} - \mathbf{T}_{12}^*)^{-1}$, provided the inverse exists. Finally, let D^r denote the r th derivative operator and

$$\mathbb{E}(m_1^{(r)}(X_i)|\mathbf{Z}) = \begin{bmatrix} \mathbb{E}(m_1^{(r)}(X_i)|Z_1) \\ \vdots \\ \mathbb{E}(m_1^{(r)}(X_i)|Z_n) \end{bmatrix}, \quad \mathbb{E}(m_2^{(r)}(Z_i)|\mathbf{X}) = \begin{bmatrix} \mathbb{E}(m_2^{(r)}(Z_i)|X_1) \\ \vdots \\ \mathbb{E}(m_2^{(r)}(Z_i)|X_n) \end{bmatrix}.$$

The vectors of fitted values at the observation points were defined in Opsomer and Ruppert [19] as $\hat{\mathbf{m}} = \hat{\boldsymbol{\alpha}} + \hat{\mathbf{m}}_1 + \hat{\mathbf{m}}_2$, in which $\hat{\boldsymbol{\alpha}} = \bar{Y}$, and $\hat{\mathbf{m}}_1$ and $\hat{\mathbf{m}}_2$ are the solutions to the set of estimating equations

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

where $\mathbf{S}_1^* = (\mathbf{I} - \mathbf{1}\mathbf{1}^T/n)\mathbf{S}_1$ and similarly for $\mathbf{S}_2^*$. In general, backfitting (Buja *et al.* [5]) is used to solve these equations. In the bivariate case, backfitting converges to the explicit solution

$$\begin{aligned} \hat{\mathbf{m}}_1 &= \left(\mathbf{I} - (\mathbf{I} - \mathbf{S}_1^* \mathbf{S}_2^*)^{-1} (\mathbf{I} - \mathbf{S}_1^*) \right) \mathbf{Y} \equiv \mathbf{W}_1 \mathbf{Y} \\ \hat{\mathbf{m}}_2 &= \left(\mathbf{I} - (\mathbf{I} - \mathbf{S}_2^* \mathbf{S}_1^*)^{-1} (\mathbf{I} - \mathbf{S}_2^*) \right) \mathbf{Y} \equiv \mathbf{W}_2 \mathbf{Y} \end{aligned} \quad (3)$$

if $\|\mathbf{S}_1^* \mathbf{S}_2^*\| < 1$ for some matrix norm $\|\cdot\|$. Sufficient conditions guaranteeing the existence of these estimators for large sample sizes are provided by:

(AS.I) The kernel K is bounded and continuous, has compact support and its first derivative has a finite number of sign changes over its support. Also, $\mu_j(K) \equiv \int u^j K(u) du = 0$ for all odd j and $\mu_2(K) \neq 0$.

(AS.II) The densities $f(x, z)$, $f_X(x)$ and $f_Z(z)$ 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 $(x, z) \in \text{supp}(f)$ and

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

(AS.III) As $n \rightarrow \infty$, $h_1, h_2 \rightarrow 0$ and $nh_1/\log(n), nh_2/\log(n) \rightarrow \infty$.

(AS.IV) The second derivatives of m_1 and m_2 exist and are bounded and continuous.

These assumptions and their applicability are discussed in Opsomer and Ruppert [19].

From Corollary 4.2 of Opsomer and Ruppert [19], it is easily derived that when the additive model is fitted by local linear regression, the asymptotic approximation to the conditional MASE, denoted by AMASE, is

$$\begin{aligned} \text{AMASE}(h_1, h_2 | \mathbf{X}, \mathbf{Z}) &= \frac{\mu_2(K)^2}{4} (h_1^4 \theta_{11} + h_1^2 h_2^2 \theta_{12} + h_2^4 \theta_{22}) \\ &\quad + \sigma^2 R(K) \left(\frac{b_x - a_x}{nh_1} + \frac{b_z - a_z}{nh_2} \right), \end{aligned} \quad (4)$$

where $R(K) = \int K(u)^2 du$,

$$\begin{aligned} \theta_{11} &= \frac{1}{n} \sum_{i=1}^n \left(\mathbf{t}_i^T D^2 \mathbf{m}_1 - \mathbf{v}_i^T \mathbb{E}(m_1''(X_i) | \mathbf{Z}) \right)^2 \\ \theta_{22} &= \frac{1}{n} \sum_{i=1}^n \left(\mathbf{v}_i^T D^2 \mathbf{m}_2 - \mathbf{t}_i^T \mathbb{E}(m_2''(Z_i) | \mathbf{X}) \right)^2 \end{aligned}$$

and

$$\theta_{12} = \frac{1}{n} \sum_{i=1}^n \left(\mathbf{t}_i^T D^2 \mathbf{m}_1 - \mathbf{v}_i^T \mathbb{E}(m_1''(X_i) | \mathbf{Z}) \right) \left(\mathbf{v}_i^T D^2 \mathbf{m}_2 - \mathbf{t}_i^T \mathbb{E}(m_2''(Z_i) | \mathbf{X}) \right).$$

We denote the values of the bandwidth parameters that minimize the AMASE as $h_{1, \text{AMASE}}$ and $h_{2, \text{AMASE}}$. While simpler than the MASE, the AMASE expression still cannot be minimized as written, since it contains the unknown $\theta_{11}, \theta_{12}, \theta_{22}$ and σ^2 . The *plug-in* estimators $\hat{h}_1$ and $\hat{h}_2$ are the solutions to the following three-step process:

Step I Estimate $\theta_{11}, \theta_{12}, \theta_{22}$ and σ^2 by $\hat{\theta}_{11}, \hat{\theta}_{12}, \hat{\theta}_{22}$ and $\hat{\sigma}^2$, respectively,

Step II plug $\hat{\theta}_{11}, \hat{\theta}_{12}, \hat{\theta}_{22}$ and $\hat{\sigma}^2$ in the AMASE, and

Step III find the values of $h_1, h_2 > 0$ that minimize the AMASE.

Typically, no explicit expressions for $\hat{h}_1$ and $\hat{h}_2$ are available. If nonparametric regression is to be used in Step I, separate bandwidth parameters will have to be estimated for $\hat{\theta}_{11}, \hat{\theta}_{12}, \hat{\theta}_{22}$ and $\hat{\sigma}^2$. We will therefore also be interested in the asymptotic MSE of these estimators.

When $\mathbf{X}$ and $\mathbf{Z}$ are independent, the bandwidth selection problem becomes simpler. The AMASE is still as in (4), but with $\theta_{12} = 0$ and

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

and similarly for θ_{22} . Under the assumption of independence, the minimizers of the AMASE are easily computed to be

$$h_{1,AMASE} = \left(\frac{R(K) \sigma^2 (b_x - a_x)}{n \mu_2(K)^2 \theta_{11}} \right)^{1/5}, \quad h_{2,AMASE} = \left(\frac{R(K) \sigma^2 (b_z - a_z)}{n \mu_2(K)^2 \theta_{22}} \right)^{1/5}. \quad (5)$$

These expressions are identical to the h_{AMISE} for univariate local linear regression in Ruppert *et al.* [21], except for the fact that the derivatives in the functionals θ_{11}, θ_{22} are now “centered.”

In Subsections 2.2 and 2.3, we will derive estimators for the unknown quantities in the AMASE. Subsection 2.4 proposes a specific choice for the plug-in estimators $\hat{h}_1, \hat{h}_2$ and studies their asymptotic properties. In Subsection 2.5, these results are extended to the additive model with D independent covariates and local polynomials of arbitrary odd degree. Outlines of all important proofs are given in the Appendix. Detailed proofs can be found in Opsomer [18].

2.2 Estimation of Regression Functionals

In this subsection, we look at regression functionals of the form

$$\theta_{11}(r, s) = \frac{1}{n} \sum_{i=1}^n \left(\mathbf{t}_i^T D^r \mathbf{m}_1 - \mathbf{v}_i^T E(m_1^{(r)}(X_i) | \mathbf{Z}) \right) \left(\mathbf{t}_i^T D^s \mathbf{m}_1 - \mathbf{v}_i^T E(m_1^{(s)}(X_i) | \mathbf{Z}) \right)$$

and similarly for $\theta_{22}(r, s)$, and

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

with $r, s \geq 1, r + s$ even. Clearly, $\theta_{kl}(2, 2) \equiv \theta_{kl}$ from the previous subsection. We are considering general r and s , since regression functionals of the form $\theta_{kl}(r, s)$ will reappear in the asymptotic approximations of σ^2 and θ_{kl} . We assume (but do not explicitly indicate) that all the smoother vectors and matrices are of degree $p > r, s$ and that $p - r, p - s$ are odd.

Let $(\mathbf{s}_{1,x}^{(r)})^T, (\mathbf{s}_{2,z}^{(r)})^T$ represent the local polynomial smoothers for the r th derivatives at x and z , respectively, with $r < p$. In the case of x , this smoother can be written as $(\mathbf{s}_{1,x}^{(r)})^T = r! \mathbf{e}_{r+1}^T (\mathbf{X}_x^T \mathbf{W}_x \mathbf{X}_x)^{-1} \mathbf{X}_x^T \mathbf{W}_x$, where $\mathbf{e}_{r+1}^T$ is the $(p+1) \times 1$ vector with a 1 in its $r+1$ coordinate and 0's elsewhere and $\mathbf{W}_x, \mathbf{X}_x$ are as before (Ruppert and Wand [22]). A similar expression holds for $(\mathbf{s}_{2,z}^{(r)})^T$, and let $\mathbf{S}_1^{(r)}, \mathbf{S}_2^{(r)}$ represent the matrices whose rows are the derivative smoothers at the observations $\mathbf{X}$ and $\mathbf{Z}$. The vectors of estimators of the derivative functions $m_1^{(r)}$ and $m_2^{(r)}$ with respect to x and z , respectively, are defined in Opsomer and Ruppert [19] as

$$\hat{\mathbf{m}}_1^{(r)} = \mathbf{S}_1^{(r)}(\mathbf{Y} - \hat{\mathbf{m}}_2) \quad \hat{\mathbf{m}}_2^{(r)} = \mathbf{S}_2^{(r)}(\mathbf{Y} - \hat{\mathbf{m}}_1),$$

where $\hat{\mathbf{m}}_1$ and $\hat{\mathbf{m}}_2$ are estimated by an additive model with local polynomial regression of degree p . The derivative estimators also have an explicit expression:

$$\begin{aligned} \hat{\mathbf{m}}_1^{(r)} &= \mathbf{S}_1^{(r)}(\mathbf{I} - \mathbf{S}_2^* \mathbf{S}_1^*)^{-1}(\mathbf{I} - \mathbf{S}_2^*)\mathbf{Y} \equiv \mathbf{W}_1^{(r)}\mathbf{Y} \\ \hat{\mathbf{m}}_2^{(r)} &= \mathbf{S}_2^{(r)}(\mathbf{I} - \mathbf{S}_1^* \mathbf{S}_2^*)^{-1}(\mathbf{I} - \mathbf{S}_1^*)\mathbf{Y} \equiv \mathbf{W}_2^{(r)}\mathbf{Y}. \end{aligned} \quad (6)$$

Letting $\mathbf{T}_{21}^* = (\mathbf{T}_{12}^*)^T$, $\theta_{11}(r, s)$ can be rewritten as

$$\begin{aligned} \theta_{11}(r, s) &= \frac{1}{n} \text{tr} \left\{ \left((\mathbf{I} - \mathbf{T}_{12}^*)^{-1} D^r \mathbf{m}_1 - (\mathbf{I} - \mathbf{T}_{21}^*)^{-1} \mathbb{E}(m^{(r)}(X_i) | \mathbf{Z}) \right) \right. \\ &\quad \left. \left((\mathbf{I} - \mathbf{T}_{12}^*)^{-1} D^s \mathbf{m}_1 - (\mathbf{I} - \mathbf{T}_{21}^*)^{-1} \mathbb{E}(m^{(s)}(X_i) | \mathbf{Z}) \right) \right\} \end{aligned}$$

A “natural” estimator for $\theta_{11}(r, s)$ can be constructed by applying the approximation $\mathbf{S}_1^* \mathbf{S}_2^* \approx \mathbf{T}_{12}^*$ from Lemma 3.1 of Opsomer and Ruppert [19], as well as $\mathbb{E}(\mathbf{S}_2 \mathbf{m}_1) \approx \mathbb{E}(m_1(X_i) | \mathbf{Z})$ and $\mathbb{E}(W_1^{(r)} \mathbf{m}_1) \approx D^r \mathbf{m}_1$, which were computed in the proofs of their Theorems 4.1 and 5.2 :

$$\begin{aligned} \hat{\theta}_{11}(r, s) &= \frac{1}{n} \text{tr} \left\{ \left((\mathbf{I} - \mathbf{S}_1^* \mathbf{S}_2^*)^{-1} - (\mathbf{I} - \mathbf{S}_2^* \mathbf{S}_1^*)^{-1} \mathbf{S}_2^* \right) \mathbf{W}_1^{(r)*} \mathbf{Y} \mathbf{Y}^T \mathbf{W}_1^{(s)*T} \right. \\ &\quad \left. \left((\mathbf{I} - \mathbf{S}_1^* \mathbf{S}_2^*)^{-T} - \mathbf{S}_2^{*T} (\mathbf{I} - \mathbf{S}_2^* \mathbf{S}_1^*)^{-T} \right) \right\} \end{aligned} \quad (7)$$

A similar reasoning leads to the following estimators for $\theta_{12}(r, s)$ and $\theta_{22}(r, s)$:

$$\hat{\theta}_{12}(r, s) = \frac{1}{n} \text{tr} \left\{ \left((\mathbf{I} - \mathbf{S}_1^* \mathbf{S}_2^*)^{-1} - (\mathbf{I} - \mathbf{S}_2^* \mathbf{S}_1^*)^{-1} \mathbf{S}_2^* \right) \mathbf{W}_1^{(r)*} \mathbf{Y} \mathbf{Y}^T (\mathbf{W}_2^{(s)*})^T \right\}$$

$$\begin{aligned} \hat{\theta}_{22}(r, s) = & \frac{1}{n} \text{tr} \left\{ \left((\mathbf{I} - \mathbf{S}_2^* \mathbf{S}_1^*)^{-1} - (\mathbf{I} - \mathbf{S}_1^* \mathbf{S}_2^*)^{-1} \mathbf{S}_1^* \right) \mathbf{W}_2^{(r)*} \mathbf{Y} \mathbf{Y}^T \mathbf{W}_2^{(s)*T} \right. \\ & \left. \left((\mathbf{I} - \mathbf{S}_2^* \mathbf{S}_1^*)^{-T} - \mathbf{S}_1^{*T} (\mathbf{I} - \mathbf{S}_1^* \mathbf{S}_2^*)^{-T} \right) \right\}. \end{aligned} \quad (8)$$

This is a generalization of the approach used by Ruppert *et al.* [21] to estimate regression functionals for local polynomial fitting.

If $\mathbf{X}$ and $\mathbf{Z}$ are independent, $\theta_{11}(r, s)$, $\theta_{22}(r, s)$ and their estimators can be simplified to

$$\theta_{11}(r, s) = \frac{1}{n} \sum_{i=1}^n \left(m_1^{(r)}(X_i) - \mathbb{E}(m_1^{(r)}(X_i)) \right) \left(m_1^{(s)}(X_i) - \mathbb{E}(m_1^{(s)}(X_i)) \right)$$

and

$$\hat{\theta}_{11}(r, s) = \frac{1}{n} \text{tr} \mathbf{W}_1^{(r)*} \mathbf{Y} \mathbf{Y}^T (\mathbf{W}_1^{(s)*})^T, \quad (9)$$

with similar expressions holding for $\theta_{22}(r, s)$.

Before stating the asymptotic bias and variance results for the estimators (7) and (8), we introduce additional notation. Let g_1 and g_2 represent the bandwidth parameters corresponding to m_1 and m_2 , respectively. Note that the same bandwidths g_1, g_2 are used in the matrices $\mathbf{S}_1^{(r)}, \mathbf{S}_2^{(r)}$ and $\mathbf{S}_1^*, \mathbf{S}_2^*$ when computing additive model derivatives. Define $\mathbf{N}_p$ as the $(p+1) \times (p+1)$ matrix whose ij th element is equal to $\mu_{i+j-2}(K)$ and $\mathbf{M}_{r,p}(u)$ be the same as $\mathbf{N}_p$, but with the $(r+1)$ th column replaced by $(1, u, \dots, u^p)^T$. Define the kernel

$$K_{(r,p)}(u) = \{r! |\mathbf{M}_{r,p}(u)| / |\mathbf{N}_p| \} K(u)$$

and let $K_{(0,p)} \equiv K_{(p)}$. Let $\mu_j(K_{(r,p)}) = \int u^j K_{(r,p)}(u) du$, $\mu(K_{(r,p)}) \equiv \mu_0(K_{(r,p)})$ and $R(K_{(r,p)}) \equiv \mu(K_{(r,p)}^2)$. Finally, let D_x^k and D_z^k represent the k th order differential operators with respect to x and z , respectively.

For the asymptotic bias and variance results to hold, we need assumptions (AS.I) and (AS.II) from the previous subsection, as well as

(AS.III') As $n \rightarrow \infty$, $g_1, g_2 \rightarrow 0$ and $ng_1^{r+s+1}/\log(n), ng_2^{r+s+1}/\log(n) \rightarrow \infty$ and there exist two positive constants C_m, C_M , such that $C_m g_1 \leq g_2 \leq C_M g_1$ for all n .

(AS.IV') The $\max(p+1, r+s)$ derivatives of m_1 and m_2 exist and are bounded and continuous.

Under these assumptions, the estimators (7) and (8) are consistent, as the following theorem shows. These results will also allow us to state the optimal rate of convergence for the estimators and their bandwidth parameters.

Theorem 2.1 Suppose that $\|\mathbf{S}_1^* \mathbf{S}_2^*\| < 1$ for some matrix norm $\|\cdot\|$. For $p, r, s \geq 1$ and $p - r, p - s$ odd, the conditional bias and variance of $\hat{\theta}_{11}(r, s)$ are:

$$\begin{aligned} E(\hat{\theta}_{11}(r, s) - \theta_{11}(r, s) | \mathbf{X}, \mathbf{Z}) &= \frac{\mu_{p+1}(K_{(r,p)})}{(p+1)!} g_1^{p-r+1} \theta_{11}(s, p+1) \\ &+ \frac{\mu_{p+1}(K_{(r,p)})}{(p+1)!} g_1^{p-s+1} \theta_{11}(r, p+1) + \sigma^2 \mu(K_{(r,p)} K_{(s,p)}) \frac{b_x - a_x}{n g_1^{r+s+1}} \\ &+ o_p \left(g_1^{p-r+1} + g_1^{-r} g_2^{p+1} + g_1^{p-s+1} + g_1^{-s} g_2^{p+1} + g_2^{\frac{r+s}{2}} + \frac{1}{n g_1^{r+s+1}} \right) \end{aligned}$$

and

$$\begin{aligned} \text{Var}(\hat{\theta}_{11}(r, s) | \mathbf{X}, \mathbf{Z}) &= 2\sigma^2 \frac{R(K_{(r,p)} * K_{(s,p)})}{n^2 g_1^{2r+2s+1}} \frac{1}{n} \sum_{i=1}^n (\sigma^2 + m(X_i, Z_i)^2 + E(m(X_i, Z_i)^2 | X_i)) f_X(X_i)^{-1} \\ &+ 4\sigma^2 \frac{(r+s)!^2}{n^2} \sum_{i=1}^n f_X(X_i)^{-2} \int D_x^{r+s}(mf)(X_i, z) dz + o_p \left(\frac{1}{n^2 g_1^{2r+2s+1}} \right). \end{aligned}$$

The conditional bias and variance of $\hat{\theta}_{12}(r, s)$ are

$$\begin{aligned} E(\hat{\theta}_{12}(r, s) - \theta_{12}(r, s) | \mathbf{X}, \mathbf{Z}) &= \frac{\mu_{p+1}(K_{(r,p)})}{(p+1)!} g_1^{p-r+1} \theta_{12}(p+1, s) \\ &+ \frac{\mu_{p+1}(K_{(s,p)})}{(p+1)!} g_2^{p-s+1} \theta_{12}(r, p+1) + \sigma^2 \frac{r!s!}{n} \frac{1}{n} \sum_{i=1}^n \frac{D_x^r D_z^s f(X_i, z_i)}{f_X(X_i) f_Z(Z_i)} \\ &+ o_p \left(g_1^{p-r+1} + g_1^{-r} g_2^{p+1} + g_2^{p-s+1} + g_1^{p+1} g_2^{-s} + g_1^{\frac{r+s}{2}} + g_2^{\frac{r+s}{2}} + g_1^{\frac{r}{2}} g_2^{\frac{s}{2}} \right). \end{aligned}$$

and

$$\begin{aligned} \text{Var}(\hat{\theta}_{12}(r, s) | \mathbf{X}, \mathbf{Z}) &= \sigma^2 \frac{r!s!}{n^2} \sum_{i=1}^n \left\{ E_{X,Z} \left(D_x^r D_z^s \frac{f(X_i, Z)}{f_X(X_i) f_Z(Z)} m(X, Z) \right) \right. \\ &\left. + E_{X,Z} \left(D_x^r D_z^s \frac{f(X, Z_i)}{f_X(X) f_Z(Z_i)} m(X, Z) \right) \right\}^2 + O_p \left(\frac{1}{\sqrt{n^5 g_1^{4r+2} g_2^{4s+2}}} \right). \end{aligned}$$

The expressions for $\hat{\theta}_{22}(r, s)$ are completely analogous to those of $\hat{\theta}_{11}(r, s)$. To find the optimal rates of convergence for $\hat{\theta}_{11}$, $\hat{\theta}_{22}$ and $\hat{\theta}_{12}$ estimated with local cubic polynomials, we set $r, s = 2$ and $p = 3$ in Theorem 2.1. Their asymptotic MSE (AMSE) can easily be derived from the theorem and minimized with respect to the bandwidth parameters g_1 and g_2 . The following corollary gives the optimal rates of convergence for the estimators and their bandwidths.

Corollary 2.1 *The optimal convergence rate for the bandwidths of $\hat{\theta}_{11}$ and $\hat{\theta}_{22}$ are given by $g_{1,AMSE} = G_1 n^{-1/7}$, $g_{2,AMSE} = G_2 n^{-1/7}$ for $G_1, G_2 > 0$. At those rates, $\hat{\theta}_{11} = \theta_{11} + O_p(n^{-2/7})$, $\hat{\theta}_{22} = \theta_{22} + O_p(n^{-2/7})$. In the case of $\hat{\theta}_{12}$, $g_{1,AMSE} = G'_1 n^{-5/28}$, $g_{2,AMSE} = G'_2 n^{-5/28}$ for $G'_1, G'_2 > 0$, and $\hat{\theta}_{12} = \theta_{12} + O_p(n^{-5/14})$. If g_1 and g_2 are set at the rates optimal for $\hat{\theta}_{11}$ and $\hat{\theta}_{22}$, then also $\hat{\theta}_{12} = \theta_{12} + O_p(n^{-2/7})$.*

2.3 Variance Estimation

The other unknown quantity in the AMASE is the variance σ^2 , which we estimate by

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n \{Y_i - \hat{m}(X_i, Z_i)\}^2. \quad (10)$$

A more sophisticated variance estimator that includes a “degrees of freedom” adjustment is discussed in Opsomer [18]. Since the latter estimator cannot be computed through backfitting, only $\hat{\sigma}^2$ will be discussed here. Fitting $\hat{\sigma}^2$ requires the selection of two bandwidth parameters, κ_1 and κ_2 .

The assumptions required for analysis of $\hat{\sigma}^2$ are (AS.I)–(AS.III) and

(AS.IV'') The $(p+1)$ th derivatives of m_1 and m_2 exist and are bounded and continuous.

We compute asymptotic approximations to the conditional bias and variance of $\hat{\sigma}^2$ in the following theorem.

Theorem 2.2 *Suppose that $\|\mathbf{S}_1^* \mathbf{S}_2^*\| < 1$ for some matrix norm $\|\cdot\|$. Assume p is an odd integer. For the homoskedastic bivariate additive model with local polynomial terms of degree p , the conditional bias and variance of $\hat{\sigma}^2$ are*

$$\begin{aligned} E(\hat{\sigma}^2 - \sigma^2 | \mathbf{X}, \mathbf{Z}) &= \left(\frac{\mu_{p+1}(K_{(p)})}{(p+1)!} \right)^2 \left(\kappa_1^{2p+2} \theta_{11}(p+1, p+1) \right. \\ &\quad \left. + \kappa_1^{p+1} \kappa_2^{p+1} \theta_{12}(p+1, p+1) + \kappa_2^{2p+2} \theta_{22}(p+1, p+1) \right) \\ &\quad + \sigma^2 (R(K_{(p)}) - 2K_{(p)}(0)) \left(\frac{b_x - a_x}{n\kappa_1} + \frac{b_z - a_z}{n\kappa_2} \right) + o_p(\kappa_1^{2p+2} + \kappa_2^{2p+2} + \frac{1}{n\kappa_1} + \frac{1}{n\kappa_2}) \end{aligned}$$

and

$$\begin{aligned} \text{Var}(\hat{\sigma}^2 | \mathbf{X}, \mathbf{Z}) &= \frac{1}{n} \text{Var}(\varepsilon_i^2) + 2 \left((R(K_{(p)}) - 2K_{(p)}(0)) \text{Var}(\varepsilon_i^2) \right. \\ &\quad \left. + 2R(K_p * K_p - 2K_p)\sigma^4 \right) \left(\frac{b_x - a_x}{n^2\kappa_1} + \frac{b_z - a_z}{n^2\kappa_2} \right) + o_p\left(\frac{1}{n^2\kappa_1} + \frac{1}{n^2\kappa_2}\right). \end{aligned}$$

As in the previous subsection, we can easily find the optimal convergence rates of this estimator. We consider the case $p = 1$ in the following corollary:

Corollary 2.2 *The optimal convergence rate for the bandwidths of $\hat{\sigma}^2$ are given by $\kappa_{1,AMSE} = K_1 n^{-1/5}$, $\kappa_{2,AMSE} = K_2 n^{-1/5}$ for $K_1, K_2 > 0$. With the bandwidths at these rates, $\hat{\sigma}^2 = \sigma^2 + O_p(n^{-1/2})$.*

2.4 Plug-in Bandwidth Estimators

We will first consider the case where $\mathbf{X}$ and $\mathbf{Z}$ are independent, since explicit expressions are available for the bandwidth estimators $\hat{h}_1, \hat{h}_2$ by replacing the unknown quantities by their estimators in expression (5). First, note that

$$\theta_{11}(r, s) \rightarrow \theta_{11}^a(r, s) = E \left(\{m_1^{(r)}(X_i) - E(m_1^{(r)}(X_i))\} \{m_1^{(s)}(X_i) - E(m_1^{(s)}(X_i))\} \right)$$

as $n \rightarrow \infty$ and similarly for $\theta_{22}(r, s)$. Both quantities converge at the rate $O_p(n^{-1/2})$.

As before, the bandwidth parameters used to fit $\hat{\sigma}^2$ will be denoted by κ_1, κ_2 and those for $\hat{\theta}_{11}, \hat{\theta}_{22}$ by g_1, g_2 . We need to assume that the second derivatives of m are not constant over $\text{supp}(f)$, i.e., that $\theta_{11}^a, \theta_{22}^a \neq 0$. For simplicity, we only consider the case where the $\hat{\theta}_{ii}$ and $\hat{\sigma}^2$ are estimated by local polynomial regressions of degree 3 and 1, respectively. We can now state the following theorem:

Theorem 2.3 *Suppose $\mathbf{X}$ and $\mathbf{Z}$ are independent. Let $\kappa_1 = K_1 n^{-1/5}$, $\kappa_2 = K_2 n^{-1/5}$, $g_1 = G_1 n^{-1/7}$ and $g_2 = G_2 n^{-1/7}$ for some constants $K_1, K_2, G_1, G_2 > 0$. Then,*

$$n^{2/7} \frac{\hat{h}_1 - h_{1,MASE}}{h_{1,MASE}} \rightarrow_P D_1, \quad n^{2/7} \frac{\hat{h}_2 - h_{2,MASE}}{h_{2,MASE}} \rightarrow_P D_2$$

where

$$\begin{aligned} D_1 &= -\frac{1}{5} \theta_{11}^{a-1} \left\{ 2 \frac{\mu_4(K_{(2,3)})}{(p+1)!} \theta_{11}^a(2, 4) G_1^2 + \sigma^2 R(K_{(2,3)})(b_x - a_x) G_1^{-5} \right\} \\ D_2 &= -\frac{1}{5} \theta_{22}^{a-1} \left\{ 2 \frac{\mu_4(K_{(2,3)})}{(p+1)!} \theta_{22}^a(2, 4) G_2^2 + \sigma^2 R(K_{(2,3)})(b_x - a_x) G_2^{-5} \right\}. \end{aligned}$$

The theorem shows that, under independence of the covariates, the proposed plug-in estimator for the bivariate additive model has the same asymptotic properties as the 1-dimensional local linear regression estimator of Ruppert *et al.* [21]. Replacing $\hat{\sigma}^2$ by the more sophisticated variance estimator referred to in Section 2.3 has no effect on the convergence rates of $\hat{h}_1, \hat{h}_2$ (see Opsomer [18]).

If $\mathbf{X}$ and $\mathbf{Z}$ are not independent, we cannot write down explicit expressions for $\hat{h}_1$ and $\hat{h}_2$, since they are only defined as the solutions to the three-step process from Subsection 2.1. Nevertheless, we can prove a number of properties about the bandwidth estimators.

Theorem 2.4 *Let $\kappa_1 = K_1 n^{-1/5}$, $\kappa_2 = K_2 n^{-1/5}$, $g_1 = G_1 n^{-1/7}$ and $g_2 = G_2 n^{-1/7}$ for some constants $K_1, K_2, G_1, G_2 > 0$. Then, $h_{1,AMASE} = H_1 n^{-1/5}$, $h_{2,AMASE} = H_2 n^{-1/5}$ for some $H_1, H_2 > 0$ and*

$$\frac{\hat{h}_1 - h_{1,AMASE}}{h_{1,AMASE}}, \frac{\hat{h}_2 - h_{2,AMASE}}{h_{2,AMASE}} = O_p(n^{-2/7}).$$

Even if $\mathbf{X}$ and $\mathbf{Z}$ are not independent, bandwidth estimators $\hat{h}_1$ and $\hat{h}_2$ can therefore be constructed, so that they have very similar convergence rate properties as in (non-additive) local linear regression. The principal difference is that the convergences can only be proven relative to $h_{1,AMASE}$ and $h_{2,AMASE}$, instead of $h_{1,MASE}$ and $h_{2,MASE}$ as in Theorem 2.3.

2.5 Extension to D Independent Covariates

In addition to having good theoretical properties, a *practical* bandwidth selection method should also be reasonably fast and applicable to as wide a range of situations as possible. Clearly, this might require giving up some estimation accuracy. In the development of our bandwidth selection method, we therefore decided to work under the assumption of independence of the covariates, fully realizing that this assumption would often be violated in practice. Nevertheless, it allows us to use much simpler expressions for the estimators, all of which are now easily computed using backfitting, and enables us to expand the method from two covariates to any number of them. As the simulations in Section 4 will show, the optimal choice of bandwidth appears quite insensitive to the lack of independence between the covariates.

If the covariates are mutually independent, Corollary 4.3 of Opsomer and Ruppert [19] showed that the asymptotic bias and variance of the bivariate additive model are equivalent to those of two separately fitted local linear regressions, except for the “centering effect,” and their Corollary 5.1 extended this to local polynomials of odd degree. We now look at the model

$$Y_i = \alpha + m_1(X_{1i}) + \dots + m_D(X_{Di}) + \varepsilon_i$$

with $E(m_d(X_{di})) = 0$ for $d = 1, \dots, D$, the ε_i independent and identically distributed with mean 0 and variance σ^2 , and all covariates X_{di} independent of each other.

We will fit this model by local polynomial regression of odd degree. Let $\mathbf{S}_d$ represent the smoother matrix for the local polynomial of odd degree p_d corresponding to the d th covariate, for $d = 1, \dots, D$. In addition to (AS.I), we require the following assumptions:

(**AS.II'**) The densities $f_{X_d}, d = 1, \dots, D$, are continuous and have compact support (a_d, b_d) . $f_{X_d}(x) > 0$ for all $x \in \text{supp}(f_{X_d})$ for each d .

(**AS.III''**) As $n \rightarrow \infty$, $h_d \rightarrow 0$ and $nh_d \rightarrow \infty$ for all d .

(**AS.IV'''**) The $(p_d + 1)$ th derivatives of $m_d, d = 1, \dots, D$ exist and are continuous.

The estimators $\hat{\mathbf{m}}_1, \dots, \hat{\mathbf{m}}_D$ are defined as the solutions to the normal equations:

$$\begin{bmatrix} \mathbf{I} & \mathbf{S}_1^* & \cdots & \mathbf{S}_1^* \\ \mathbf{S}_2^* & \mathbf{I} & \cdots & \mathbf{S}_2^* \\ \vdots & & \ddots & \vdots \\ \mathbf{S}_D^* & \mathbf{S}_D^* & \cdots & \mathbf{I} \end{bmatrix} \begin{bmatrix} \hat{\mathbf{m}}_1 \\ \hat{\mathbf{m}}_2 \\ \vdots \\ \hat{\mathbf{m}}_D \end{bmatrix} = \begin{bmatrix} \mathbf{S}_1^* \\ \mathbf{S}_2^* \\ \vdots \\ \mathbf{S}_D^* \end{bmatrix} \mathbf{Y},$$

and $\hat{\alpha} = \bar{Y}$. These estimators are most easily computed through the backfitting algorithm. They exist and are unique if the matrix

$$\mathbf{M} = \begin{bmatrix} \mathbf{I} & \mathbf{S}_1^* & \cdots & \mathbf{S}_1^* \\ \mathbf{S}_2^* & \mathbf{I} & \cdots & \mathbf{S}_2^* \\ \vdots & & \ddots & \vdots \\ \mathbf{S}_D^* & \mathbf{S}_D^* & \cdots & \mathbf{I} \end{bmatrix} \quad (11)$$

is invertible. If $D = 2$, Lemmas 3.2 and 5.1 of Opsomer and Ruppert [19] guarantee that this is indeed the case for sufficiently large n . The same approach cannot be used to prove existence for arbitrary D using the current approach, however. We therefore add the assumption

(**AS.V**) The matrix $\mathbf{M}$ is invertible.

In the following theorem, we compute the asymptotic approximations to the conditional bias and variance of the estimators for the additive component functions. Since the plug-in bandwidth selection method ignores boundary effects, we also state the theorem without including them. For a result that encompasses them, see Opsomer [18].

Theorem 2.5 *If $\mathbf{X}_1, \dots, \mathbf{X}_D$ are mutually independent, the conditional bias and variance of the additive model estimators $\hat{m}_d(X_{di})$, $d = 1, \dots, D$, $i = 1, \dots, n$ are*

$$\begin{aligned} \mathbb{E}(\hat{m}_d(X_{di}) - m_d(X_{di}) | \mathbf{X}_1, \dots, \mathbf{X}_D) = \\ \frac{h_d^{p_d+1}}{(p_d+1)!} \mu_{p_d+1}(K_{(p_d)}) \left(m_d^{(p_d+1)}(X_{di}) - \mathbb{E}(m_d^{(p_d+1)}(X_{di})) \right) + O_p\left(\frac{1}{\sqrt{nh_d}}\right) + o_p\left(\sum_{k=1}^D h_k^{p_k+1}\right) \end{aligned}$$

and

$$\text{Var}(\hat{m}_d(X_{di}) | \mathbf{X}_1, \dots, \mathbf{X}_D) = \frac{1}{nh_d} f_{X_d}(X_{di})^{-1} R(K_{(p_d)}) + o_p\left(\frac{1}{nh_d}\right).$$

The AMASE for the estimated function $\hat{m} = \hat{\alpha} + \sum_{d=1}^D \hat{m}_d$ is given in the following corollary.

Corollary 2.3 *The conditional MASE for the additive model with D independent covariates is approximated by*

$$\begin{aligned} \text{MASE}(\mathbf{h} | \mathbf{X}_1, \dots, \mathbf{X}_D) = \sum_{d=1}^D \frac{\mu_{p_d+1}(K_{(p_d)})^2}{(p_d+1)!} h_d^{2p_d+2} \theta_{dd}(p_d+1, p_d+1) \\ + \sigma^2 \sum_{d=1}^D R(K_{(p_d)}) \frac{b_d - a_d}{nh_d} + O_p\left(\sum_{d=1}^D \frac{1}{nh_d}\right) + o_p\left(\sum_{d=1}^D \left(h_d^{2p_d+2} + \frac{1}{nh_d}\right)\right). \end{aligned}$$

When $D = 2$, Corollary 5.1 of Opsomer and Ruppert [19] did not contain the $O_p(1/\sqrt{nh_d})$ bias term for the backfitting estimator, making it likely that this term is actually 0 and that its presence in the result for general D is only due to the inability of the current approach to adequately deal with the matrix $\mathbf{M}$ when $D > 2$. Therefore, although this term is of the same order as the asymptotic standard deviation of $\hat{m}_d$, we will ignore it for the purpose of developing a plug-in bandwidth selection method and use

$$h_{d,AMASE} = \left(\frac{(p_d+1) p_d!^2 R(K_{(p_d)}) \sigma^2 (b_d - a_d)}{2n \mu_{p_d+1}(K_{(p_d)})^2 \theta_{dd}(p_d+1, p_d+1)} \right)^{\frac{1}{2p_d+3}} \quad (12)$$

as the “minimizer” of the AMASE.

In order to estimate the $\theta_{dd}(p_d+1, p_d+1)$ and σ^2 with sufficient accuracy for the plug-in, we first need to find the asymptotically optimal bandwidths for both estimators and fit separate additive models for each. To simplify notation, write $\mathbf{p} = (p_1, \dots, p_D)$ for the degree of the local polynomials and let $\hat{\mathbf{h}} = (\hat{h}_1, \dots, \hat{h}_D)$, $\boldsymbol{\theta}(\mathbf{p} + \mathbf{1}, \mathbf{p} + \mathbf{1}) = (\theta_{11}(p_1+1, p_1+1), \dots, \theta_{DD}(p_D+1, p_D+1))$ and similarly for $\hat{\boldsymbol{\theta}}(\mathbf{p} + \mathbf{1}, \mathbf{p} + \mathbf{1})$. Let $\mathbf{g} = (g_1, \dots, g_D)$

and $\boldsymbol{\kappa} = (\kappa_1, \dots, \kappa_D)$ represent the bandwidth parameters for estimating $\hat{\boldsymbol{\theta}}(\mathbf{p} + \mathbf{1}, \mathbf{p} + \mathbf{1})$ and $\hat{\sigma}^2$, respectively.

We first consider estimating $\hat{\boldsymbol{\theta}}(\mathbf{p} + \mathbf{1}, \mathbf{p} + \mathbf{1})$ by fitting the additive model with local polynomials of degree $\mathbf{p} + \mathbf{2}$. A straightforward generalization of Theorem 2.1 to the multivariate independent case shows that the conditional Asymptotic Mean Squared Error (AMSE) of $\hat{\theta}_{dd}(p_d + 1, p_d + 1)$ is minimized by

$$g_{d,AMSE} = \left(C_{\theta}(p_d) \frac{(p_d + 3)! R(K_{(p_d+1, p_d+2)}) \sigma^2(b_d - a_d)}{2n |\theta_{dd}(p_d + 1, p_d + 3)| \mu_{p_d+3}(K_{(p_d+1, p_d+2)})} \right)^{\frac{1}{2p_d+5}} \quad (13)$$

for $d = 1, \dots, D$, where

$$C_{\theta}(p_d) = \begin{cases} 1 & \text{if } \theta_{dd}(p_d + 1, p_d + 3) < 0 \\ \frac{2p_d+3}{2} & \theta_{dd}(p_d + 1, p_d + 3) > 0. \end{cases}$$

In the case of $\hat{\sigma}^2$, we generalize Theorem 2.2 to the multivariate independent case. For given local polynomials of degree $\mathbf{p}$, we find that the AMSE is minimized by

$$\kappa_{d,AMSE} = \left(C_{\sigma}(p_d) \frac{(p_d + 1)!^2 |R(K_{(p_d)}) - 2K_{(p_d)}(0)| \sigma^2(b_d - a_d)}{n \theta_{dd}(p_d + 1, p_d + 1) \mu_{p_d+1}(K_{(p_d)})^2} \right)^{\frac{1}{2p_d+3}} \quad (14)$$

for $d = 1, \dots, D$, where

$$C_{\sigma}(p_d) = \begin{cases} 1 & \text{if } R(K_{(p_d)}) - 2K_{(p_d)}(0) < 0 \\ \frac{1}{2p_d+2} & R(K_{(p_d)}) - 2K_{(p_d)}(0) > 0. \end{cases}$$

3 Bandwidth Selection and Model Fitting Algorithm

We are now ready to outline the proposed methodology. The bandwidth selection component will be based on the *Direct Plug-In* (DPI) method of Ruppert *et al.* [21], while the computation of the additive fits will rely on the backfitting algorithm of Buja *et al.* [5]. The computer code for the implementation of the method is written in the *Matlab* v4.2c [17] programming environment and is available from the first author.

The outline of the proposed method is depicted in Figure 1. In Step 1, a piecewise polynomial regression is performed to provide crude estimates of σ^2 and $\boldsymbol{\theta}(\mathbf{p} + \mathbf{1}, \mathbf{p} + \mathbf{3})$, with the number of “pieces” N_d for each covariate selected by minimizing Mallows’ C_p . For each value of $\mathbf{N} = (N_1, \dots, N_D)$, this is a straightforward parametric fit and computations are very fast. To avoid overfitting, we set $N_d \leq N_{max} = 5$. An exhaustive search for the optimal $\mathbf{N}$ would take $(N_{max})^D$ separate piecewise regressions, which

Figure 1: Bandwidth selection and model fitting algorithm

rapidly becomes prohibitive as D increases. We therefore replace the exhaustive search by the following simple method:

1. Let $\mathbf{N}^{(0)} = (1, \dots, 1)$,
2. For $d = 1, \dots, D$: find the minimizer $N_d^{(t)}$ of C_p for each covariate, holding all others at $N_k = N_k^{(t-1)}$,
3. Repeat (2) until $\mathbf{N}^{(t)} = \mathbf{N}^{(t-1)}$.

Unless the covariates are strongly correlated, this search typically converges after 2 or 3 iterations.

In Step 2 of Figure 1, the bandwidths $\hat{\mathbf{g}}$ are computed by plugging estimates $\hat{\sigma}_P^2$ and $\hat{\boldsymbol{\theta}}(\mathbf{p} + \mathbf{1}, \mathbf{p} + \mathbf{3})$ into expression (13) and are used to fit an additive model with local

polynomials of degree $\mathbf{p} + \mathbf{2}$. The estimators $\hat{\boldsymbol{\theta}}(\mathbf{p} + \mathbf{1}, \mathbf{p} + \mathbf{1})$ are computed according to (9). Step 3 plugs $\hat{\boldsymbol{\theta}}(\mathbf{p} + \mathbf{1}, \mathbf{p} + \mathbf{1})$ and $\hat{\sigma}_p^2$ into (14) to compute $\hat{\mathbf{k}}$ and uses these bandwidths to fit an additive model with local polynomials of degree $\mathbf{p}$ and compute $\hat{\sigma}^2$ as in definition (10). Finally, Step 4 plugs $\hat{\boldsymbol{\theta}}(\mathbf{p} + \mathbf{1}, \mathbf{p} + \mathbf{1})$ and $\hat{\sigma}^2$ into (12) and fits an additive model with local polynomials of degree $\mathbf{p}$.

As can be seen in equations (12), (13) and (14), these bandwidths depend on the distribution of the observations only through their ranges, $(b_d - a_d)$. Singularity problems can arise in regions of sparse data, since a minimum of $p + 1$ observations are needed to fit a polynomial of degree p . Each of the computed bandwidths are therefore compared to the size of the intervals between the observations, and if necessary increased. We will call this increase the *data distribution adjustment*.

In each of the Steps 2–4, the additive model is fitted with the backfitting algorithm. The usual backfitting algorithm requires computation of $3Dn$ separate regressions (for the construction of $3D$ smoother matrices) and can be quite slow for large datasets. It can be speeded up dramatically with only a small loss in accuracy by estimating the functions $m_d(\cdot)$ over a regularly spaced grid and interpolating from that grid to the observation points to find $\hat{\mathbf{m}}_d$ (Fan and Marron [10]). For a grid with M points, the smoother matrices are $M \times n$, so that the number of regressions is limited to $3DM$. If we write $\mathbf{L}_d, d = 1, \dots, D$ for the $n \times M$ interpolation matrices, backfitting cycles through the following set of equations:

$$\begin{aligned} \hat{\mathbf{m}}_{1,M} &= \mathbf{C}\mathbf{L}_1\mathbf{S}_{1,M}(\mathbf{Y} - \boldsymbol{\alpha} - \sum_{k \neq 1} \hat{\mathbf{m}}_{k,M}) \\ &\vdots \\ \hat{\mathbf{m}}_{D,M} &= \mathbf{C}\mathbf{L}_D\mathbf{S}_{D,M}(\mathbf{Y} - \boldsymbol{\alpha} - \sum_{k \neq D} \hat{\mathbf{m}}_{k,M}) \end{aligned}$$

where the $\mathbf{S}_{d,M}$ are the smoother matrices over the gridpoints and $\mathbf{C} = \mathbf{I} - \mathbf{1}\mathbf{1}^T/n$ guarantees that centering is preserved. For the present application, we decided to set $M = 100$ and use cubic interpolation.

In addition to its theoretical properties discussed in Section 2, this plug-in algorithm should provide a dramatic increase in speed over cross-validation methods in estimating the optimal bandwidth. As shown in Figure 1, it requires exactly 3 recomputations of the additive model for any number of covariates, which compares very favorably with a D -dimensional grid search as required by cross-validation.

4 Simulations

In this section, we report one set of simulation experiments in which a bivariate model is fitted by local linear regression, with the correlation varied to evaluate robustness to departures from independent covariates. Experiments with more than two covariates and higher degree local polynomials can be found in Opsomer [18]. We use the following example functions:

$$m_1(t) = 1 - 6t + 36t^2 - 53t^3 + 22t^5 \quad , \quad m_2(t) = \sin(5\pi t).$$

These functions are similar to the functions #1 and #2 in Ruppert *et al.* [21]. As the first part of Table 1 shows, the functions also have very different amounts of curvature (as measured by θ) and different asymptotically optimal bandwidths. Assigning the same amount of smoothness to each component function would clearly not work well here. This is nevertheless common practice today, and is advocated by several authors (e.g. Hastie and Tibshirani [16], Chambers and Hastie [6]).

In this experiment, we evaluate five levels of correlation: independence (or 0), -0.25 , -0.5 , 0.25 and 0.5 . The covariates are generated from a joint distribution with marginals $\mathcal{N}(1/2, 1/9)$ and the desired correlation. Since the domain of the covariates is assumed to be bounded by (AS.II), we rejected all observations for which one of the covariates fell outside $\pm 1.5\sigma$ (or, equivalently, outside the interval $[0, 1]$), and replaced them by new observations. From Figure 1 in Opsomer and Ruppert [19], the correlations $-0.5, 0.5$ are outside the bounds set by assumption (A.II), but this did not appear to affect the convergence. This supports our view that correlation within these bounds, though sufficient, is not necessary for convergence of backfitting. The data were generated by the model $Y_i = m_1(X_i) + m_2(Z_i) + \varepsilon_i$, where the errors ε_i are distributed $\mathcal{N}(0, 1)$. We consider samples of 200 and 500 observations, each with 400 replicates. The densities for various estimators of interest were estimated by using the EBBS algorithm (Ruppert [20]) on these replicates.

Of primary interest is the behavior of the estimators $\hat{h}_1$ and $\hat{h}_2$ relative to their target values. Therefore, a grid search is performed to approximate the true optimal bandwidth parameters $h_{1,MASE}$ and $h_{2,MASE}$. For a grid of values for h_1, h_2 , the Average Squared Error ($ASE = \frac{1}{n} \sum_{i=1}^n (\hat{m}(X_i, Z_i) - m(X_i, Z_i))^2$) is computed for 400 replicates and averaged to estimate the MASE. The function $MASE(h_1, h_2)$ is then approximated by using cubic interpolation over the grid and the estimates of $h_{1,MASE}$ and $h_{2,MASE}$ minimize this function. The second part of Table 1 contains the estimated $h_{1,MASE}$ and $h_{2,MASE}$ for the different correlation levels. It is clear that the optimal bandwidths are

		m_1		m_2	
		$n = 200$	$n = 500$	$n = 200$	$n = 500$
Under independence:	$\theta(2, 2)$	2,182		30,387	
	h_{AMASE}	0.128	0.107	0.076	0.063
Estimated h_{MASE} for:	$\rho = 0$	0.168	0.128	0.080	0.066
	$\rho = -0.25$	0.168	0.132	0.080	0.066
	$\rho = -0.5$	0.170	0.130	0.080	0.064
	$\rho = 0.25$	0.166	0.130	0.082	0.066
	$\rho = 0.5$	0.174	0.128	0.080	0.064

Table 1: Values for θ , h_{AMASE} and Monte Carlo estimates of h_{MASE}

quite insensitive to the amount of correlation between the covariates, at least in the range considered here.

Figure 2 shows the densities of $\log(\hat{h}_i) - \log(h_{i,MASE})$ for both bandwidth estimates for the five correlation values at sample sizes of 200 and 500. The densities for the different correlation levels are not individually labelled, since they are so close together. The same conclusion is further reinforced by Figure 3, where the densities of the ASE corresponding to the models fit with $\hat{h}_1$ and $\hat{h}_2$ remain virtually unchanged for both sample sizes at the different correlation levels.

The simulations also allow us to compare the asymptotic approximation AMASE with the true MASE. As can be seen from Table 1, the difference between h_{AMASE} and h_{MASE} is much larger for m_1 than for m_2 . It is therefore not surprising that the bandwidths estimated by the proposed method in Figure 2 also display a much larger bias for the former function. This bias becomes smaller for the larger sample size, but is still evident. In the case of m_2 , the densities display less variability and much less bias at both sample sizes. Since m_1 is a polynomial of low degree while m_2 is a sine function, these results are somewhat counter-intuitive. A likely cause for this behavior is the fact that h_{AMASE} ignores boundary effects, while the finite-sample h_{MASE} does not. Since the larger bandwidth for m_1 results in a bigger boundary region than for m_2 , this could help explain the difference in behavior observed in this experiment.

Ultimately, a good bandwidth estimate should lead to accurate estimation of the unknown functions. Figure 4 shows five “typical” fitted functions for both sample sizes as well as the underlying functions, centered around their mean. For these graphs, we selected $\rho = 0$, but the curves would look the same for any of the other correlation levels considered. While the basic shape of g_1 and g_2 are reproduced by the estimated curves, the smaller sample size still leads to a significant amount of variability in the estimates.

Figure 2: Densities of the bandwidth estimators at five correlation levels.

At $n = 500$, the estimated curves are quite close to the true functions (except at the boundaries).

5 Example

The proposed algorithm is used on a real dataset to demonstrate its behavior in a typical application, as well as to point out some of the adjustments that can be made to the analysis to improve the fitted functions. We will also describe the *additive variable plot*, a useful diagnostic tool to visually check whether patterns observed in a fitted function are likely to be real or spurious. The full dataset consists of the median value of homes in 506 census tracts in the Boston Standard Metropolitan Statistical Area in 1970 and 13 accompanying socio-demographic and related variables. It was originally studied in Harrison and Rubinfeld [15], who estimated a marginal willingness-to-pay model for

Figure 3: Densities of the Average Squared Error at five correlation levels.

housing. Belsley *et al.* [1] provide a complete listing of the data and use it as an example for regression diagnostics, and Breiman and Friedman [3] analyze it with ACE. The dataset is available online on Statlib.

When the additive model with local linear terms is run on the full dataset, the severe collinearity in the data meant that even the OLS “pilot” estimate of Step 1 of our method results in a singular design matrix. When this step is avoided by using fixed values for $\theta(2, 4)$ and σ_P^2 , the backfitting algorithm still does not converge, further indicating the presence of *concurvity* between the variables. Concurvity is a term coined by Buja *et al.* [4] and refers to nonlinear dependencies which lead to degeneracy (non-uniqueness of the estimators) in additive models. It has recently been further explored by Donnell *et al.* [8]. One way around the concurvity problem is to use a stepwise method that selects a subset of the variables based on some criterion that maximizes the correlation between the fitted values and the selected covariates, with a penalty for overfitting. Breiman

Figure 4: Examples of estimated functions (solid lines) and the true functions (dash-dot lines).

and Friedman [3] applied their ACE method for this purpose on the data, and selected four variables. We will use the same subset for the current analysis. The dependencies between these variables are still quite high, with all correlations still above 0.40. This did not appear to cause problems in fitting the additive model.

The dependent variable and the covariates of interest are:

- MV: median value of owner-occupied homes (in \$000),
- RM: average number of rooms in owner unit,
- TAX: full property tax rate (\$/\$10,000),
- PTRATIO: pupil/teacher ratio by town school district,
- LSTAT: proportion of population that is of “lower status” (%),

which are further defined in Harrison and Rubinfeld [15].

Figure 5: Scatterplots of the Boston Housing data

In an original application of the algorithm to these data (see Opsomer [18]), a number of problems became apparent. Six outliers were identified, all belonging to two upscale Boston neighborhoods. Of more serious concern was the presence of large regions with little or no data in the ranges of TAX and LSTAT, causing a large data distribution adjustment in the computed bandwidths. For the current analysis, we decided to remove the outliers and perform a logarithmic transformation on both “problem” covariates before fitting the model. The effect of the logarithmic transformation is to reduce the gap between the observations and presumably provide a better bandwidth estimate. The fitted functions can easily be expressed in the original scales by reversing the transformation. In order to improve the stability of the matrix operations in the algorithm, we rescaled all the variables to the $[0,1]$ interval, which is also easily reversed to express the results in the original scales. Figure 5 shows the scatterplots for the dataset after these

adjustments, and we will fit the additive model

$$\begin{aligned} E(MV|RM, TAX, PTRATIO, LSTAT) = & \alpha + m_1(RM) + m_2(\log(TAX)) \\ & + m_3(PTRATIO) + m_4(\log(LSTAT)) \end{aligned}$$

using local linear terms.

The selected bandwidth parameters are given in Table 2. To make comparison of bandwidths across the different variables meaningful, the bandwidths in the transformed $([0,1])$ scale are also provided in parentheses. Figure 6 shows the estimated additive component functions. The residual plots, the normal probability plots and plots of the residuals against the covariates displayed no serious problems with the residuals (see Opsomer [18]).

Variable	Bandwidth
RM	0.6816 (0.1320)
$\log(TAX)$	$0.2199^{\ddagger}$ (0.3792)
PTRATIO	1.669 (0.1776)
$\log(LSTAT)$	$0.1932^{\ddagger}$ (0.1440)

‡ In the logarithmic scale.

Table 2: Bandwidth and variance estimates for Boston Housing data.

The bandwidth parameter for $\log(TAX)$ remains inflated by the data distribution adjustment, as the gap between the observations is still present. This type of problems is due to the use of a single (global) bandwidth parameter for the whole range of each variable, and is unlikely to occur with variable bandwidth estimates, which are becoming increasingly common for local polynomial regression (e.g. Fan and Gijbels [9], Ruppert [20]).

After fitting a model with any nonparametric regression technique, it is sometimes unclear whether apparent features of the fitted functions are “real.” As an example, consider the spike in the lower end of $\hat{m}_3$. Is this a real feature of the data or only due to boundary effect variability? A useful tool for evaluating this kind of behavior in the estimated functions is a variant of the *added variable plots* (Cook and Weisberg [7]), in which the additive model “residuals”

$$\hat{A}_{di} = Y_i - \hat{\alpha} - \sum_{k \neq d} \hat{m}_k(X_{ki})$$

are plotted against each covariate d . By comparing this “residual” with the backfitting algorithm as described in Hastie and Tibshirani [16], it is easy to see that $\hat{\mathbf{m}}_d = \mathbf{S}_d^* \hat{\mathbf{A}}_d$

Figure 6: Fitted additive functions for Boston Housing dataset

for $d = 1, \dots, D$, after convergence of the algorithm. We will call the $\hat{A}_{di}$ the *additive residuals* and the plots of the additive residuals against the individual covariates the *additive variable plots*. This diagnostic is very similar to the AMALL residual plots recently proposed in Berk and Booth [2]. The additive variable plots are shown in Figure 7. Since these plot contains the “residuals” after removal of the effects of the other covariates, they can be directly compared to the estimated functions. The “kink” in the PTRATIO function appears indeed caused by the three lowest observations and hence is unlikely to reflect a fundamental departure from the otherwise monotone decreasing trend in the data.

The functions in Figure 6 are similar to those found by Breiman and Friedman [3]. This is not too surprising, as ACE is also based on local linear fits. The variables RM and $LSTAT$ appear to explain most of the variation seen in housing prices in the restricted

Figure 7: Additive variable plots for Boston Housing data.

dataset, since the ranges of their fitted functions are much larger than those of the other covariates. One of the uses for nonparametric regression, and for additive modelling in particular, is as exploratory analysis for the selection of an appropriate parametric model for inference. In this case, a linear term for PTRATIO and logarithmic terms for TAX and LSTAT seem reasonable based on Figure 6, while the choice for RM is less clear.

A Appendix: Outlines of Proofs

Detailed proofs are available in Opsomer [18].

Outline of proof of Theorem 2.1: We only prove the results for $\hat{\theta}_{11}(r, s)$. Its expectation is equal to

$$\begin{aligned} E(\hat{\theta}_{11}(r, s)) &= \frac{1}{n} \text{tr} \left\{ [(I - S_1^* S_2^*)^{-1} - (I - S_2^* S_1^*)^{-1} S_2^*] W_1^{(r)*} E(Y) E(Y)^T \right. \\ &\quad \left. W_1^{(s)*T} [(I - S_1^* S_2^*)^{-T} - S_2^{*T} (I - S_2^* S_1^*)^{-T}] \right\} \\ &\quad + \sigma^2 \frac{1}{n} \text{tr} \left\{ [(I - S_1^* S_2^*)^{-1} - (I - S_2^* S_1^*)^{-1} S_2^*] W_1^{(r)*} W_1^{(s)*T} \right. \\ &\quad \left. [(I - S_1^* S_2^*)^{-T} - S_2^{*T} (I - S_2^* S_1^*)^{-T}] \right\}. \end{aligned} \quad (15)$$

Write $\theta_{11}(k, l) = \frac{1}{n} G_1^{(k)} G_1^{(l)T}$, with $G_1^{(k)} = (I - T_{12}^*)^{-1} m_1^{(k)} - (I - T_{21}^*)^{-1} E(m_1^{(k)}(X_i) | Z)$. Using results from Theorem 5.3 from Opsomer and Ruppert [19],

$$\begin{aligned} [(I - S_1^* S_2^*)^{-1} - (I - S_2^* S_1^*)^{-1} S_2^*] W_1^{(r)*} E(Y) &= \\ G_1^{(r)} + \frac{1}{(p+1)!} \mu_{p+1}(K_{(r,p)}) g_1^{p-r+1} G_1^{(p+1)} &+ o_p(g_1^{p-r+1} + g_1^{-r} g_2^{p+1} + g_2^{\frac{r+s}{2}}). \end{aligned}$$

For the second term, we note that

$$[(I - S_1^* S_2^*)^{-1} - (I - S_2^* S_1^*)^{-1} S_2^*] W_1^{(r)*} = S_1^{(r)} + o_p(\mathbf{1}^T / n g_1^{r+1}).$$

and

$$[S_1^{(r)} S_1^{(s)T}]_{ii} = \mu(K_{(r,p)} K_{(s,p)}) f_X(X_i)^{-1} \frac{1}{n g_1^{r+s+1}} + o_p\left(\frac{1}{n g_1^{r+s+1}}\right).$$

Plugging those approximations in (15) leads to the desired bias expression.

For the variance of $\hat{\theta}_{11}(r, s)$, rewrite $\hat{\theta}_{11}(r, s)$ as

$$\hat{\theta}_{11}(r, s) = \frac{1}{n} \sum_{i=1}^n \sum_{j=1}^n L_{rs}^{ij} Y_i Y_j$$

where

$$\begin{aligned} L_{rs}^{ij} &= \sum_{k=1}^n [\mathbf{s}_{1,X_i}^{(r)}]_k [\mathbf{s}_{1,X_j}^{(s)}]_k \\ &= \frac{1}{n g_1^{r+s+2}} \int K_{(r,p)}\left(\frac{X_i - x}{g_1}\right) K_{(s,p)}\left(\frac{X_j - x}{g_1}\right) f_X(x)^{-1} dx (1 + o_p(1)). \end{aligned}$$

Hence, the variance of $\hat{\theta}_{11}(r, s)$ is

$$\begin{aligned} \text{Var}(\hat{\theta}_{11}(r, s)) &= \frac{1}{n^2} \sum_{i=1}^n \sum_{j=1}^n \sum_{k=1}^n \sum_{l=1}^n L_{rs}^{ij} L_{rs}^{kl} \text{Cov}(Y_i Y_j, Y_k Y_l) \\ &= \frac{1}{n^2} \sum_{i=1}^n (L_{rs}^{ii})^2 \text{Var}(Y_i^2) + \frac{1}{n^2} \sum_{i \neq j} \left((L_{rs}^{ij})^2 + (L_{rs}^{ji})^2 \right) \text{Var}(Y_i Y_j) \end{aligned}$$

$$\begin{aligned}
& + \frac{1}{n^2} \sum \sum_{i \neq j} \left(L_{rs}^{ii} L_{rs}^{ij} + L_{rs}^{ij} L_{rs}^{ji} \right) \text{Cov}(Y_i^2, Y_i Y_j) \\
& + \frac{1}{n^2} \sum \sum_{i \neq j \neq k} \left(L_{rs}^{ij} L_{rs}^{ik} + L_{rs}^{ji} L_{rs}^{ki} + 2 L_{rs}^{ij} L_{rs}^{ki} \right) \text{Cov}(Y_i Y_j, Y_i Y_k) \\
& = A + B_1 + B_2 + C_1 + C_2 + D_1 + D_2 + 2D_3.
\end{aligned}$$

Each of these terms can be approximated in the same manner. We demonstrate the computations for B_1 . Let $\approx$ denote approximation of order $(1 + o_p(1))$ in the following derivations.

$$\begin{aligned}
B_1 &= \frac{\sigma^2}{n^3 g_1^{2r+2s+4}} \sum_{i=1}^n \int \int \int \int f_X(x_1)^{-1} f_X(x_2)^{-1} K_{(r,p)}\left(\frac{X_i - x_1}{g_1}\right) \\
&\quad K_{(s,p)}\left(\frac{x_3 - x_1}{g_1}\right) K_{(r,p)}\left(\frac{X_i - x_2}{g_1}\right) K_{(s,p)}\left(\frac{x_3 - x_2}{g_1}\right) (\sigma^2 + m(x_3, z_3))^2 \\
&\quad + m(X_i, Z_i)^2 f(x_3, z_3) dx_1 dx_2 dx_3 dz_3 \\
&= \frac{\sigma^2}{n^3 g_1^{2r+2s+4}} \sum_{i=1}^n \int \int \left\{ \int f_X(x_1)^{-1} K_{(r,p)}\left(\frac{X_i - x_1}{g_1}\right) K_{(s,p)}\left(\frac{x_3 - x_1}{g_1}\right) dx_1 \right\}^2 \\
&\quad (\sigma^2 + m(x_3, z_3))^2 + m(X_i, Z_i)^2 f(x_3, z_3) dx_2 dx_3 dz_3 \\
&\approx \frac{\sigma^2}{n^3 g_1^{2r+2s+2}} \sum_{i=1}^n \int \int f_X(X_i)^{-2} \left\{ K_{(r,p)} * K_{(s,p)}\left(\frac{X_i - x_3}{g_1}\right) \right\}^2 \\
&\quad (\sigma^2 + m(x_3, z_3))^2 + m(X_i, X_i)^2 f(x_3, z_3) dx_3 dz_3 \\
&\approx \frac{\sigma^2}{n^3 g_1^{2r+2s+1}} \sum_{i=1}^n \int f_X(X_i)^{-2} R(K_{(r,p)} * K_{(s,p)}) (\sigma^2 + m(X_i, z_3))^2 \\
&\quad + m(X_i, Z_i)^2 f(X_i, z_3) dz_3 \\
&= \frac{\sigma^2}{n^2 g_1^{2r+2s+1}} R(K_{(r,p)} * K_{(s,p)}) \frac{1}{n} \sum_{i=1}^n f_X(X_i)^{-1} (\sigma^2 + m(X_i, Z_i)^2 \\
&\quad + E(m(X_i, Z_i)^2 | X_i)).
\end{aligned}$$

B_2 is approximated by the same quantity. Other significant terms are D_1, D_2 and D_3 . ■

Outline of proof of Theorem 2.2: The bias result follows immediately from Theorem 5.1 of Opsomer and Ruppert [19]. For the variance, we use the same reasoning as in Hall and Marron [12]. We write

$$\text{Var}(\hat{\sigma}^2) = \frac{1}{n^2} \left\{ \sum_{j=1}^n E(\Delta_j^2) + 2\sigma^4 \sum \sum_{j \neq k} t_{jk}^2 \right\},$$

where

$$\Delta_j = \left(\delta_j - \sum_{i=1}^n \delta_i w_{ij} \right) \varepsilon_j + \left(1 - 2w_{jj} + \sum_{i=1}^n w_{ij}^2 \right) (\varepsilon_j^2 - \sigma^2)$$

$$\delta_i = m(X_i, Z_i) - \sum_{j=1}^n w_{ij} m(X_j, Z_j) \quad t_{jk} = \sum_{i=1}^n w_{ij} w_{ik} - 2w_{jk}.$$

Let $\nu_j = 1 - 2w_{jj} + \sum_{i=1}^n w_{ij}^2$ and note that $\sum_{i=1}^n \nu_i^2 = n(1 + o_p(1))$. Hence,

$$\Delta_j = \nu_j(\varepsilon_j^2 - \sigma^2) + o_p(\lambda_1^p + \lambda_2^p)\varepsilon_j$$

and

$$\frac{1}{n^2} \sum_{j=1}^n \mathbb{E}(\Delta_j^2) = \frac{1}{n} \text{Var}(\varepsilon_j^2) + o_p\left(\frac{1}{n} + \frac{1}{n^2 \lambda_1} + \frac{1}{n^2 \lambda_2}\right).$$

For t_{jk} , we have

$$\begin{aligned} t_{jk} &\approx \frac{1}{n\kappa_1} (K_{(p)} * K_{(p)} - 2K_{(p)}) \left(\frac{X_j - X_k}{\kappa_1} \right) f_X(X_j)^{-1} \\ &\quad + \frac{1}{n\kappa_2} (K_{(p)} * K_{(p)} - 2K_{(p)}) \left(\frac{Z_j - Z_k}{\kappa_2} \right) f_Z(Z_j)^{-1}. \end{aligned}$$

and

$$\begin{aligned} \frac{1}{n^2} \sum_{j \neq k} t_{jk}^2 &= R(K_{(p)} * K_{(p)} - 2K_{(p)}) \sum_{j=1}^n \left\{ f_X(X_j)^{-1} \frac{1}{n^2 \kappa_1} + f_Z(Z_j)^{-1} \frac{1}{n^2 \kappa_2} \right\} \\ &\quad + o_p\left(\frac{1}{n^2 \kappa_1} + \frac{1}{n^2 \kappa_2}\right). \end{aligned}$$

Putting together these expressions yields the desired expression for the variance. ■

Outline of proof of Theorem 2.3: First, note that $h_{1,MASE} = h_{1,AMASE} + O_p(n^{-3/5})$ which can be shown by using the same reasoning as in Hall *et al.* [13] for the kernel density estimation problem. This approximation puts an upper bound on the relative rate of convergence of $\hat{h}_1$ to $h_{1,MASE}$:

$$\frac{\hat{h}_1 - h_{1,MASE}}{h_{1,MASE}} = \frac{\hat{h}_1 - h_{1,AMASE}}{h_{1,AMASE}} + O_p(n^{-2/5}).$$

Hence, by Corollaries 2.1 and 2.2,

$$\begin{aligned} \frac{\hat{h}_1 - h_{1,MASE}}{h_{1,MASE}} &\approx (\hat{\theta}_{11} - \theta_{11}) \frac{\partial h_{1,AMASE}}{\partial \theta_{11}} \frac{1}{h_{1,AMASE}} \\ &\quad + (\hat{\sigma}^2 - \sigma^2) \frac{\partial h_{1,AMASE}}{\partial \sigma^2} \frac{1}{h_{1,AMASE}} + O_p(n^{-2/5}) \\ &= O_p(n^{-2/7}). \end{aligned}$$
■

Proof of Theorem 2.4: Note that $\theta_{12}^2 \leq \theta_{11}\theta_{22}$ by the Cauchy-Schwartz inequality, so that necessarily $h_{i,AMASE} = H_i n^{-1/5}$, as in the independent case. The reasoning for finding the relative rates of convergence of $\hat{h}_1, \hat{h}_2$ is analogous to that for Theorem 2.3. ■

Outline of proof of Theorem 2.5: Note that $\mathbf{S}_d^* \mathbf{m}_k = O_p(1/\sqrt{nh_d})$ if $k \neq d$. Since $\mathbf{M}$ is assumed invertible,

$$\begin{bmatrix} E(\hat{\mathbf{m}}_1) \\ \vdots \\ E(\hat{\mathbf{m}}_D) \end{bmatrix} = \mathbf{M}^{-1} \begin{bmatrix} \mathbf{S}_1 \mathbf{m}_1 \\ \vdots \\ \mathbf{S}_D \mathbf{m}_D \end{bmatrix} + \mathbf{R}_{vec},$$

where $\mathbf{R}_{vec} = \{\mathbf{R}_1^T, \dots, \mathbf{R}_D^T\}^T$ and $\mathbf{R}_d = O_p(\mathbf{1}^T/\sqrt{nh_d}) + o_p(\sum_{k=1}^D \mathbf{h}_k^{p_k+1})$. The asymptotic approximation to $\mathbf{S}_d \mathbf{m}_d$ is found through a straightforward application of Theorem 4.1 of Ruppert and Wand [22], yielding the desired bias result.

To compute the variance of $\hat{m}_d(X_{di})$, we first note that

$$\mathbf{M} \begin{bmatrix} \mathbf{S}_1^* \\ \vdots \\ \mathbf{S}_D^* \end{bmatrix} = \begin{bmatrix} \mathbf{S}_1^* \\ \vdots \\ \mathbf{S}_D^* \end{bmatrix} + \mathbf{T}_{vec}$$

with $\mathbf{T}_{vec} = o_p(\{\mathbf{1}\mathbf{1}^T/n, \dots, \mathbf{1}\mathbf{1}^T/n\}^T)$. If we let $\mathbf{S}_{vec} = [\mathbf{S}_1^{*T}, \dots, \mathbf{S}_D^{*T}]$, we have

$$\mathbf{M}^{-1} \mathbf{S}_{vec} = \mathbf{S}_{vec} + \mathbf{T}_{vec}.$$

The variance of $\hat{m}_d(X_{di})$ is σ^2 multiplied by the $(nd + i)$ th element of the diagonal of

$$\mathbf{M}^{-1} \mathbf{S}_{vec} \mathbf{S}_{vec}^T \mathbf{M}^{-T} = \mathbf{S}_{vec} \mathbf{S}_{vec}^T + \mathbf{T}_{vec} \mathbf{T}_{vec}^T,$$

which can be approximated using the same approach as in Theorem 5.1 in Opsomer and Ruppert [19]. ■

References

- [1] D. A. Belsley, E. Kuh, and R. E. Welsch. *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity*. John Wiley & Sons, New York, 1980.
- [2] K.N. Berk and D.E. Booth. Seeing a curve in multiple regression. *Technometrics*, 37:385–398, 1995.

- [3] L. Breiman and J. H. Friedman. Estimating optimal transformations for multiple regression and correlation. *Journal of the American Statistical Association*, 80:580–619, 1985.
- [4] A. Buja, D. Donnell, and W. Stuetzle. Additive principal components. Technical report, Department of Statistics, University of Washington, Seattle, WA, 1986.
- [5] A. Buja, T. J. Hastie, and R. J. Tibshirani. Linear smoothers and additive models. *Annals of Statistics*, 17:453–555, 1989.
- [6] J. M. Chambers and T. J. Hastie, editors. *Statistical models in S*. Wadsworth and Brooks/Cole, Pacific Grove, CA, 1992.
- [7] R. D. Cook and S. Weisberg. *Residuals and Influence in Regression*. Chapman and Hall, Washington, D.C., 1982.
- [8] D. J. Donnell, A. Buja, and W. Stuetzle. Analysis of additive dependencies and concurvities using smallest additive principal components. *Annals of Statistics*, 22:1635–1668, 1994.
- [9] J. Fan and I. Gijbels. Data-driven bandwidth selection in local polynomial regression: variable bandwidth selection and spatial adaptation. *Journal of the Royal Statistical Association, Series B*, 57:371–394, 1995.
- [10] J. Fan and J. S. Marron. Fast implementation of nonparametric curve estimators. *Journal of Computational and Graphical Statistics*, 3:35–56, 1994.
- [11] C. Gu and G. Wahba. Minimizing gcv/gml scores with multiple smoothing parameters via the newton method. Technical Report 847, Department of Statistics, University of Wisconsin at Madison, 1988.
- [12] P. Hall and J. S. Marron. On variance estimation in nonparametric regression. *Biometrika*, 77:415–419, 1990.
- [13] P. Hall, S. J. Sheather, M. C. Jones, and J. S. Marron. On optimal data-based bandwidth selection in kernel density estimation. *Biometrika*, 78:263–269, 1991.
- [14] W. Härdle, P. Hall, and J. S. Marron. How far are automatically chosen regression smoothing parameters from their optimum? *Journal of the American Statistical Association*, 83:86–95, 1988.

- [15] D. Harrison and D. L. Rubinfeld. Hedonic housing prices and the demand for clean air. *Journal of Economics and Management*, 5:81–102, 1978.
- [16] T. J. Hastie and R. J. Tibshirani. *Generalized Additive Models*. Chapman and Hall, Washington, D.C., 1990.
- [17] The MathWorks, Inc., Natick, MA. *Matlab Reference Guide*, 1994.
- [18] J.-D. Opsomer. *Optimal Bandwidth Selection for Fitting an Additive Model by Local Polynomial Regression*. PhD thesis, Cornell University, 1995. Available via anonymous FTP at ftp.orie.cornell.edu.
- [19] J.-D. Opsomer and D. Ruppert. Fitting a bivariate additive model by local polynomial regression. Preprint 95–27, Department of Statistics, Iowa State University, 29 August 1995.
- [20] D. Ruppert. Empirical-bias bandwidths for local polynomial nonparametric regression and density estimation. Technical Report TR1137, School of Operations Research and Industrial Engineering, Cornell University, 1995. Available via anonymous FTP at ftp.orie.cornell.edu.
- [21] D. Ruppert, S. J. Sheather, and M. P. Wand. An effective bandwidth selector for local least squares regression. *Journal of the American Statistical Association*, 90:1257–1270, 1996.
- [22] D. Ruppert and M. P. Wand. Multivariate locally weighted least squares regression. *Annals of Statistics*, 22:1346–1370, 1994.
- [23] M. P. Wand and M. C. Jones. *Kernel Smoothing*. Chapman and Hall, London, 1995.

Department of Statistics
 Snedecor Hall
 Iowa State University
 Ames, IA, 50011-1210

College Of Engineering
 Rhodes Hall
 Cornell University
 Ithaca, NY 14853-3801