A root-n consistent backfitting estimator for semiparametric additive modelling

Jean D. Opsomer
Department of Statistics
Iowa State University
Ames, IA 50011

David Ruppert
School of Operations Research & Industrial Engineering
Cornell University
Ithaca, NY 14853

March 10, 1999

Abstract

We explore additive models that combine both parametric and nonparametric terms and propose a $\sqrt{n}$-consistent backfitting estimator for the parametric component of the model. The theoretical properties of the estimator are developed for the case with a single nonparametric term and extended to an arbitrary number of nonparametric additive terms. An estimator for the optimal bandwidth making minimal use of asymptotic expressions for bias and variance is proposed, and a fast implementation algorithm for model fitting and bandwidth selection is developed. The practical behavior of the estimator and bandwidth selection is illustrated by simulation experiments.

Key Words: local polynomial regression, bandwidth selection, EBBS, partially linear model.

1 Introduction

Additive models are a popular and flexible class of nonparametric regression methods (Hastie and Tibshirani (1990)), which assume that the conditional mean function can be represented as

$$E(Y|Z_1, \ldots, Z_D) = m_1(Z_1) + \ldots + m_D(Z_D).$$

Because they allow multidimensional smoothing to reduce to a sequence of one-dimensional smoothing steps (Stone (1985), Opsomer (1996)), they allow analysis of multidimensional problems which would be arduous or even impossible to approach with "full-dimensional" nonparametric methods. They also maintain the ease of interpretation of univariate nonparametric smooths, since the estimates of the component functions can be plotted separately. The additive model is particularly useful in cases where the model underlying the data can be assumed to be (approximately) additive in the covariates.

An interesting special case of the additive model is the *semiparametric additive model* (SAM in what follows), where some of the additive component functions are parametric terms while the remaining ones are unspecified and are estimated nonparametrically. Such a model can be written as

$$\mathrm{E}(Y|X_1,\ldots,X_P,Z_1,\ldots,Z_D)=\sum_{p=1}^{P}\beta_p X_p+\sum_{d=1}^{D}m_d(Z_d). \tag{1}$$

Such semiparametric additive model might occur, for instance, when the main interest of a study is to precisely quantify the effect of a number of covariates $X_1,\ldots,X_P$ on the dependent variable Y, but the relationship is observed in the presence of "nuisance" covariates $Z_1,\ldots,Z_D$. The use of parametric terms for the X_p, if properly specified, allows for easily interpretable inference about its effect on Y, while by modelling the $m_d(Z_d)$ nonparametrically, one avoids the potential introduction of bias in the estimated relationship between the X_p and Y. This problem can be particularly severe when the X_p and Z_d are correlated. Another case would be when the researcher is fairly confident about the shape of the relationship between the X_p and Y, but not about that of the other covariates. By modeling some terms by the appropriate parametric model, the risk of overfitting the model is reduced by decreasing the overall "degrees of freedom" of the fit (Hastie and Tibshirani (1990)). A common example of the latter case is when the X_p are dummy (0/1) variables, so that nonparametric regression would be meaningless.

Since SAM is a special case of the additive model, it is highly desirable to be able to compute estimates of the former with the same method as the latter, i.e., with the *backfitting algorithm* of Buja *et al.* (1989). If $D=1$, backfitting is unnecessary, since an equivalent, non-iterative solution exists for the estimators of a SAM (see Section 2.1 below). This is no longer the case for higher-dimensional models.

For $D = 1$, the SAM, often called the *partially linear model* in this case, has been previously studied by several authors. Early results by Heckman (1986) for the balanced analysis of covariance case indicated that a penalized likelihood approach (equivalent to backfitting) indeed resulted in a $\sqrt{n}$-consistent estimator for the vector $(\beta_1, \ldots, \beta_P)$. Rice (1986) showed that this result does not hold in general. In the case of kernel regression, Speckman (1988) also showed that when the "typical" rates of convergence are used for the nonparametric terms, $\sqrt{n}$-consistency cannot be achieved. The latter author therefore suggests an alternative estimator which achieves this rate, but is no longer equivalent to a backfitting estimator. The asymptotic properties of that alternative estimator are further studied by Gao (1995) and Linton (1995). The latter author also proposes a plug-in bandwidth selection method. A drawback of that estimator is that it does not easily generalize to the case $D > 1$. Recently, Carroll *et al.* (1997) studied *generalized partially linear single-index models*, a class of models that include the partially linear model as a special case. The estimator the propose is also different from the backfitting estimator and does not generalize easily to the case $D > 1$.

In this article, we show that the reason for the lack of $\sqrt{n}$-consistency of backfitting estimators is due to the wrong choice of bandwidth, and that, by selecting a more appropriate bandwidth, the optimal convergence rate can be achieved without adjustments to the estimator itself. This same result was found by Carroll *et al.* (1997) for their estimator. We also show that traditional bandwidth selection methods such as cross-validation and plug-in are unable to estimate this optimal bandwidth, though for unrelated reasons. A different approach is therefore necessary to select the bandwidth. We propose a bandwidth selection method based on the *Empirical Bias Bandwidth Selection* (EBBS) method of Ruppert (1997), which is used in conjunction with backfitting to estimate the parameters in SAM.

Asymptotics suggest that the bandwidth that is optimal for estimating the β_i's will be smaller than the bandwidth optimal for estimating m, and this is what we found in our finite-sample study. In the simple case where $P = D = 1$, asymptotics show that the bias of $\hat{\beta}$ caused by bias in $\hat{m}$ depends on the covariance between X and $m''(Z)$, and this effect was also found in the simulations.

In Section 2, the theoretical properties of the SAM estimator are studied, both when the nonparametric model term is univariate (Section 2.1) and multivariate (Section 2.2). Section 3 describes the proposed bandwidth selection method, and Section

4 reports on a simulation experiment. All proofs are in the Appendix.

2 Asymptotic Properties

2.1 A Simple Case

We first study the case where the Z_i are univariate, so that the model under consideration is

$$Y_i = \boldsymbol{X}_i^T \boldsymbol{\beta} + m(Z_i) + \varepsilon_i,$$

with $\boldsymbol{X}_i = (X_{i1}, \ldots, X_{iP})^T$, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_P)^T$ and where the errors ε_i are iid with mean 0 and variance σ^2. This is the partially linear model studied by Speckman (1988). The backfitting estimators for $\boldsymbol{\beta}$ and $\boldsymbol{m} = (m(Z_1), \ldots, m(Z_n))^T$ are equivalent to

$$\begin{aligned}
\hat{\boldsymbol{\beta}} &= (\boldsymbol{X}^T(\boldsymbol{I} - \boldsymbol{S})\boldsymbol{X})^{-1}\boldsymbol{X}^T(\boldsymbol{I} - \boldsymbol{S})\boldsymbol{Y} \\
\hat{\boldsymbol{m}} &= \boldsymbol{S}(\boldsymbol{Y} - \boldsymbol{X}^T\hat{\boldsymbol{\beta}}),
\end{aligned}$$

with $\boldsymbol{X} = [\boldsymbol{X}_1 \ldots \boldsymbol{X}_n]^T$, $\boldsymbol{Y} = (Y_1, \ldots, Y_n)^T$ and similarly for $\boldsymbol{Z}$, and $\boldsymbol{S}$ the *smoother matrix* for estimating $\mathrm{E}(\cdot|\boldsymbol{Z})$ by backfitting. We will only consider local linear regression as the smoothing method in what follows. Generalization to local polynomials of higher degree is straightforward (see Opsomer and Ruppert (1997)).

The rows of the smoother matrix correspond to the smoother vectors $\boldsymbol{s}_z^T$ evaluated at the observation points, $z = Z_1, \ldots, Z_n$. For local linear regression, we write $\boldsymbol{S} = [\boldsymbol{s}_{Z_1} \cdots \boldsymbol{s}_{Z_n}]^T$, where

$$\boldsymbol{s}_z^T = \boldsymbol{e}_1^T(\boldsymbol{Z}_z^T\boldsymbol{W}_z\boldsymbol{Z}_z)^{-1}\boldsymbol{Z}_z^T\boldsymbol{W}_z$$

with $\boldsymbol{e}_1^T = (1, 0)$, $\boldsymbol{W}_z = \mathrm{diag}\{\frac{1}{h}K(\frac{Z_1 - z}{h}), \ldots, \frac{1}{h}K(\frac{Z_n - z}{h})\}$ for some kernel function K and

$$\boldsymbol{Z}_z = \begin{bmatrix} 1 & (Z_1 - z) \\ \vdots & \vdots \\ 1 & (Z_n - z) \end{bmatrix}.$$

One often overlooked fact about this type of models is that in general, the estimators are *not* well-defined when one of the parametric terms is taken to be an intercept (X_{ip} is a constant for all i, for some variable p). In that case, $(\boldsymbol{X}^T(\boldsymbol{I} - \boldsymbol{S})\boldsymbol{X})$ is singular, since it contains a column of zeroes. This is true for any smoothing method which preserves constants, including local linear regression. Simple solutions to this problem are to *center* the smoothers, i.e., replace $\boldsymbol{S}$ by $\boldsymbol{S}^* = (\boldsymbol{I} - \mathbf{1}\mathbf{1}/n)\boldsymbol{S}$, and/or center all variables around their sample means. The former is a common adjustment for additive models, and it is in general necessary to ensure uniqueness of the solutions of the backfitting algorithm (Opsomer and Ruppert (1997)). For simplicity, we will assume here that the covariates in the model have been centered.

The moments of K are written as $\mu_r(K) = \int u^r K(u)du$. We make the following assumptions:

(AS.I) The kernel K is bounded and continuous and has compact 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 density $f_Z(z)$ is bounded and continuous and has compact support. Also, $f_Z(z) > 0$ for all $z \in \text{supp}(f_Z)$.

(AS.III) As $n \to \infty$, $h \to 0$ and $nh \to \infty$.

(AS.IV) The second derivative of m exists and is bounded and continuous.

These assumptions are standard for local linear regression (see Ruppert and Wand (1994)). Under these assumptions, we prove in the Appendix:

Theorem 2.1 *The conditional bias and variance of $\hat{\boldsymbol{\beta}}$ can be approximated by:*

$$\mathrm{E}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}|\boldsymbol{X}, \boldsymbol{Z}) = -\frac{\mu_2(K)}{2}h^2 \mathrm{E}(\mathrm{var}(\boldsymbol{X}_i|Z_i))^{-1}\mathrm{cov}(\boldsymbol{X}_i, m''(Z_i)) + o_p(h^2)$$

and

$$\mathrm{var}(\hat{\boldsymbol{\beta}}|\boldsymbol{X}, \boldsymbol{Z}) = \frac{\sigma^2}{n}\mathrm{E}(\mathrm{var}(\boldsymbol{X}_i|Z_i))^{-1} + O_p\left(\frac{h^2}{n} + \frac{1}{n^2h}\right)$$

Theorem 2.1 shows that for the "usual" $n^{-1/5}$ rate for the bandwidth, the estimate of $\boldsymbol{\beta}$ is not $\sqrt{n}$-consistent, since the bias is $O_p(n^{-2/5})$. This is the same result found by Speckman (1988). It can be remedied by selecting a different bandwidth rate, as the following corollary makes more precise:

Corollary 2.1 *If $h \propto n^r$ for $-1 < r < -1/4$, then*

$$\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \Rightarrow \mathcal{N}\left(0, \sigma^2\mathrm{E}(\mathrm{var}(\boldsymbol{X}_i|Z_i))^{-1}\right).$$

Corollary 2.1 shows that for the right choice of bandwidths, the backfitting estimator of $\boldsymbol{\beta}$ achieves the semiparametric efficiency bound and is in that sense equivalent to the estimator proposed by Speckman (1988).

Compared to the equivalent expressions from OLS regression, we can see that the effect of the nonparametrically modelled covariate Z_i on $\mathrm{var}(\hat{\boldsymbol{\beta}})$ is to replace $\mathrm{var}(\boldsymbol{X}_i)$ by $\mathrm{E}(\mathrm{var}(\boldsymbol{X}_i|Z_i))$ in the denominator. When $\boldsymbol{X}_i$ and Z_i are independent, the asymptotic variance of $\hat{\boldsymbol{\beta}}$ is therefore the same as when the model is fully parametric. Another way to interpret this result is that $\mathrm{var}(\hat{\boldsymbol{\beta}})$ for the SAM estimator is asymptotically the same as that found when one uses OLS to regress Y_i on $(\boldsymbol{X}_i - \mathrm{E}(\boldsymbol{X}_i|Z_i))$ without Z_i in the model, which makes sense since Z_i and $(\boldsymbol{X}_i - \mathrm{E}(\boldsymbol{X}_i|Z_i))$ are uncorrelated.

Theorem 2.1 shows that the effect of the correlation between $\boldsymbol{X}_i$ and Z_i on the asymptotic bias is through the terms $\mathrm{cov}(\boldsymbol{X}_i, m''(Z_i))$ and $\mathrm{E}(\mathrm{var}(\boldsymbol{X}_i|Z_i))$. The former term also depends on the curvature of m through its second derivative. This bias disappears when $\boldsymbol{X}_i$ and Z_i are independent.

These asymptotic results are not directly useful to develop a "traditional" plug-in bandwidth selection method as in Gasser *et al.* (1991), Ruppert *et al.* (1995) or Opsomer and Ruppert (1998), because of the lack of explicit expressions for the bandwidth-dependent terms in the Mean Squared Error.

In the case of cross-validation, the difficulties involved in selecting a bandwidth appropriate for $\hat{\boldsymbol{\beta}}$ are even more fundamental. The cross-validation function is defined as

$$\mathrm{CV}(h) = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{\boldsymbol{\beta}}^{(-i)}\boldsymbol{X}_i - \hat{m}^{(-i)}(Z_i))^2,$$

where the superscripts $(-i)$ denote the ith leave-one-out estimators. In this quantity, the slowest converging estimator is $\hat{m}^{(-i)}$, so that as $n \to \infty$, its behavior will dominate that of $\hat{\boldsymbol{\beta}}^{(-i)}$. While this method might therefore be appropriate for estimating m, it does not provide satisfactory bandwidth choice for β.

2.2 Extension to General Backfitting Estimators

Consider model (1) with $P, D > 1$, to be estimated by backfitting with local linear regression. Since there are now several nonparametric terms, the model is not identifiable unless restrictions are placed on the functions m_d. A common assumption is that $\mathrm{E}(m_d(Z_{di})) = 0$ for $d = 1, \ldots, D$, which is achieved by using the centering adjustment for the smoother matrices (described in Section 2.1) and by explicitly including an intercept term in the model:

$$\mathrm{E}(Y|X_1,\ldots,X_P,Z_1,\ldots,Z_D) = \alpha + \sum_{p=1}^{P}\beta_p X_p + \sum_{d=1}^{D} m_d(Z_d). \tag{2}$$

While it is not strictly required, we will again assume that the expectation of the parametric terms is zero, i.e. that $\mathrm{E}(X_{pi}) = 0$ for $p = 1,\ldots,P$, so that $\mathrm{E}(Y_i) = \alpha$. In practice, this implies replacing the covariates X_{pi} by $X_{pi} - \bar{X}_p$ and estimating α by $\bar{Y}$. To simplify the notation, we assume here that both the X_{pi} and Y_i are centered around 0 and ignore α in the discussion that follows.

Let $\boldsymbol{S}_d^*, d = 1,\ldots,D$ represent centered local linear smoother matrices with corresponding bandwidths h_d. The estimators for the SAM are defined as the solutions of the backfitting algorithm on the following set of equations

$$\begin{cases} \hat{\boldsymbol{\beta}} & = & (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T(\boldsymbol{Y} - \sum_{k=1}^{D}\hat{\boldsymbol{m}}_k) \\ \hat{\boldsymbol{m}}_d & = & \boldsymbol{S}_d^*(\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}} - \sum_{k\neq d}\hat{\boldsymbol{m}}_k), \quad d = 1,\ldots,D, \end{cases}$$

where $\hat{\boldsymbol{m}}_d = (\hat{m}_d(Z_{d1}),\ldots,\hat{m}_d(Z_{d1}))^T$. Alternatively, we can define the SAM estimators non-iteratively as follows. Let $\boldsymbol{W}_d$ represent the *additive model smoother* for additive component function m_d, i.e. the matrix which maps $\boldsymbol{Y}$ to $\hat{\boldsymbol{m}}_d$. Explicit expressions when $D = 2$ are given in Hastie and Tibshirani (1990), and recursive expressions for higher-dimensional models can be found in Opsomer (1996). Let $\boldsymbol{W}_M = \sum_{d=1}^{D}\boldsymbol{W}_d$ represent the *additive model smoother* corresponding to $m_M = m_1 + \ldots + m_D$. Then,

$$\begin{aligned} \hat{\boldsymbol{\beta}} &= (\boldsymbol{X}^T(\boldsymbol{I} - \boldsymbol{W}_M)\boldsymbol{X})^{-1}\boldsymbol{X}^T(\boldsymbol{I} - \boldsymbol{W}_M)\boldsymbol{Y}, \\ \hat{\boldsymbol{m}}_M &= \boldsymbol{W}_M(\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}). \end{aligned}$$

These expressions are equivalent to the solutions of the backfitting algorithm, when that solution is unique. Direct computation of the estimators is rarely performed with these equations, since backfitting is much more efficient. They allow the theory developed in Opsomer (1996) to be used to find explicit expressions for the asymptotic bias and variance of $\hat{\boldsymbol{\beta}}$. Since these expressions would be recursive as well as very complicated, and could not be used to develop a bandwidth selection method, we only derive the asymptotic convergence rates, showing that the leading terms in this setting generalize those of Section 2.1. In addition to assumption (AS.I), we need:

(AS.II′) The density $f_Z(\boldsymbol{z})$ is bounded and continuous and has compact support. Also, $f_Z(\boldsymbol{z}) > 0$ for all $\boldsymbol{z} \in \mathrm{supp}(f_Z)$.

(AS.III′) As $n \to \infty$, $h_d \to 0$ and $nh_d \to \infty$ for $d = 1, \ldots, D$.

(AS.IV′) The second derivatives of $m_d, d = 1, \ldots, D$ exist and are bounded and continuous.

Theorem 2.2 *The leading terms of the conditional bias and variance are of the following orders:*

$$\mathrm{E}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta} | \boldsymbol{X}, \boldsymbol{Z}) = O_p(\sum_{d=1}^{D} h_d^2) \text{ and } \mathrm{var}(\hat{\boldsymbol{\beta}} | \boldsymbol{X}, \boldsymbol{Z}) = O_p(\frac{1}{n}),$$

so that if $h_d \propto n^r$ for some $-1 < r < -1/4$, $d = 1, \ldots, D$, then $\hat{\boldsymbol{\beta}}$ is $\sqrt{n}$-consistent.

The proof of Theorem 2.2 is in the Appendix. Theorem 2.2 shows that the SAM estimator achieves the semiparametric efficiency bound for any D. Using the approach in Opsomer (1996) and this theorem, it is possible to derive the conditional bias and variance for the nonparametric terms m_d, $d = 1, \ldots, D$ and show that their convergence rates are the same as in the fully nonparametric additive model.

3 Implementing a Bandwidth Selection Method

We develop a bandwidth selection method for estimating the parameters $\beta_1, \ldots, \beta_P$ in the general model (2). The optimal choice for the bandwidth parameters is taken to be the minimizer of $\mathrm{MSE}(\boldsymbol{c}^T \hat{\boldsymbol{\beta}} | \boldsymbol{X}, \boldsymbol{Z})$, for a given vector of coefficients $\boldsymbol{c}$. The specification of $\boldsymbol{c}$ is needed to reduce the criterion for optimality of the bandwidth to a one-dimensional quantity. Possible choices for $\boldsymbol{c}$ are $\boldsymbol{c} = \boldsymbol{e}_p = (0, \ldots, 0, 1, 0, \ldots, 0)^T$, if one primarily interested in estimating β_p accurately, or $\boldsymbol{c} = \bar{\boldsymbol{X}} = (\bar{X}_1, \ldots, \bar{X}_P)^T$, if prediction for the overall parametric portion of the model is considered most important. Similarly, to predict $\boldsymbol{Y}_{n+1}$, the next Y, one would use $\boldsymbol{c} = \boldsymbol{X}_{n+1}$, the next $\boldsymbol{X}$.

The bandwidth selection method is based on the *Empirical Bias Bandwidth Selection* (EBBS) method of Ruppert (1997). There are two important advantages to this approach here:

1. No asymptotic approximation to the variance of $\boldsymbol{c}^T \hat{\boldsymbol{\beta}}$ is required. Instead, the exact variance expression is used with only σ^2, the variance of ϵ, to be estimated from the data. This is especially important for estimation of β, since unlike the

case where m is estimated, the leading term for the variance in Theorem 2.1 does not depend on h. The dependence of the bandwidth on h occurs only in the remainder terms. (Actually, there is one asymptotic approximation used to speed computation of the exact variance. At one place in the algorithm, we assume that $\boldsymbol{W}_M$ is symmetric, though, in fact, this smoother matrix is only approximately symmetric.)

2. The bias-bandwidth function is estimated directly from the data using as a model the polynomial form predicted by the asymptotic theory in Section 2. This is similar in principle to the approach of Fan and Gijbels (1995).

Although EBBS requires several recalculations of the SAM estimates and potentially involves a D-dimensional numerical minimization, the overall computational burden can be significantly decreased by the use of two simplifications.

First, the backfitting algorithm is sped up by computing the local linear fits on a grid and them interpolating them to the Z-values in the data. See Opsomer and Ruppert (1998) for discussion. Second, to avoid a D-dimensional minimization, we let the bandwidth h_d for estimating m_d in the backfitting algorithm be of the form

$$h_d = b s_d$$

where s_d is the sample standard deviation of Z_d and b is a univariate smoothing parameter that we call the "bandwidth factor." We choose b by minimizing an estimate of the mean square error (MSE) of $\boldsymbol{c}^T\hat{\boldsymbol{\beta}}$. Since the main interest in the procedure is the minimization of the MSE for $\hat{\boldsymbol{\beta}}$, not that for the $\hat{m}_d$, this restriction on the bandwidths is not likely to result in a significant increase in MSE while it dramatically reduces the computational burden, especially for models with high-dimensional nonparametric components.

Here is how EBBS is implemented. Let $B = \{b(1), \ldots, b(N)\}$ be a grid of possible values of b. Typically N is between 10 and 20 and the values in B are equally spaced on the log scale, i.e., "log-spaced." For each $b(j) \in B$, we compute $\hat{\boldsymbol{\beta}}$ and $\hat{m}_M$, which we call $\hat{\boldsymbol{\beta}}(j)$ and $\hat{m}_M(\cdot; j)$. We then estimate σ^2 by

$$\hat{\sigma}^2(j) = n^{-1} \sum_{i=1}^{n} \Big(Y_i - \boldsymbol{X}_i^T \hat{\boldsymbol{\beta}}(j) - \hat{m}_M(\boldsymbol{Z}_i; j)\Big)^2. \qquad (3)$$

Next, $\mathrm{Var}(\hat{\boldsymbol{\beta}}(j)) = \sigma^2 \boldsymbol{\Sigma}(j)$ with

$$\boldsymbol{\Sigma}(j) = \{\boldsymbol{X}^T(\boldsymbol{I} - \boldsymbol{W}_M)\boldsymbol{X}\}^{-1}\{\boldsymbol{X}^T(\boldsymbol{I} - \boldsymbol{W}_M)(\boldsymbol{I} - \boldsymbol{W}_M)^T\boldsymbol{X}\}\{\boldsymbol{X}^T(\boldsymbol{I} - \boldsymbol{W}_M)\boldsymbol{X}\}^{-1},$$

where the dependence of $\boldsymbol{\Sigma}(j)$ on $b(j)$ is through $\boldsymbol{W}_M$. The smoother matrix $\boldsymbol{W}_M$ is not actually computed. Rather, the product $\boldsymbol{W}_M\boldsymbol{X}$ is found by smoothing each column of $\boldsymbol{X}$ against $Z_1, \ldots, Z_D$ by backfitting. Also, we use the approximation $(\boldsymbol{W}_M)^T\boldsymbol{X} \approx \boldsymbol{W}_M\boldsymbol{X}$ which is justified by the near symmetry of $\boldsymbol{W}_d$ for each d.

To estimate the bias of $\boldsymbol{c}^T\hat{\boldsymbol{\beta}}$ at $b(j)$, we make use of the approximation

$$E(\boldsymbol{c}^T\hat{\boldsymbol{\beta}}) = \boldsymbol{c}^T\boldsymbol{\beta} + \sum_{t=1}^{T} a_t b(j)^{1+t} + o(b^{1+T}(j)) \text{ as } b(j) \to 0, \tag{4}$$

where the a_t are constants. Although an asymptotic expression could be found for each a_t and estimates of the quantities in this expression could be estimated, we prefer not to use such a "plug-in" approach. Rather, at each $b(j)$, a_t, $t = 1, \ldots, T$ are directly estimated by fitting the "data" $\{(b(k), \boldsymbol{c}^T\hat{\boldsymbol{\beta}}(k)) : k = j - k_1, \ldots, j + k_2\}$. Here k_1, k_2, and T are tuning parameters that must be chosen by the user such that $k_1 + k_2 \geq T$ since we will estimate $T + 1$ parameters using $k_1 + k_2$ "data" points. A rather large simulation experiment reported in the next section shows that the performance of EBBS is insensitive to the choice of tuning parameters, at least within the range of values in the experimental region. We recommend $N = 18$, $k_1 = 1$, $k_2 = 2$, and $T = 1$ as reasonable values within that region.

The estimates of $a_1, \ldots, a_T$ at $b(j)$, $\hat{a}_t(j), t = 1, \ldots, T$, are found by minimizing

$$\sum_{k=j-k_1}^{j+k_2} \{\mathbf{c}^T\hat{\boldsymbol{\beta}} - \sum_{t=0}^{T} a_t(j) b(j)^{1+t}\}^2.$$

over $a_t(j), j = 0, \ldots, T$. The extra parameter, $a_0(j)$, represents $\mathbf{c}^T\boldsymbol{\beta}$ (see (4)) but $\hat{a}_0(j)$ is not actually used as an estimate of $\mathbf{c}^T\boldsymbol{\beta}$.

Then the bias of $\boldsymbol{c}^T\hat{\boldsymbol{\beta}}$ when $b(j)$ is the bandwidth is estimated by $\sum_{t=1}^{T} \hat{a}_t(j) b(j)^{1+t}$. Thus

$$\widehat{MSE}(b(j)) = \Big(\sum_{t=1}^{T} (\hat{a}_t(j) b^{1+t}(j)\Big)^2 + \hat{\sigma}^2(j)\boldsymbol{\Sigma}(j),$$

for $j = k_1, \ldots, M - k_2$. The final bandwidth factor $\hat{b}$ is the value of $b(j)$ that minimizes $\widehat{MSE}(b(j))$ over $\{b(k_1+1), \ldots, b(N-k_2)\}$. In Ruppert (1997), the first local minimizer of the estimated MSE is used, rather than the global minimizer. A disadvantage of using the global minimizer is that very large bandwidths might be selected since the bias model (4) breaks down at large bandwidths. However, this problem did not occur in our experiment and we found the global minimizer superior to the local minimizer.

As an alternative to estimating σ^2 with the current trial bandwidth as in (3), one could use residuals using a bandwidth optimized for estimation of σ^2 as in Opsomer and Ruppert (1998). One could also adjust the divisor n in (3) to correct for the bias due to estimation of β and m. However, these refinements would increase the computational cost and do not seem necessary if, as here, σ^2 is merely a nuisance parameter.

4 Simulations

4.1 A study of the EBBS tuning parameters

EBBS has tuning parameters N, T, k_1, and k_2. To find good values for these parameters, we performed a Monte Carlo experiment. In the experiment we used N, T, k_1, and V as factors. Here we define $V = k_1 + k_2 - T$ as the excess number of "data" points over the number of parameters. Data were generated according to the model

$$Y_i = \sum_{p=1}^{P} \beta_p X_{ip} + 2\sin(fZ_i) + \epsilon_i, \;\; i = 1, \ldots, n \tag{5}$$

where the ϵ_i where $N(0, \sigma^2)$. The Z_i were equally spaced on [0, 1] and $X_{ip} = \rho Z_i + (1-\rho)U_{ip}$ where the U_{ij} where iid Unif$(0,1)$. We consider only estimation of the first component of β so that $\mathbf{c}= (1, 0, \ldots)$. Although the Z_i are equally-spaced rather than random, the theory in Section 2 is the same as if they were iid Unif$(0,1)$. We chose them equally-spaced in this simulation study so that the minimum possible bandwidth did not depend on the sample

To study a wide variety of sampling situations, f, n, σ, P, and ρ were varied as factors, which we will call the "data factors" to distinguish them from N, T, k_1, and V which we will calling "EBBS factors." The levels of all factors are shown in Table 1. Each data factor was put at two widely spaced levels to represent rather different conditions. For the factor f we chose values 3 and 6, because with $f = 6$ there is a large covariance between X and $m''(Z)$ while for $f = 3$ there is little covariance. Thus, we expect $f = 3$ and 6 to be low bias and high bias cases, respectively. The EBBS factors V, N, and T were each set at three levels with the aim that one of the three levels would be nearly optimal. The factor k_1 was only set at two levels, 0 and 1, since k_1 was expected to be of lesser importance. A complete unreplicated design of 1728 runs was used, with the higher order interactions being used to estimate error.

Factor	Levels
n	50, 250
ρ	0.2, .6
σ	0.5, 2
f	3, 6
P	1, 4
k_1	0, 1
V	0, 1, 2
N	12, 18, 24
T	1, 2, 3

Table 1: Factor levels for the study of the EBBS tuning parameters.

The response was $\log(|\hat{\beta}_1 - \beta_1| + .001)$. The log transformation was used with the hope that it would reduce interactions, since the asymptotic expected absolute error is multiplicative in n and σ.

EBBS used bandwidth factors between a lower limit of .05 ($n = 250$) or .1 ($n = 50$) and an upper limit of 2. We looked at two variations on EBBS, one that used the bandwidth that globally minimized the estimated MSE and one that selected the first local minimizer. The globally minimizer proved more accurate, so that global minimization is recommended in practice and we will only report results for this method.

Main effects and interactions among the data factors were mostly as expected, e.g., n, σ, and P had large effects and smaller but significant interactions. However, the effects of ρ and f were not significant, which was somewhat surprising. Although these factors affect bias, apparently they do not affect the accuracy of $\hat{\beta}$ when using EBBS. EBBS merely selects small bandwidths in high bias cases; see the example below.

We found no main effects or interactions among the EBBS factors and no interactions between the EBBS factors and the data factors. *Since we varied five data factors over wide ranges, this is strong evidence that the choice of tuning parameters is not important.* We feel that any combination of the tuning parameter values in Table 1 can be recommended.

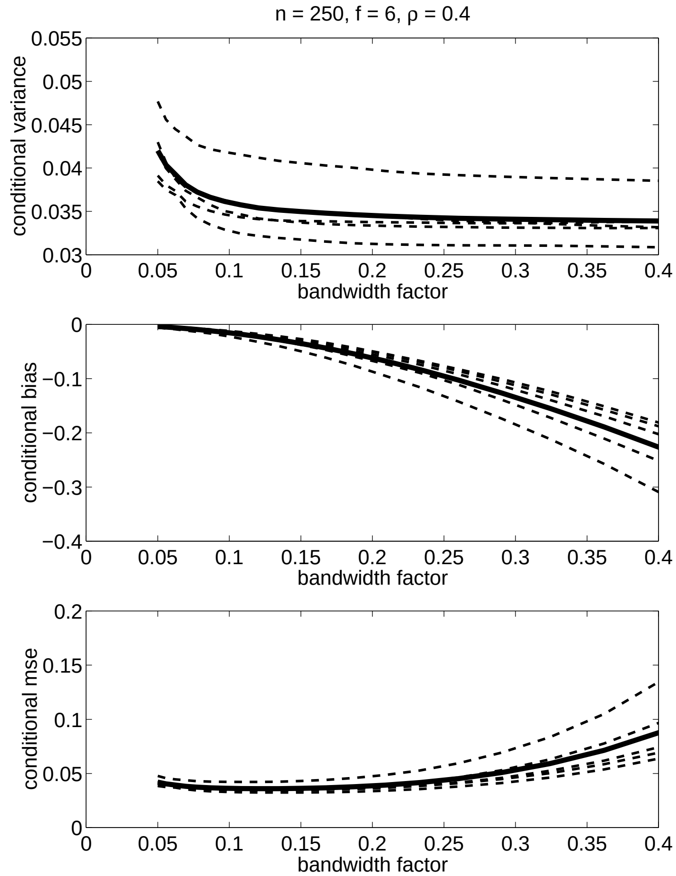

Figure 1: *Conditional bias, variance, and MSE for five independent samples (dashed) and their averages (solid).*

4.2 An example

We looked in detail at a special case of the above study with $n = 250$, $P = 1$, $\rho = .4$, and $\sigma = .5$.

First we studied the variability in the conditional bias and variance. Figure 1 shows the exact conditional bias, variance, and MSE for each of five independent samples (dashed) and their averages over the five samples (solid) for the case $f = 6$. One can see that there is considerable variance in the conditional variance and, for larger bandwidths, in the conditional bias. However, for each sample there is a range of the bandwidth factor from 0.1 to 0.2 where the exact conditional MSE is nearly minimized. EBBS does aim to find the bandwidth that minimizes conditional EBBS, which seems proper, but EBBS should also be finding the bandwidth that minimizes unconditional MSE.

We examined the cases $f = 3$ and 6 to study the effects of bias on EBBS. The tuning parameters were $N = 18$, $k_1 = 1$, $k_2 = 2$, and $T = 1$. EBBS used 18 bandwidth factors log-spaced between .05 and 2. Since $k_1 = 1$ and $k_2 = 2$, the smallest and two largest bandwidth factor could not be selected by EBBS. Thus, the bandwidth factors that could be selected ranged from $b(k_1+1) = 0.062$ to $b(N-k_2) = 1.29$. In Figure 2 we show the squared bias, variance, and MSE for estimation of β as functions of a fixed bandwidth as estimated from 500 samples. Panels (a) and (b) are the low and high bias cases, respectively. In panel (a) we see that bias is negligible. As would be expected, the variance increases as the bandwidth factor converges to 0. The variance also increases slightly as the bandwidth factor increases beyond 1; since the remainder term in the variance expression in Theorem 2.1 is not monotonic in h so this variance behavior is not so surprising. The variance of $\hat{\beta}$ does not depend on m so that variance function in (b) is the same as in (a). The bias in (b) increases rapidly as the bandwidth factor increases beyond 0.2.

We applied EBBS to 500 samples each from $f = 3$ and 6. In Figure 3, histograms of the EBBS-chosen bandwidth factors are given. Note that in the high bias case, panel (b), EBBS avoids the high bias region while in panel (a) EBBS tends to select larger bandwidths but largely avoid the region where the variance begins to increase.

We looked at $\hat{m}$ in the case $f = 6$ for various values of the bandwidth factor b. For $b = .1$, a value typical of what EBBS chooses and where the MSE of $\hat{\beta}$ is essentially minimized (see Figure 2), $\hat{m}$ is noticeably undersmoothed. For $b = .2$, the largest bandwidth factor that EBBS has much probability of choosing and the point where the MSE starts to increase, $\hat{m}$ appears to have the right amount of smoothness. The bias in $\hat{m}$ does not become noticeable until b is in the range of .3 to .4. These results agree with our asymptotics that show that the bandwidth optimal for $\hat{\beta}$ converges to zero more quickly that the bandwidth optimal for $\hat{m}$.

4.3 Computation time and performance as $n \to \infty$

To check the execution time we experimented with $P = D = 1$ and with 100 values of n log-spaced between 50 and 40,000. Clock time for execution of a MATLAB routine on a SPARC Ultra 1 was very nearly linear in n: time (in seconds) $\approx .83 + .0142\,n$. Thus, when $n = 500$ execution time is under 8 seconds while execution time is 2.25 seconds for $n = 100$.

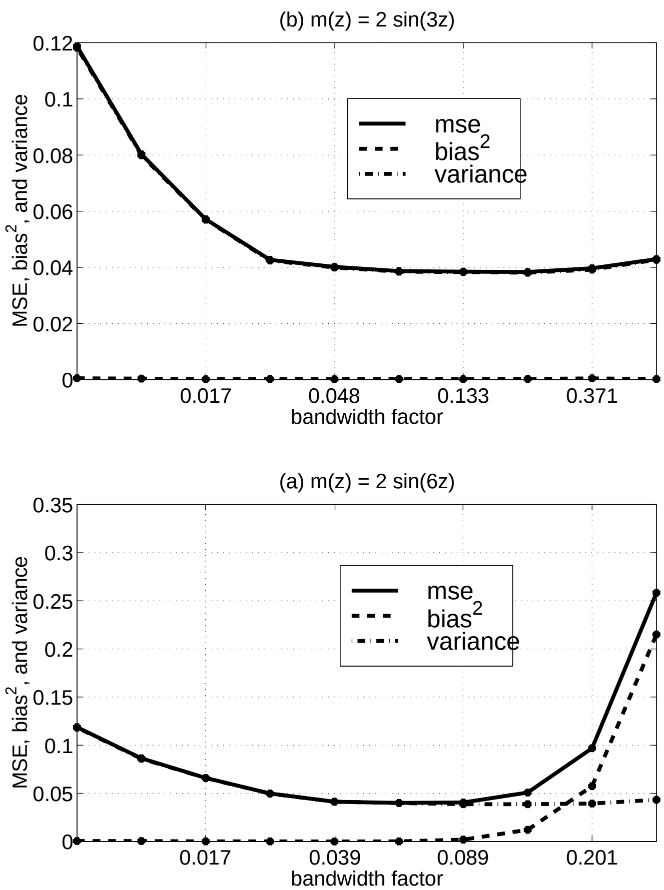

Figure 2: *Estimated squared bias (dashed), variance (dotted), and mean square error (solid) of* $\hat{\beta}$ *as functions of the bandwidth factor for smoothing* Z*. The estimates are based on 500 Monte Carlo samples, each of 250 observations. (a)* $m(z) = 2\sin(3z)$*; the squared bias is nearly zero so the mean square error overprints the variance and the squared bias overprints the horizontal axis. (b)* $m(z) = 2\sin(6z)$*; the squared bias grows rapidly with the bandwidth.*

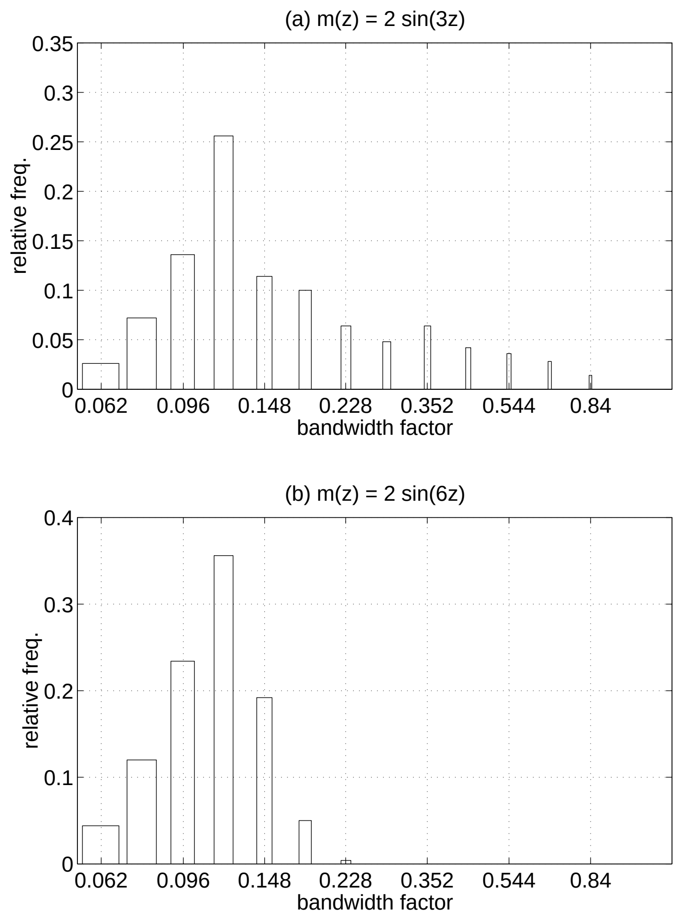

Figure 3: *Histograms of EBBS selected bandwidth factors from 500 samples each from, (a) the low bias case of* $f = 3$*, and (b) the high bias case of* $f = 6$*. The bins are of equal widths, but this is not apparent because the x-axis is logarithmic.*

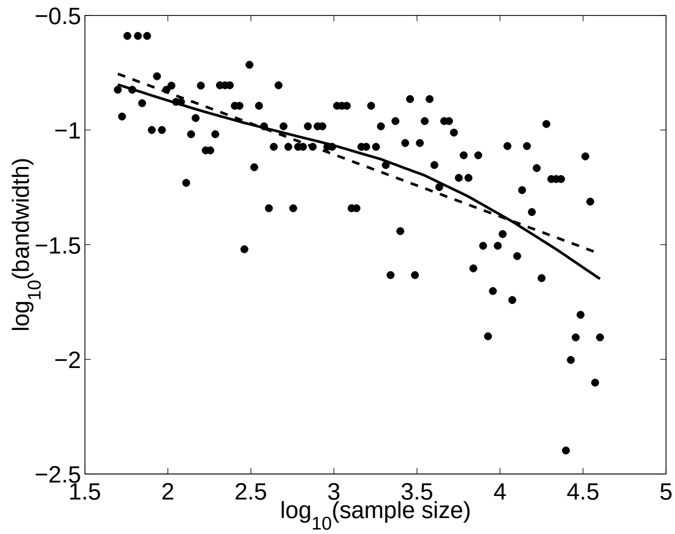

Figure 4: *Plot of* $\log_{10}(\hat{b})$ *against* $\log_{10} n$ *with a spline smooth (solid) and a linear fit (dashed) to the data.*

We also plotted $\log_{10}(\hat{b})$ against $\log_{10}(n)$ to see if the behavior of $\hat{b}$ was similar to what asymptotic theory predicts, i.e., whether the plot was linear with a slope of less than $-1/4$. Recall from Corollary 2.1 that $r < -1/4$ is necesary for $\hat{\boldsymbol{\beta}}$ to be $\sqrt{n}$-consistent. In Figure 4 we see that the plot is somewhat linear, though a spline smooth indicates some curvature. A linear fit to the data has a slope of $-.27$ which is just below $-1/4$. If one fits to only the last 70% of the data is slope is nearly $-1/3$.

Figure 5 is a plot of the logarithms of the conditional biases for five samples versus the logarithms of the bandwidth factor. There is a reference line with a slope of 2. One can see that the plots are nearly linear with slopes of 2 as predicted by asymptotics.

In summary, the results in Figure 4 are roughly consistent with asymptotics.

5 Discussion

Our theoretical work shows that backfitting can provide $\sqrt{n}$-consistent estimates of the parametric part, $\boldsymbol{\beta}$, of a semiparametric additive model (SAM), provided the bandwidth is chosen properly, i.e., not by traditional methods such as cross-validation.

We proposed a method of bandwidth selection and studied this method in a detailed simulation where five data factors and four tuning parameters were varied. The

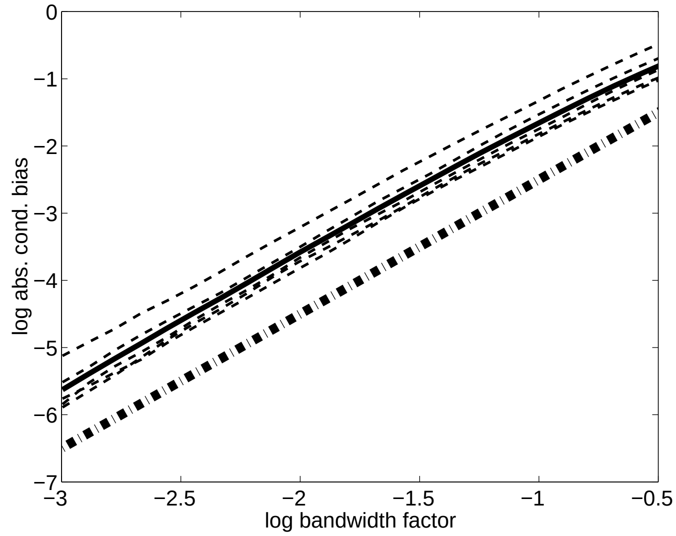

Figure 5: *Log of absolute conditional bias versus log of the bandwidth factor. Dashed curves are for five independent samples and the solid curve is their average. A thick dashed and dotted line with a slope of 2 is plotted for reference.*

method is insensitive to the choice of tuning parameters over the range of the data factors, indicating that the method can be applied in practice without needing to specify the tuning parameters carefully. Nonetheless, to provide guidance for data analysis, we do suggest default values of the tuning parameters.

The estimator $\hat{\beta}$ is relatively insensitive to the choice of bandwidth, provided the bandwidth is not so large that bias becomes a factor. In the simulation study, the proposed bandwidth methods avoided the high bias region.

The bandwidth selector requires that $\hat{\boldsymbol{\beta}}$ be computed on a grid of trial bandwidths, but this is no more computation than required by, say, cross-validation. The computational burden of our selector is not onerous for data sets less than, say, 1000 observations. If much larger data sets were routinely used for SAM modeling, then a binned implementation of backfitting might be contemplated.

One feature of the bandwidth selector is that when we are estimating $\boldsymbol{c}^T\boldsymbol{\beta}$ for some fixed vector $\boldsymbol{c}$, then the selected bandwidth depends on $\boldsymbol{c}$. This is to be expected since the optimal bandwidth will also depend on $\boldsymbol{c}$. In practice, one is usually interested in many choices of $\boldsymbol{c}$ and computing a separate bandwidth for each is not feasible. In that case, one could estimate a bandwidth for one interesting value of $\boldsymbol{c}$, say $\boldsymbol{c} = \bar{\boldsymbol{X}}$, and then use that bandwidth for all other $\boldsymbol{c}$'s as well.

In this article, we have treated the nonparametric component, $m(z)$, as a nuisance parameter. Obviously, there will be situations where m is at least as much interest as $\boldsymbol{\beta}$, but to keep our paper focused we have not considered this case.

As mentioned in Section 2.2, backfitting is more computationally efficient than direct estimation for the estimators studied in this article. Recently, estimators based on "marginal integration" have been proposed, e.g., by Linton and Nielsen (1995) and Fan, Härdle, and Mammen (1998). These estimators can be calculated explicitly rather than by backfitting. Backfitting has become common in practice, probably due to the influence of Hastie and Tibshirani's (1990) book and the availability of software in Splus. A comparison of backfitting with marginal integration methods would be quite useful but is beyond the scope of this article.

Acknowledgment

We thank the reviewers and associate editor for many very helpful suggestions that pushed us to clarify our ideas. Ruppert's research was supported by NSF Grants DMS-9626762 and DMS-9804058.

A Appendix: Outlines of Proofs

Proof of Theorem 2.1: For the conditional bias, we need to compute

$$\mathrm{E}(\hat{\boldsymbol{\beta}}|\boldsymbol{X},\boldsymbol{Z})-\boldsymbol{\beta}=\left(\frac{1}{n}\boldsymbol{X}^T(\boldsymbol{I}-\boldsymbol{S})\boldsymbol{X}\right)^{-1}\frac{1}{n}\boldsymbol{X}^T(\boldsymbol{I}-\boldsymbol{S})\boldsymbol{m}.$$

Consider first the matrix $\frac{1}{n}\boldsymbol{X}^T(\boldsymbol{I}-\boldsymbol{S})\boldsymbol{X}=\frac{1}{n}\boldsymbol{X}^T\boldsymbol{X}-\frac{1}{n}\boldsymbol{X}^T\boldsymbol{S}\boldsymbol{X}$. Using the same reasoning as in the proof of Theorem 2.1 in Ruppert and Wand (1994), it is straightforward to show that

$$\begin{aligned}\frac{1}{n}\boldsymbol{X}^T\boldsymbol{S}\boldsymbol{X} &= \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n}\boldsymbol{X}_i\boldsymbol{X}_j^T[\boldsymbol{s}_{Z_i}]_j \\ &= \mathrm{E}(\boldsymbol{X}_i\mathrm{E}(\boldsymbol{X}_i^T|Z_i))+O_p\left(h^2+\frac{1}{nh}\right). \qquad (6)\end{aligned}$$

The $O_p(h^2)$ term in the approximation comes from the off-diagonal terms in the sum $(i\neq j)$, while the $O_p(\frac{1}{nh})$ term comes form the diagonal ones $(i=j)$. Hence,

$$\left(\frac{1}{n}\boldsymbol{X}^T(\boldsymbol{I}-\boldsymbol{S})\boldsymbol{X}\right)^{-1}=\left(\mathrm{E}(\boldsymbol{X}_i(\boldsymbol{X}_i-\mathrm{E}(\boldsymbol{X}_i|Z_i))^T)\right)^{-1}+O_p\left(h^2+\frac{1}{nh}\right).$$

The order of this approximation can be derived using (6) and a cofactor expansion for the inverse (Horn and Johnson (1985), p.21).

From Opsomer and Ruppert (1997), we know that $(\boldsymbol{I}-\boldsymbol{S})\boldsymbol{m} = -\frac{\mu_2(K)}{2}h^2 D_z^2 \boldsymbol{m} + o_p(h^2)$, with D_z^r the rth derivative operator with respect to Z, and hence

$$\frac{1}{n}\boldsymbol{X}^T(\boldsymbol{I}-\boldsymbol{S})\boldsymbol{m} = -\frac{\mu_2(K)}{2}h^2 \mathrm{E}(\boldsymbol{X}_i m''(Z_i)) + o_p(h^2),$$

leading immediately to the desired bias approximation, since $\mathrm{cov}(\boldsymbol{X}_i, m''(Z_i)) = \mathrm{E}(\boldsymbol{X}_i m''(Z_i))$ by centering.

For the conditional variance of $\hat{\boldsymbol{\beta}}$, we need to approximate

$$\begin{aligned}
\mathrm{var}(\hat{\boldsymbol{\beta}}|\boldsymbol{X},\boldsymbol{Z}) &= \sigma^2\left(\boldsymbol{X}^T(\boldsymbol{I}-\boldsymbol{S})\boldsymbol{X}\right)^{-1}(\boldsymbol{X}^T(\boldsymbol{I}-\boldsymbol{S})(\boldsymbol{I}-\boldsymbol{S}^T)\boldsymbol{X})\left(\boldsymbol{X}^T(\boldsymbol{I}-\boldsymbol{S})\boldsymbol{X}\right)^{-1} \\
&= \sigma^2\left(\boldsymbol{X}^T(\boldsymbol{I}-\boldsymbol{S})\boldsymbol{X}\right)^{-1} \\
&\quad + \left(\boldsymbol{X}^T(\boldsymbol{I}-\boldsymbol{S})\boldsymbol{X}\right)^{-1}\left(\boldsymbol{X}^T\boldsymbol{S}^T\boldsymbol{X} - \boldsymbol{X}^T\boldsymbol{S}\boldsymbol{S}^T\boldsymbol{X}\right)\left(\boldsymbol{X}^T(\boldsymbol{I}-\boldsymbol{S})\boldsymbol{X}\right)^{-1}.
\end{aligned}$$

We therefore need to compute $\frac{1}{n}\boldsymbol{X}^T\boldsymbol{S}^T\boldsymbol{X}$ and $\frac{1}{n}\boldsymbol{X}^T\boldsymbol{S}\boldsymbol{S}^T\boldsymbol{X}$. The former is approximated by the same quantity as $\frac{1}{n}\boldsymbol{X}^T\boldsymbol{S}\boldsymbol{X}$ as computed in (6), while for the latter,

$$\frac{1}{n}\boldsymbol{X}^T\boldsymbol{S}\boldsymbol{S}^T\boldsymbol{X} = \frac{1}{n}\sum_{i=1}^n\sum_{j=1}^n\sum_{k=1}^n \boldsymbol{X}_i\boldsymbol{X}_j^T[\boldsymbol{s}_{Z_i}]_k[\boldsymbol{s}_{Z_j}]_k. \tag{7}$$

Let $K*K$ denote convolution of functions the kernel K with itself. Then, using the approach as in the proof of Theorem 2.1 in Ruppert and Wand (1994) for the asymptotic variance, we can show that

$$\frac{1}{n}\sum_{k=1}^n[\boldsymbol{s}_{Z_i}]_k[\boldsymbol{s}_{Z_j}]_k = \frac{1}{nh}(K*K)\left(\frac{Z_i-Z_j}{h}\right)\frac{1}{f_Z(Z_i)}(1+O_p(h^2)),$$

so that the terms in (7) for which $i \neq j \neq k$, converge to $\mathrm{E}(\boldsymbol{X}_i\mathrm{E}(X_i|Z_i)^T) + O_p(h^2)$. The rate at which these terms in (7) converge to this result is bounded by the rate of their standard deviation. We can show that

$$\mathrm{E}\left(\boldsymbol{X}_i\boldsymbol{X}_j^T\frac{1}{h^2}K\left(\frac{Z_i-Z_k}{h}\right)K\left(\frac{Z_j-Z_k}{h}\right)\frac{1}{f_Z(Z_i)}\frac{1}{f_Z(Z_j)}\right)^2 = O_p(\frac{1}{h^2}),$$

so that the standard deviation of the non-diagonal terms in (7) is $O_p(1/\sqrt{n^3h^2})$. All other terms in the sum (7) can be shown to be of order $O_p(\frac{1}{nh})$. Therefore,

$$\frac{1}{n}\boldsymbol{X}^T\boldsymbol{S}^T\boldsymbol{X} - \frac{1}{n}\boldsymbol{X}^T\boldsymbol{S}\boldsymbol{S}^T\boldsymbol{X} = O_p(h^2+\frac{1}{nh})$$

and the variance result immediately follows.

■

Proof of Theorem 2.2: For the bias result, we note

$$\mathrm{E}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta} | \boldsymbol{X}, \boldsymbol{Z}) = -(\boldsymbol{X}^T(\boldsymbol{I} - \boldsymbol{W}_M)\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{B}_M,$$

where $\boldsymbol{B}_M$ is the bias for fitting an additive model to the D nonparametric terms in (2). A reasoning analogous to that in the proof of Theorem 3.1 of Opsomer (1996) shows

$$\frac{1}{n}\boldsymbol{X}^T\boldsymbol{B}_M = \frac{1}{n}\boldsymbol{X}^T\bar{m}_M + \frac{1}{n}\sum_{d=1}^{D}\boldsymbol{X}^T(\boldsymbol{I} - \boldsymbol{S}_d^*\boldsymbol{W}_M^{(-d)})^{-1}(\boldsymbol{Q}_d^* - \boldsymbol{S}_d^*\boldsymbol{B}_M^{(-d)}) + o_p(h_d^2),$$

where the superscript $(-d)$ indicate the smoother and bias for a $(D-1)$-dimensional additive model with mean function $m_M - m_d$. Since $\frac{1}{n}\boldsymbol{X}^T\bar{m}_M = o_p(1/\sqrt{n})$ and

$$(\boldsymbol{I} - \boldsymbol{S}_d^*\boldsymbol{W}_M^{(-d)})^{-1} = \boldsymbol{I} + O_p(\frac{1}{n}), \tag{8}$$

we can show recursively that $\frac{1}{n}\boldsymbol{X}^T\boldsymbol{B}_M = O_p(\sum_{d=1}^{D} h_d^2\mathbf{1}_P)$. Similarly,

$$\begin{aligned}\frac{1}{n}\boldsymbol{X}^T(\boldsymbol{I} - \boldsymbol{W}_M)\boldsymbol{X} &= \frac{1}{n}\sum_{d=1}^{D}\boldsymbol{X}^T(\boldsymbol{I} - \boldsymbol{S}_d^*\boldsymbol{W}_M^{(-d)})^{-1}(\boldsymbol{I} - \boldsymbol{S}_d^*)\boldsymbol{X} \\ &= \mathrm{E}(\boldsymbol{X}_i\sum_{d=1}^{D}\mathrm{E}(\boldsymbol{X}_i^T|Z_{di}))\,(1 + O_p(\frac{1}{n})).\end{aligned}$$

For the variance, (8) and the fact that $\boldsymbol{X}^T\boldsymbol{S}_d^* = \mathrm{E}(\boldsymbol{X}_i|Z_{di})(1 + O_p(1))$ are used to establish the rate of the leading term.

■

References

[1] A. BUJA, T.J. HASTIE, and R.J. TIBSHIRANI (1989). Linear smoothers and additive models. *Annals of Statistics* **17**, 453–555.

[2] R.J. CARROLL, J. FAN, I. GIJBELS and M.P. WAND (1997). Generalized partially linear single-index models. *Journal of the American Statistical Association* **92**, 477–489.

[3] J. FAN and I. GIJBELS(1995). 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.

[4] J. FAN, W. HÄRDLE, and E. MAMMEN (1998). Direct estimation of low-dimensional components in additive models. *Annals of Statistics* **26**, 943–971.

[5] J. GAO (1995). Asymptotic theory for partly linear models. *Communications in Statistics – Theory and Methods* **24**, 1985–2009.

[6] T. GASSER, A. KNEIP, and W. KÖHLER (1991). A flexible and fast method for automatic smoothing. *Journal of the American Statistical Association* **86**, 643–652.

[7] T. J. HASTIE and R. J. TIBSHIRANI (1990). *Generalized Additive Models.* Chapman and Hall, Washington, D.C.

[8] N.E. HECKMAN (1986). Spline smoothing in a partly linear model. *Journal of the Royal Statistical Association, Series B* **48**, 244–248.

[9] R. A. HORN and C. A. JOHNSON (1985). *Matrix Analysis.* Cambridge University Press, Cambridge, U.K.

[10] O. B. LINTON (1995). Second order approximation in the partially linear regression model. *Econometrica* **63**, 1079–1112.

[11] O. B. LINTON and J. P. NIELSEN (1995). A kernel method of estimating structured nonparametric regression based on marginal integration. *Biometrika* **82**, 93–101.

[12] J.-D. OPSOMER (20 February 1996). On the properties of backfitting estimators. Preprint 96–12, Department of Statistics, Iowa State University.

[13] J.-D. OPSOMER and D. RUPPERT (1997). Fitting a bivariate additive model by local polynomial regression. *Annals of Statistics* **25**, 186–211.

[14] J.-D. OPSOMER and D. RUPPERT (1998). A fully automated bandwidth selection method for fitting additive models. *The Journal of the American Statistical Association* **93**, 605–620.

[15] J.A. RICE (1986). Convergence rates for partially splined models. *Statistics and Probability Letters* **4**, 203–208.

[16] D. RUPPERT (1997). Empirical-bias bandwidths for local polynomial nonparametric regression and density estimation. *Journal of the American Statistical Association* **92**, 1049–1062.

[17] D. RUPPERT, S. J. SHEATHER, and M. P. WAND (1995). An effective bandwidth selector for local least squares regression. *Journal of the American Statistical Association* **90**, 1257–1270.

[18] D. RUPPERT and M. P. WAND (1994). Multivariate locally weighted least squares regression. *Annals of Statistics* **22**, 1346–1370.

[19] P.E. SPECKMAN (1988). Regression analysis for partially linear models. *Journal of the Royal Statistical Association, Series B* **50**, 413–436.

[20] C.J. STONE (1985). Additive regression and other nonparametric models. *Annals of Statistics* **13**, 689–705.