Local additive estimation

Juhyun Park and Burkhardt Seifert *

Lancaster University and University of Zürich

April 15, 2008

Abstract

Additive models are popular in high–dimensional regression problems because of flexibility in model building and optimality in additive function estimation. Moreover, they do not suffer from the so-called *curse of dimensionality* generally arising in nonparametric regression setting. Less known is the model bias incurring from the restriction to the additive class of models. We introduce a new class of estimators that reduces additive model bias and at the same time preserves some stability of the additive estimator. This estimator is shown to partially relieve the dimensionality problem as well. The new estimator is constructed by localizing the assumption of additivity and thus named *local additive estimator*. Implementation can be easily made with any standard software for additive regression. For detailed analysis we explicitly use the smooth backfitting estimator by Mammen, Linton and Nielsen (1999).

*Juhyun Park was Postdoctoral Research Fellow (Email:*juhyun.park@lancaster.ac.uk*), Burkhardt Seifert is Professor (Email:*seifert@ifspm.uzh.ch*) in Biostatistics unit, Institute for Social and Preventive Medicine, University of Zürich, Zürich, Switzerland. Funding for this work was provided by Swiss National Science Foundation grant 20020-103743.

KEY WORDS: Nonparametric regression, additive models, backfitting, local polynomial smoothing.

1 Introduction

Application of additive models is numerous from econometrics, social sciences to environmental sciences (Deaton and Muellbauer 1980; Hastie and Tibshirani 1990). Separability of each component is well suited for flexible and interpretable model building in modern high dimensional problems with many covariates. The main advantage of additive regression is that it allows us to deal with high-dimensional regression in one-dimensional precision.

Since the recognition of potential of additive models in 80s, several additive estimators have been developed in various contexts of smoothing. Earlier methods tend to be more algorithmic in nature because of nontrivial analyses required to understand the behaviour of estimators (see Opsomer and Ruppert 1997; Opsomer 2000). More recent methods include marginal integration by Linton and Nielsen (1995) and smooth backfitting by Mammen, Linton and Nielsen (1999). The smooth backfitting estimator (SBE) is shown to be oracle optimal for the additive function estimation, that is, it achieves the same precision as in one–dimensional regression. The SBE is also applicable when additivity is only approximately valid by means of a projection idea (Mammen et al. 2001).

Less known is the model bias incurring from the restriction to the additive class of models. Additive models miss important (nonadditive) features by considering the nonadditive part nuisance or noise. This is also related to the fact that fitting additive models and diagnostics are less trivial in that it involves various issues concerning model selection and stability (Breiman 1993).

Models without additive restriction fall in the broad category of nonparametric

regression models. Their properties have been well established in several earlier works, one of which points out that local linear estimator is minimax optimal in more than one–dimensional regression problem (Fan et al. 1997). However, as the dimension of the variables grows, the stability of the estimation becomes increasingly an issue, which brings about *curse of dimensionality* (see, e.g. Stone 1980, 1982).

This situation leads to the question whether or how to combine advantages of those estimators, the stability of additive estimator and the optimality of local linear one. The approach proposed in Studer et al. (2005) uses penalty to the nonadditive part, which produces a family of *regularised* estimators. In this paper, we introduce another class of estimators by *localizing* the additivity assumption and this will be named *local additive estimator*.

Let $(\mathbf{X}, Y)$ be random variables of dimensions d and 1, respectively and let $(\mathbf{X}_i, Y_i), i = 1, \cdots, n$, be independent and identically distributed random variables from $(\mathbf{X}, Y)$. Denote the design density of $\mathbf{X}$ by $f(\mathbf{x})$. We assume that $\mathbf{X}$ has compact support $[-1, 1]^d$. The regression function $r(\mathbf{x}) = E[Y|\mathbf{X} = \mathbf{x}]$ is assumed to be smooth. The additive model has the relation

$$r(\mathbf{x}) = r_0 + r_1(x_1) + \cdots + r_d(x_d)\,. \tag{1}$$

This is a global assumption on the shape of the regression function and thus quite restrictive.

Given $\mathbf{x}$, consider a $\mathbf{w}$-neighborhood of $\mathbf{x}$. If $||\mathbf{w}||$ is small enough, by Taylor theorem, we would have

$$r(\mathbf{x}) \approx r_0 + r_1(x_1) + \cdots + r_d(x_d)\,.$$

Note that this is not an *assumption* on the model. The accuracy of the approximation clearly depends on the $\mathbf{w}$-neighborhood. We will call this approximate additive relation *local additivity*.

The above argument naturally leads to an estimator that can be constructed from additive estimator using data in the neighborhood of interest. For a given point $\mathbf{x}_0$, construct an additive estimator using data in the $\mathbf{w}$-neighborhood of $\mathbf{x}_0$. The new estimator is defined as the predictor of the additive estimator at $\mathbf{x} = \mathbf{x}_0$. This will be termed *local additive estimator*, denoted by $\hat{r}_{ladd}(\mathbf{x}_0)$. A formal definition is given in Section 2.

By not directly imposing the additive restriction, we reduce model bias. On the other hand, the merit of additivity that allows us to deal with high–dimensional regression in one–dimensional precision is partially lost. The main advantages of the new estimator can be summarized as follows. 1) Additivity is approximately valid *locally* even when the true regression function is not additive. This helps keep bias small for general regression function. 2) The local additive approximation is more flexible than the local linear one. Thus, the local region for the additive estimator can be chosen larger than that for the local linear one, which improves variance of the estimator. 3) Standard software for additive estimators is directly applicable.

The paper is organized as follows. We formulate main results in Section 2, followed by asymptotic comparison to the local linear estimator, $\hat{r}_{ll}$, and the additive estimator, $\hat{r}_{add}$ as an illustration. Smoothing parameter selection is also discussed. Numerical studies are found in Section 3 with an application to a real data example. An extended version of simulation studies and some proofs of Section 2.5 are found in Park and Seifert (2008).

2 Local additive estimation

2.1 Preliminaries

Let $\mathbf{x}_0$ be a fixed interior output point. For $\mathbf{w} = (w_1, \cdots, w_d)$, we apply an additive estimator $\hat{r}_{add}$ using data in a $\mathbf{w}$-neighborhood of $\mathbf{x}_0$. Our analysis is based on d–dimensional rectangular region $[\mathbf{x}_0 \pm \mathbf{w}] = \{\mathbf{X}_i, \mathbf{X}_i \in [\mathbf{x}_0 - \mathbf{w}, \mathbf{x}_0 + \mathbf{w}]\}$. Denote the number of observations $\mathbf{X}_i$ in $[\mathbf{x}_0 \pm \mathbf{w}]$ by $\tilde{n}$. Properties of the local additive estimator can be developed by rescaling the region $[\mathbf{x}_0 \pm \mathbf{w}]$ to $[-1, 1]^d$ and then using results known for $\hat{r}_{add}$. We will consider additive estimators that reach the optimal order $O(n^{-4/5})$. For technical reasons, we will focus on linear estimators, which enable us to compute expectations under Taylor expansions. The SBE by Mammen et al. (1999) is known to be oracle optimal under general conditions, and other estimators inherit this optimality under more special situations (Linton and Nielsen 1995; Opsomer and Ruppert 1997; Opsomer 2000). Throughout the article, we will assume that

(A.1) The regression function r and the design density f are twice continuously differentiable.

The special case of uniform design will be separately dealt with later in this section.

When additive estimator is viewed as a componentwise one–dimensional smoother, it has inherently a smoothing parameter associated with it. It may refer to smoothing window h as in kernel smoothers, smoothing parameter λ as in smoothing splines, or generally degrees of freedom df as in equivalent linear smoothers. We will stick to h for a smoothing parameter, as the local linear smoother is used later in our analysis.

Suppose that all w_j's are of same order. For simplicity of notation let $w_j = w$. Let $w \to 0$ and $h_j/w \to 0$. Write

$$\mathbf{U} = \frac{\mathbf{X} - \mathbf{x}_0}{w}\,, \tag{2}$$

for the rescaled random variable on $[-1,1]^d$ with density

$$\begin{aligned}\tilde{f}(\mathbf{u}) &= f(\mathbf{x}_0 + w\mathbf{u}) / \int_{[-1,1]^d} f(\mathbf{x}_0 + w\mathbf{u})\, d\mathbf{u} \\ &= \frac{f(\mathbf{x}_0 + w\mathbf{u})}{2^d f(\mathbf{x}_0)} + O(w^2)\,. \end{aligned} \tag{3}$$

The corresponding regression function is

$$\tilde{r}(\mathbf{u}) = r(\mathbf{x}_0 + w\mathbf{u}) \tag{4}$$

and the transformed bandwidth is

$$\tilde{h}_j = h_j / w\,. \tag{5}$$

The local additive estimator at $\mathbf{x}_0$ is defined as $\hat{r}_{ladd}(\mathbf{x}_0) = \widehat{\tilde{r}}_{add}(\mathbf{0})$.

Denote 1st and 2nd partial derivatives of r by $r'_j(\mathbf{x})$, $r''_{j,k}(\mathbf{x})$ and the $d \times d$ matrix of 2nd derivatives by $\mathbf{r}''$. $\hat{r}_{ll}(\mathbf{x}_0)$ and the local additive estimator by $\hat{r}_{ladd}(\mathbf{x}_0)$. We write E, B, V, MSE, ISE, ASE, MISE and MASE for the conditional expectation, bias, variance, mean squared error, integrated squared error, average squared error, integrated mean squared error and average mean squared error, respectively. Define a matrix norm $||\cdot||$ for a symmetric matrix $A = \{a_{ij}\}$ as $||A|| = max_{i,j}|a_{ij}|$ and write $||\cdot||_2$ for the usual L_2 norm.

Let us first consider a bilinear function of components u_j and u_k as

$$b^{jk}(\mathbf{u}) = (u_j - \bar{U}_j)(u_k - \bar{U}_k)\,,$$

where $\bar{U}_j$ and $\bar{U}_k$ are jth and kth marginal averages of $\mathbf{U}$ in (2). Note that $\bar{U}_j$ and $\bar{U}_k$ are considered constants given $\mathbf{U}$. We will see that studying this function is revealing when applying Taylor expansions in the proof of our main results. Let f_w be a sequence of design densities that converges to uniform. This can be constructed, for example, as in (3) by defining $f_w(\mathbf{u}) = \frac{f(\mathbf{x}_0 + w\mathbf{u})}{2^d f(\mathbf{x}_0)}$ for a density f satisfying (A.1). Let

$\hat{b}^{jk}_{add,w}$ be the corresponding additive estimator. If u_j and u_k are uniformly distributed, as $n \to \infty$, $b^{jk}(\mathbf{0}) \to 0$ and $\hat{b}^{jk}_{add,0} \to 0$. Thus, $\hat{b}^{jk}_{add,w}(\mathbf{0})$ should converge to zero too. Surprisingly enough, the case of vanishing second partial derivatives needs special attention. Denote

$$A_{j,k} = \left\{ \mathbf{x} \in [-1,1]^d \mid r''_{j,k}(\mathbf{x}) = 0 \right\} . \tag{6}$$

Without higher order smoothness assumption, the results below are only valid for $\mathbf{x}_0$ outside the borders $\partial A_{j,k}$ of $A_{j,k}$. We claim however that these borders are small and can be ignored for most practical situations, as explained in the remarks following Proposition 1 in Section 2.5.

In addition to $(A.1)$, the following assumptions are made.

$(A.2)$ The kernel K is bounded, has compact support, is symmetric around 0 and is Lipschitz continuous.

$(A.3)$ The density f of $\mathbf{x}$ is bounded away from zero and infinity on $[-1,1]^d$.

$(A.4)$ For some $\theta > 5/2, E[|Y|^\theta] < \infty$.

$(A.5)$ $\tilde{h}_j \to 0$ such that $\tilde{n}\tilde{h}_j^d / \ln \tilde{n} \to \infty$ as $\tilde{n} \to \infty$.

2.2 Main result

Theorem 1. *Assume that $\hat{r}_{add}$ is linear in Y and oracle optimal. Let f_w be a sequence of design densities that converges to uniform f_0 and $\hat{r}_{add,w}$ be the corresponding additive estimator. Assume that $\hat{r}_{add,w}$ converges as f_w converges and satisfies*

$$|\hat{b}^{jk}_{add,w}(\mathbf{0}) - \hat{b}^{jk}_{add,0}(\mathbf{0})| \leq L||f_w - f_0||_2^2 \text{ for all } j \neq k ,$$

where L is a constant. Then, for all $\mathbf{x}_0 \notin \bigcup_{j,k} \partial A_{j,k}$ defined in (6),

$$\begin{aligned} B^2[\hat{r}_{ladd}(\mathbf{x}_0)] &= max\{O(h^4), O(w^8 + w^4 \max_{j,k} |\hat{b}^{jk}_{add,0}(\mathbf{0})|^2)\} \\ V[\hat{r}_{ladd}(\mathbf{x}_0)] &= O((nw^{d-1}h)^{-1}) . \end{aligned}$$

Proof. Here, we will present the main ideas for bias. Because the estimator is linear, we have

$$E[\hat{r}_{add}(\mathbf{x}_0)] = \frac{1}{n}\sum_{i=1}^{n} W_i(\mathbf{x}_0, \mathbf{X}_i) r(\mathbf{X}_i)\,.$$

Similarly, for the local additive estimator, we have

$$E[\hat{r}_{add,w}(\mathbf{x}_0)] = \frac{1}{\tilde{n}}\sum_{i=1}^{\tilde{n}} \tilde{W}_i(\mathbf{0}, \mathbf{U}_i) \tilde{r}(\mathbf{U}_i)\,,$$

where $\mathbf{U}$ is given in (2).

$$\begin{aligned}
\tilde{r}(\mathbf{U}_i) &= r(\mathbf{x}_0 + w\mathbf{U}_i) = r(\mathbf{x}_0) + w\sum_j r'_j(\mathbf{x}_0)U_{ij} + \frac{w^2}{2}\sum_{j,k} r''_{j,k}(\mathbf{x}_0)U_{ij}U_{ik} + R(\mathbf{x}_0, \mathbf{U}_i) \\
&= additive + \frac{w^2}{2}\sum_{j\neq k} r''_{j,k}(\mathbf{x}_0)U_{ij}U_{ik} + R(\mathbf{x}_0, \mathbf{U}_i)\,.
\end{aligned}$$

Thus,

$$\begin{aligned}
B[\hat{r}_{ladd}(\mathbf{x}_0)] &= \frac{1}{\tilde{n}}\sum_{i=1}^{\tilde{n}} \tilde{W}_i(\mathbf{0}, \mathbf{U}_i)\tilde{r}(\mathbf{U}_i) - r(\mathbf{x}_0) \\
&= B[additive] + \frac{w^2}{2}\sum_{j\neq k} r''_{j,k}(\mathbf{x}_0)\Big(\frac{1}{\tilde{n}}\sum_{i=1}^{\tilde{n}} \tilde{W}_i(\mathbf{0}, \mathbf{U}_i)U_{ij}U_{ik}\Big) \\
&\quad + \frac{1}{\tilde{n}}\sum_{i=1}^{\tilde{n}} \tilde{W}_i(\mathbf{0}, \mathbf{U}_i) R(\mathbf{x}_0, \mathbf{U}_i)\,.
\end{aligned}$$

Because of oracle optimality of the estimator, the bias of the additive part becomes

$$\begin{aligned}
B[additive] &= \frac{\tilde{h}^2}{2}\Big(\frac{w^2}{2}\sum_j 2r''_{j,j}(\mathbf{x}_0)\Big) + o(\tilde{h}^2 w^2) \\
&= \frac{h^2}{2}\sum_j r''_{j,j}(\mathbf{x}_0) + o(h^2)\,, \qquad (7)
\end{aligned}$$

the latter equality following from (5). For the leading nonadditive term, first consider

$$\frac{1}{\tilde{n}}\sum_{i=1}^{\tilde{n}} \tilde{W}_i(\mathbf{0}, \mathbf{U}_i)U_{ij}U_{ik}\,.$$

Observe that

$$U_{ij}U_{ik} = (U_{ij} - \bar{U}_j)(U_{ik} - \bar{U}_k) + \bar{U}_j U_{ik} + \bar{U}_k U_{ij} + \bar{U}_j \bar{U}_k \,. \tag{8}$$

Given $\mathbf{U}_i$, the last three terms are linear and thus do not add additional bias. Therefore, we focus on

$$\frac{1}{\tilde{n}} \sum_{i=1}^{\tilde{n}} \tilde{W}_i(\mathbf{0}, \mathbf{U}_i)(U_{ij} - \bar{U}_j)(U_{ik} - \bar{U}_k) = \hat{b}^{jk}_{add,w}(\mathbf{0}) \,.$$

This is nothing but the additive estimator at $\mathbf{0}$ when the design density is f_w and the true regression function is the bilinear function b^{jk}. It may be written as

$$\hat{b}^{jk}_{add,w}(\mathbf{0}) = \hat{b}^{jk}_{add,w}(\mathbf{0}) - \hat{b}^{jk}_{add,0}(\mathbf{0}) + \hat{b}^{jk}_{add,0}(\mathbf{0}) \,.$$

Thus,

$$\begin{aligned} |\hat{b}^{jk}_{add,w}(\mathbf{0})| &\leq L||f_w - f_0||_2^2 + |\hat{b}^{jk}_{add,0}(\mathbf{0})| \\ &= O(w^2 + |\hat{b}^{jk}_{add,0}(\mathbf{0})|) \,. \end{aligned} \tag{9}$$

Therefore, the second term is of order $O(w^2)O(w^2 + |\hat{b}^{jk}_{add,0}(\mathbf{0})|)$. The last remainder term may be written as

$$w^2 \sum_{j,k} \Big(\frac{1}{\tilde{n}} \sum_{i=1}^{\tilde{n}} \tilde{W}_i(\mathbf{0}, \mathbf{U}_i) U_{ij} U_{ik} \int_0^1 (1-\theta) \{ r''_{j,k}(\mathbf{x}_0 + \theta w \mathbf{U}_i) - r''_{j,k}(\mathbf{x}_0) \} \, d\theta \Big) \,.$$

As $\mathbf{r}''$ is continuous, the integrands are $o(1)$ and the corresponding terms become negligible compared to the main term above, if $r''_{j,k}(\mathbf{x}_0) \neq 0$. If $r''_{j,k}(\mathbf{x}) = 0$ in a neighborhood of $\mathbf{x}_0$, the corresponding integrand vanishes. Hence, the result follows from (7) and (9). □

To demonstrate the idea of our result, we make a rough comparison to the existing results in the following two sections by differentiating a situation with additive regression function from that with general regression function.

2.3 Behavior for additive regression function

When the true regression function is additive, the additive estimator $\hat{r}_{add}$ has MSE of $O(n^{-4/5})$ and the local linear estimator $\hat{r}_{ll}$ has MSE of $O(n^{-4/(4+d)})$. We can see this from

$$V\big[\hat{r}_{ll}(\mathbf{x}_0)\big] = O\big((nh^d)^{-1}\big)\,, \qquad B^2\big[\hat{r}_{ll}(\mathbf{x}_0)\big] = O\big(h^4||\mathbf{r}''||^2\big) = O\big(h^4\big)\,,$$
$$V\big[\hat{r}_{add}(\mathbf{x}_0)\big] = O\big((nh)^{-1}\big)\,, \qquad B^2\big[\hat{r}_{add}(\mathbf{x}_0)\big] = O\big(h^4(||\mathbf{r}''||)^2\big) = O(h^4)\,.$$

The local additive estimator $\hat{r}_{ladd}$ should beat the local linear estimator and come as close to the additive one as possible. With the same principle, the local additive estimator would have

$$\begin{aligned} V\big[\hat{r}_{ladd}(\mathbf{x}_0)\big] &= O\big((\tilde{n}\tilde{h})^{-1}\big) = O\big((nw^{d-1}h)^{-1}\big)\,, \\ B^2\big[\hat{r}_{ladd}(\mathbf{x}_0)\big] &= O\big(\tilde{h}^4(||\tilde{\mathbf{r}}''||)^2\big) = O(h^4)\,. \end{aligned}$$

Obviously, the additive estimator is optimal, the local linear estimator is worst, and the local additive estimator is in between.

2.4 Behavior for general regression function

Now consider the general case. Note that properties of additive estimators for general regression functions are not well studied. Nevertheless, when the true regression function is not additive, bias of the additive estimator is $O(1)$. Variance does not depend on the regression function and thus remains the same. Thus we have

$$V\big[\hat{r}_{ll}(\mathbf{x}_0)\big] = O\big((nh^d)^{-1}\big)\,, \qquad B^2\big[\hat{r}_{ll}(\mathbf{x}_0)\big] = O\big(h^4||\mathbf{r}''||)^2\big) = O\big(h^4\big)\,,$$
$$V\big[\hat{r}_{add}(\mathbf{x}_0)\big] = O\big((nh)^{-1}\big)\,, \qquad B^2\big[\hat{r}_{add}(\mathbf{x}_0)\big] = O\big(||\mathbf{r}''||^2\big) = O(1)\,.$$

Applying the same principle to the local additive estimator would lead to

$$\begin{aligned} V\big[\hat{r}_{ladd}(\mathbf{x}_0)\big] &= O\big((\tilde{n}\tilde{h})^{-1}\big) = O\big((nw^{d-1}h)^{-1}\big)\,, \\ B^2\big[\hat{r}_{ladd}(\mathbf{x}_0)\big] &= O\big(||\tilde{\mathbf{r}}''||^2\big) = O(w^4)\,. \end{aligned}$$

We will show (Theorem 2) that the limit for the bias of $\hat{r}_{ladd}(\mathbf{x}_0)$ can be further improved to $B^2[\hat{r}_{ladd}(\mathbf{x}_0)] = O(w^8)$ using the SBE.

2.5 Local additive estimator based on the SBE

When the regression function is additive, it can be shown that there is no loss in bias with local additive estimator compared to additive estimator. For general case, the local additive estimator based on the SBE satisfies the requirements of Theorem 1. Note that for the SBE, existence and convergence occur with probability tending to one (see Mammen et al. 1999), thus our statements imply the same without explicitly mentioning it.

The results are valid under quite general distributions, see assumption (A.4). For simplicity of notation we will assume that the residuals ε have constant variance σ^2 whenever appropriate.

Theorem 2. *The local additive estimator $\hat{r}_{ladd}$ based on the smooth backfitting estimator fulfills Theorem 1 with $\hat{b}^{jk}_{add,w} = O(w^2)$ and*

$$V[\hat{r}_{ladd}(\mathbf{x}_0)] = 2\mu_0(K^2)\sigma^2 \sum_{j=1}^{d} \left(nw^{d-1}h_j\right)^{-1} (1 + o(1)) .$$

Corollary 1. *For all $\mathbf{x}_0 \notin \bigcup_{j,k} \partial A_{j,k}$ defined in (6), the local additive estimator $\hat{r}_{ladd}$ based on the smooth backfitting estimator has $B^2[\hat{r}_{ladd}(\mathbf{x}_0)] = \max\{O(h^4), O(w^8)\}$.*

In brief, the projection property of the SBE together with (A.1) helps reducing the bias for the general regression function. In summary we have for general regression function

$$MSE[\hat{r}_{ladd}(\mathbf{x}_0)] = O(h^4 + w^8 + (nw^{d-1}h)^{-1}) . \tag{10}$$

Corollary 2. *Assume that $d \leq 8$. Optimal orders of w and h of the local additive estimator $\hat{r}_{ladd}$ based on the smooth backfitting estimator are given by*

$$w \sim n^{-1/(9+d)}\,, \quad h \sim n^{-2/(9+d)}\,,$$

leading to

$$MSE[\hat{r}_{ladd}(\mathbf{x}_0)] \sim n^{-8/(9+d)} = n^{-4/\left(4+\frac{d+1}{2}\right)}\,.$$

In comparison, the optimal local linear estimator achieves $O(n^{-4/(4+d)})$. The reduction of dimensionality is explained by the factor $\tilde{d} = \frac{d+1}{2}$, the *equivalent dimension*. For example when $d = 3$ the local additive estimator behaves similar to a local linear estimator with $\tilde{d} = 2$, and when $d = 5$ it will be reduced to $\tilde{d} = 3$. Thus, local additive estimation provides some relaxation of dimensionality in nonparametric regression compared to the minimax local linear estimator.

It turns out that the existence of second derivatives is not sufficient to derive explicit coefficients for leading terms. Below we deal with the special situation of a uniform design with higher order smoothness assumption.

($A.1'$) The regression function r is four times continuously differentiable and f is uniform.

Proposition 1. *Suppose that ($A.1'$) holds. Bias of the local additive estimator $\hat{r}_{ladd}$ based on the smooth backfitting estimator is given by*

$$B[\hat{r}_{ladd}(\mathbf{x}_0)] = \Big(\frac{\mu_2(K)}{2}\sum_{j=1}^{d} h_j^2 r''_{j,j}(\mathbf{x}_0) - \frac{w^4}{4!\cdot 9}\sum_{j\neq k} r''''_{j,j,k,k}(\mathbf{x}_0)\Big) + o\left(h^2 + w^4\right)\,.$$

Contrary to Theorems 1 and 2, the Proposition is valid without any exclusion of boundaries $\partial A_{j,k}$, which implies that the restriction there is related to irregular points of the regression function only. It should be mentioned however that irrespective of condition ($A.1'$) the MSE is always of order $O\left(h^4 + w^6 + (nw^{d-1}h)^{-1}\right)$ if r'' is

Lipschitz continuous. Thus, the local additive estimator *works* also at the remaining boundaries. Proposition 1 additionally shows that higher order smoothness assumption would not help further reduce bias. Moreover, it can be deduced from the proof (not shown) that the existence of $\mathbf{r}''$ is not sufficient to derive leading terms.

The optimal smoothing parameters are determined in the following. Define

$$a = \frac{\mu_2(K)}{2} \sum_j r''_{j,j}(\mathbf{x}_0)\,, \quad b = \frac{1}{4! \cdot 9} \sum_{j \neq k} r''''_{j,j,k,k}(\mathbf{x}_0)\,, \quad c = 2d\mu_0(K^2)\sigma^2\,.$$

Proposition 2. *Suppose that (A.1′) holds. Assume that $h_j = h$ and let $h = C_h w^2$. The smoothing parameter w that minimizes asymptotic MSE is given by*

$$w = \left(\frac{c(d+1)}{8C_h(aC_h^2 - b)^2} \right)^{1/(9+d)} n^{-1/(9+d)}\,.$$

Proposition 3. *Under the same assumptions as in Proposition 2, the optimal choice of C_h is given by*

$$C_h = \sqrt{\frac{2}{d-1}\Big(-\frac{b}{a}\Big)}\,.$$

provided that $ab < 0$.

Properties of the local additive estimator based on the SBE are studied in detail in Park and Seifert (2008). Proofs of Propositions 1–3 are found there and results of Theorem 2 and Corollary 1 can be deduced directly from results formulated there.

2.6 Data-adaptive parameter selection

We consider smoothing parameter selection based on model selection criteria for general regression function estimation.

Although asymptotic equivalence of classical model selection criteria has long been recognized (Härdle et al. 1988), because of small sample behavior, several versions of model selection criteria exist (Hurvich and Simonoff 1998). Still most discussions were limited to one dimensional problem.

For additive models with ordinary backfitting estimator, Opsomer and Ruppert (1998) proposed a plug-in bandwidth selector and Wood (2000) proposed generalized cross-validation approach for additive models with penalized regression splines. For additive models with smooth backfitting estimator, Nielsen and Sperlich (2005) discussed cross validation, while Mammen and Park (2005) proposed a bandwidth selection method which minimizes a penalized sum of squared residuals

$$PLS = \hat{\sigma}^2 \left(1 + 2 \sum_j \frac{1}{nh_j} K(0) \right)$$

and noted that it is computationally more feasible than cross validation. They also conjectured about model misspecification (p. 1263) that *...the penalized least squares bandwidth will work reliably also under misspecification of the additive model. This conjecture is supported by the definition of this bandwidth...* but pointed out the difficulty involved in the theory (p. 1267).

For nonadditive models Studer et al. (2005), in the context of penalized additive regression approach, investigated parameter selection based on AIC-type model selection criteria such as AIC, GCV, and AIC_C (Hurvich et al. 1998) and established asymptotic equivalence of these estimators in multivariate local linear regression for $d \leq 4$ where the estimator satisfies stability condition. Note that the additive SBE uses only two-dimensional marginal densities and thus such restriction is not necessary.

We investigate smoothing parameter selection based on AIC-type model selection criteria and show that PLS is equivalent to AIC-type model selection criteria. Because the local additive estimator based on the SBE uses two-dimensional densities in the rescaled window, the formulas (6.18)-(6.21) in Mammen and Park (2005) can be used to show that (A.5) is sufficient for the local additive estimator to be stable. In view of Corollary 2, (A.5) is necessary too.

Consider

$$AIC(h,w) = \log(\hat{\sigma}^2) + 2tr(H)/n\,,$$

where $\hat{\sigma}^2 = \frac{1}{n}||\mathbf{Y} - H\mathbf{Y}||^2$, $\mathbf{Y}$ is the column vector of responses on design points with a hat matrix H and $tr(H)$ is the trace of the hat matrix H. Using

$$\log(\hat{\sigma}^2) = \log(\sigma^2) + \frac{\hat{\sigma}^2}{\sigma^2} - 1 + O_p\left((\hat{\sigma}^2 - \sigma^2)^2\right)\,.$$

Studer et al. (2005) defined the Taylor approximation of AIC$-\log(\sigma^2)$ by

$$\mathrm{AIC}_T = \frac{\hat{\sigma}^2}{\sigma^2} - 1 + \frac{2}{n}tr(H)\,. \tag{11}$$

It can be shown that AIC and AIC_T are equivalent for the optimal parameters in Corollary 2. Using the fact that for additive regression functions

$$tr(H) \to K(0)\sum_j 1/h_j\,,$$

(see (6.11) in Mammen and Park 2005), we establish below that PLS and AIC_T are equivalent as long as $\hat{\sigma}^2$ is consistent and $\hat{r}$ is stable.

Proposition 4. *The PLS defined by Mammen and Park (2005) is equivalent to AIC_T defined by Studer et al. (2005).*

A decomposition of AIC_T leads to

Proposition 5.

$$\begin{aligned} &AIC_T - \Big(\frac{1}{n\sigma^2}\varepsilon'\varepsilon - 1\Big) \\ &= \quad \frac{1}{n\sigma^2}||(I-H)\mathbf{r}||^2 + \frac{1}{n\sigma^2}E[||H\varepsilon||^2] + O_p(\frac{h^2+w^4}{\sqrt{n}}) + O_p\big(\frac{1}{n\sqrt{w^{d-1}h}}\big) \end{aligned}$$

The first term on the right hand side of the decomposition of AIC_T is the mean squared bias, whereas the second term is the variance of $\hat{r}_{ladd}$, both divided by σ^2. Thus, smoothing parameter selection based on AIC-type model selection criteria leads to asymptotically optimal bias variance compromise.

Proofs of Propositions 4 and 5 are given in Appendix.

3 Numerical performance

3.1 Simulation studies

We are interested in investigating how the smoothing parameters are related to performance of the estimators of general regression function in terms of conditional MISE. For general multivariate nonparametric regression problem, there are limited simulation studies reported in the literature. For example, Banks et al. (2003) reported comparison results of a broad class of multivariate nonparametric regression techniques. Some additive model simulation studies can be found in Dette et al. (2005) and Martins-Filho and Yang (2006). Here we focus on comparison to local linear and additive estimators as a benchmark on either extremes. Local linear estimator is optimal for general regression function estimation so the comparison to it allows us to assess the behavior for nonadditive regression function estimation. Likewise additive estimator is used to study the behavior for additive regression function estimation. Results are based on Monte-Carlo approximation of MISE.

d=2: A random uniform design on $[-1,1]^2$ and normally distributed residuals $\mathcal{N}(0,\sigma^2)$ were assumed with sample sizes 200, 400, and 1600. Estimators are evaluated at an equidistant output grid of 21×21 points. For fitting the SBE, we used *SBF2* package of *R* developed in conjunction with Studer et al. (2005), which is freely available from www.biostat.uzh.ch/research/software/.

The main factor of consideration in our simulation studies is the regression function, covering a range of additive and nonadditive functions. To illustrate the behavior of the local additive estimator, we first consider the regression function

$$r(\mathbf{x}) = x_1^2 + x_2^2 + \frac{\alpha}{1-\alpha} x_1 x_2 \,, \tag{12}$$

where α controls the amount of nonadditive structure in the function.

Figure 1 about here.

Performance of the local additive estimator is illustrated in Figure 1. Estimation is based on 400 observations with $\alpha = 0.4$ and $\sigma = 0.5$. All estimators used their MISE-optimal smoothing parameters. As expected, the additive estimator (lower right panel) does not capture the nonadditive structure. The local linear estimator (upper right) reveals the diagonal structure but has a quite large bias due to its large MISE-optimal bandwidth ($h = 0.64$). Because of local additive, instead of local linear, approximation of the regression function, the local additive estimator uses more observations ($w = 0.94$), resulting in an improved variance, whereas the bandwidth is smaller ($h = 0.47$), resulting in an improved bias. As a consequence, the local additive estimator inherits the optimal properties in a *local* sense.

For smoothing parameter selection in practice, Figure 2 presents comparison of ASE-optimal parameters to AIC_C optimal ones for the local additive estimator based on one realization drawn from the same design used in Figure 1 with $\sigma = 0.5$. The range of smoothing parameters suggested by both criteria largely agrees and we find AIC_C comparable for practical use.

Figure 3 about here.

The effect of nonadditivity α in the regression function on MISE can be seen in Figure 3, on a log scale. MISE (first row) is decomposed into the integrated squared bias (second row) and variance (third row). Different columns correspond to different σs. In each panel, the optimal MISE is plotted as a function of α, with an individual optimal choice of smoothing parameters found in the above simulations. Solid line is for local additive estimator, dashed line for local linear estimator and dotted line for additive estimator. As is expected, the regression function has little effect on the

local linear estimator but had a dramatic impact on the additive estimator because of growing nonadditivity. Local additive estimator shows relatively robust performance, adapting the best of the former estimators.

MISE behavior for other regression functions is summarized in Table 1. Regression functions used are additive peaks

$$r(\mathbf{x}) = \frac{1}{2}\sum_{k=1}^{2}\left(0.3\exp(-2(x_k+0.5)^2) + 0.7\exp(-4(x_k-0.5)^2) + 0.5\exp(-\frac{x_k^2}{2})\right),$$

superposed peaks

$$r(\mathbf{x}) = 0.3\exp(-2\|\mathbf{x}+0.5\|^2) + 0.7\exp(-4\|\mathbf{x}-0.5\|^2) + 0.5\exp(-\frac{\|\mathbf{x}\|^2}{2}),$$

and periodic nonadditive function

$$r(\mathbf{x}) = \cos(\pi||\mathbf{x}||).$$

MISE-values are multipled by 1000. MISE-optimal smoothing parameters are also supplied, with MISE ratios.

Table 1 about here.

We considered variants of these scenarios for other regression functions and design densities such as fixed uniform, fixed uniform jittered, linearly skewed one and linearly skewed jittered designs and observed similar phenomena stable across designs considered. More simulation results are found in Park and Seifert (2008). There, one can also find simulations for $d = 3$. Because of dimensionality, the candidate regions of smoothing parameters are narrower than those for $d = 2$, but the behavior of the estimators is similar and thus the same conclusions apply.

d=10: For higher–dimensional case, we considered the regression function

$$r(\mathbf{x}) = x_1^2 + \alpha x_1\big(\sum_{j=2}^{10} x_j\big) \quad \alpha = 0, 0.5, \text{ or } 1, \tag{13}$$

with 2000 observations on a random uniform design and $\sigma = 0.2$. Local estimation in 10 dimensions calls for boundary correction. Otherwise, the expected number of observations in a corner would be $w^{10}n$ compared to $1024w^{10}n$ in the center. To illustrate the behavior of local additive estimator using an additive estimator other than the SBE, we used the function *gam* in the *mgcv* package of *R*. Although optimality of the penalized splines used there is not known, the idea of local additive estimator can be easily applied. Moreover, *gam* has computational advantages; implementation with *gam* particularly facilitates selection of smoothing parameter using generalized cross validation (GCV). Unconditional MASE was approximated with 20 runs of simulation. To reduce computational burden, estimators are evaluated at 50 design points randomly chosen at each simulation. The resulting relative standard error of MASE estimators is about 3-5%.

Figure 4 about here.

Figure 4 shows performance of estimators for three different values of α. Dashed line is for local linear estimator, solid line for local additive estimator. The letter "a" at the end of solid line represents additive estimator. The x-axis represents smoothing parameter; for local linear estimator, it is the bandwidth h and for local additive estimator, it is w, and the GCV-optimal value of h given w was chosen internally by *gam*. Performance of local linear estimator does not depend on the regression function, while local additive estimator adapts to additivity, exhibiting lower curves as the panel moves to the right. We can conclude that overall performance of local additive estimator exceeds that of others, adapting to nonadditivity.

In summary, we have observed that when the regression function is additive or close to additive the local additive estimator is compatible to the additive estimator, and when the regression function is nonadditive it mimics the local linear estimator whenever possible. We also have noticed that the lowest possible bandwidth that local

additive estimator could exploit is limited by the number of observations required to obtain a stable estimator for every output point. A boundary correction sometimes helps to stabilize an estimator but it works differently for different estimators and thus we decided not to include it except for $d = 10$.

3.2 Real data example

We use the ozone dataset from the *R* package (Section 10.3, Hastie and Tibshirani (1990)) to make comparison to the previous analysis. With nine predictors, an additive regression model would be a natural choice. When a new approach which can deal with nonadditive structure is applied, the model can be further refined or simplified. Studer et al. (2005) pointed out that the additive model with nine predictors is almost equivalent, in terms of adjusted R^2, to an additive model with a subset of predictors, allowing bivariate interaction terms. They applied penalized regression approach to uncover behavior of the bivariate interaction, noting serious departure from additive model assumption.

To make it comparable, we adopt the same framework as Studer et al. (2005), where the dependent variable is defined as the logarithm of the upland ozone concentration (up03) and three predictors, humidity (hmdt), inversion base height (ibtp), and calendar day (day) are chosen which maximize adjusted R^2 among fitted additive models with bivariate interaction terms with 16 degrees of freedom each, using *gam* in *R* package *mgcv*. Then the three variables were scaled to [0,1]. As noted in the previous analysis, one observation (92) that contains excessive value of wind speed was removed prior to the analysis.

We consider local additive model and additive with bivariate interaction model for comparison. The additive with interaction model was fitted using *gam* with internally chosen optimal smoothing parameters. To fit the local additive model based on the

SBE, univariate bandwidths h_1, h_2 and h_3 are initially chosen to have four degrees of freedom each as in Studer et al. (2005). These are shown to lie close together with mean $h = \sqrt[3]{h_1 h_2 h_3} = 0.237$. Bandwidths for the local additive estimator are set to be (ch_1, ch_2, ch_3). Parameters c and w are then selected based on AIC_C.

Figure 5 about here.

These estimators are compared in Figure 5. For reference, we also reproduced the local linear estimator from Studer et al. (2005). The univariate components on the top show similar trend, although the local linear estimates show occasional kinks and the additive with interaction models tends to smooth out quickly, especially for hmdt. The bottom row shows the largest bivariate interaction, that between ibtp and hmdt, for each estimator. We see that in both terms the local additive estimator provides a good compromise. Interested readers are referred to Section 5.3 and Figure 5 in Studer et al. (2005) for further comparison and issues with regularisation.

References

[1] Banks, D. L., Olszewski, R. T., and Maxion, R. A. (2003). Comparing methods for multivariate nonparametric regression. *Computations in Statistics*, **32**, 541-571.

[2] Breiman, L. (1993). Fitting additive models to regression data: Diagnostics and alternative views. *Computational Statistics and Data Analysis*, **15**, 13-46.

[3] Deaton, A. and Muellbauer, J. (1980). Econometrics and consumer behaviour. *Cambridge University Press: Cambridge.*

[4] Dette, H., Von Lieres, Carsten, and Sperlich, S. (2005). A comparison of different nonparametric methods for inference on additives. *Nonparametric Statistics*, **17**, 57-81.

[5] Fan, J. (1993). Local linear regression smoothers and their minimax efficiency. *Annals of Statistics*, **21**, 196-216.

[6] Fan, J., Gasser, T., Gijbels, I., Brockmann, M. and Engel, J. (1997). Local polynomial regression: optimal kernels and asymptotic minimax efficiency. *Annals of the Institute of Statistical Mathematics*, **49**, 79-99.

[7] Hastie, T. and Tibshirani, R. (1990). Generalized additive models. *Chapman and Hall, London.*

[8] Hurvich, C., Simonoff, J. and Tsai, C. (1998). Smoothing parameter selection in nonparametric regression using an improved Akaike information criterion. *Journal of Royal Statistical Society, B*, **60**, 271-293.

[9] Linton, O. and Nielsen, J. (1995). A kernel method of estimating structured nonparametric regression based on marginal integration. *Biometrika*, **82**, 93-100.

[10] Mammen, E., Linton, O. and Nielsen, J. (1999). The existence and asymptotic properties of a backfitting projection algorithm under weak conditions. *Annals of Statistics*, **27**, 1443-1490.

[11] Mammen, E., Marron, J. S., Turlach, B. A. and Wand, M. P. (2001). A general projection framework for constrained smoothing. *Statistical Science*, **16** (3), 232-248.

[12] Mammen, E. and Park, B. U. (2005). Bandwidth selection for smooth backfitting in additive models. *Annals of Statistics*, **33**(3), 1260-1294.

[13] Martins-Filho, C. and Yang, K. (2006). Finite sample performance of kernel-based regression methods for nonparametric additive models under common bandwidth selection criterion. ,

[14] Nielsen, J. P. and Sperlich, S. (2005). Smooth backfitting in practice. *Journal of the Royal Statistical Society, B*, **60**, 43-61.

[15] Opsomer, J. (2000). Asymptotic properties of backfitting estimators. *Journal of Multivariate Analysis*, **73**, 166–179.

[16] Opsomer, J., and Ruppert, D. (1997). Fitting a bivariate additive model by local polynomial regression. *Annals of Statistics*, **25**, 186–212.

[17] Opsomer, J., and Ruppert, D. (1998). A fully automated bandwidth selection method for fitting additive models. *Journal of American Statistical Association*, **93**, 605–619.

[18] Park, J. and Seifert, B. (2008). On properties of local additive estimation based on the smooth backfitting estimator. *technical report*, available on `http://www.maths.lancs.ac.uk/~parkj1/paper/locaddSBE.pdf`

[19] Stone, C. J. (1980). Optimal rates of convergence for nonparametric estimators. *Annals of Statistics*, **8**, 1348-1360.

[20] Stone, C. J. (1982). Optimal global rates of convergence for nonparametric regression. *Annals of Statistics*, **10**, 1040-1053.

[21] Stone, C. J. (1985). Additive regression and other nonparametric models. *Annals of Statistics*, **13**, 689-705.

[22] Studer, M., Seifert, B. and Gasser, T. (2005). Nonparametric regression penalizing deviations from additivity. *Annals of Statistics*, **33**, 1295-1329.

[23] Wood, S. N. (2000). Modelling and smoothing parameter estimation with multiple quadratic penalties. *Journal of Royal Statistical Society, B*, **62**, 413–428.

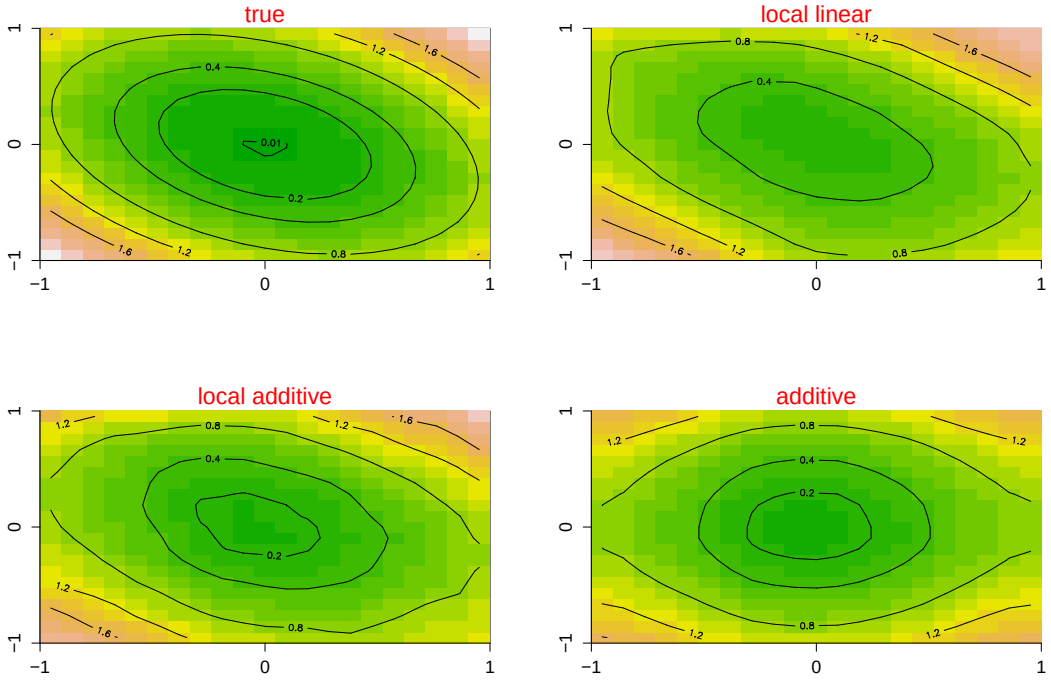

Figure 1: Contour plot of regression function (12) and estimators. Parameters are chosen to be MISE-optimal from simulation with $\alpha = 0.4$ and $\sigma = 0.5$. Additive estimator fails to capture nonadditive structure. While local linear estimator and local additive estimator show compatible performance, local additive estimator incurs smaller bias at the center due to smaller bandwidth.

Additive peaks			
σ	local linear (h_{opt})	local additive (h_{opt}, w_{opt})	additive (h_{opt})
0.1	3.9=315% (h=0.260)	1.2=100% (h=0.123, w=0.870)	1.3=107% (h=0.143)
0.5	22.1=136% (h=0.473)	16.2=100% (h=0.350, w=0.988)	15.4=95% (h=0.350)
1.0	39.6=111% (h=1.000)	35.6=100% (h=0.741, w=0.933)	32.5=91% (h=0.861)
Superposed peaks			
σ	local linear (h_{opt})	local additive (h_{opt}, w_{opt})	additive (h_{opt})
0.1	2.6=117% (h=0.260)	2.2=100% (h=0.193, w=0.242)	7.0=311% (h=0.350)
0.5	14.9=124% (h=0.741)	12.0=100% (h=0.638, w=0.716)	13.3=110% (h=0.638)
1.0	30.9=123% (h=1.000)	25.1=100% (h=0.741, w=0.741)	24.4=97% (h=0.861)
Periodic nonadditive			
σ	local linear (h_{opt})	local additive (h_{opt}, w_{opt})	additive (h_{opt})
0.1	4.8=130% (h=0.260)	3.7=100% (h=0.193, w=0.242)	96.8=2611% (h=0.166)
0.5	32.7=97% (h=0.350)	33.6=100% (h=0.260, w=0.260)	111.7=333% (h=0.302)
1.0	85.7=91% (h=0.473)	93.9=100% (h=0.407, w=0.407)	139.2=148% (h=0.473)

Table 1: Comparison of MISE performance based on 400 observations at different standard deviations–optimal parameters are given in the parentheses. Outperformance of local additive estimator is consequence of smaller h than that for local linear estimator and smaller additive region ($w < 1$) than that for additive estimator.

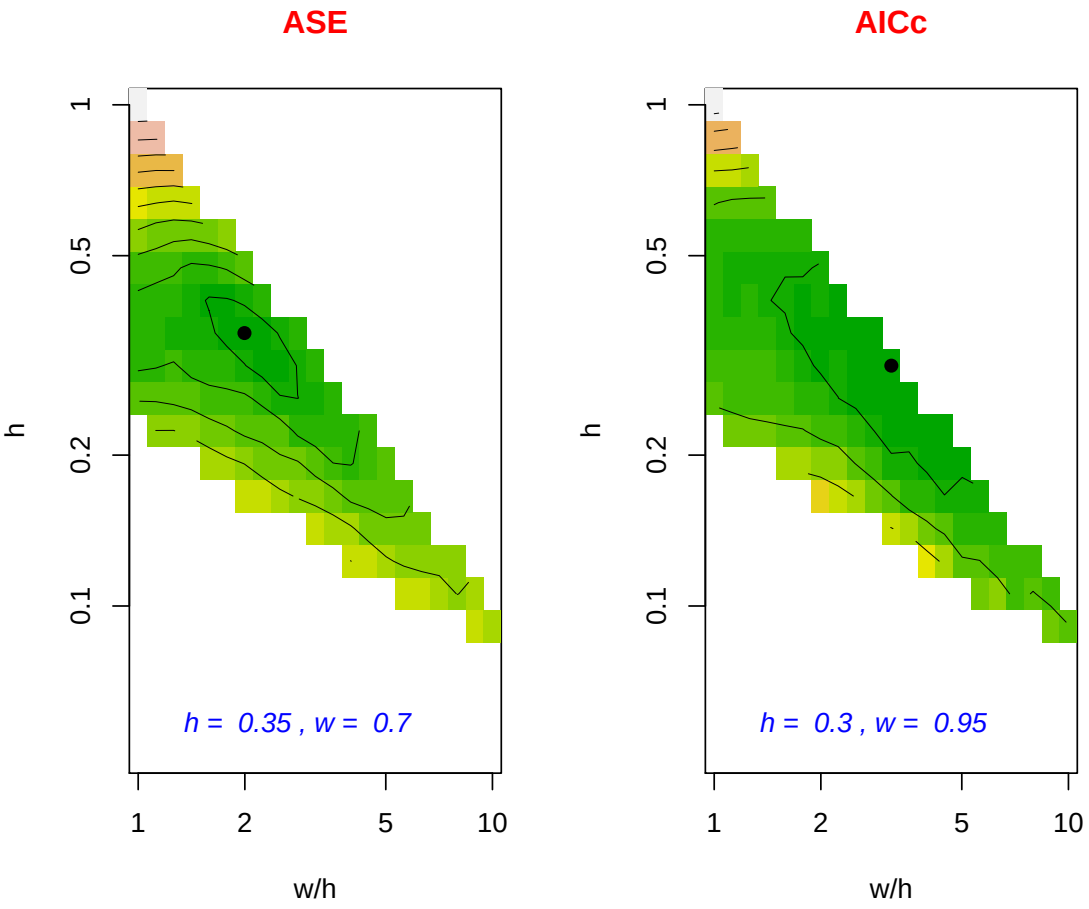

Figure 2: ASE and parameter selection by AIC_C for regression function (12) and design used in Figure 1 with $\sigma = 0.5$.

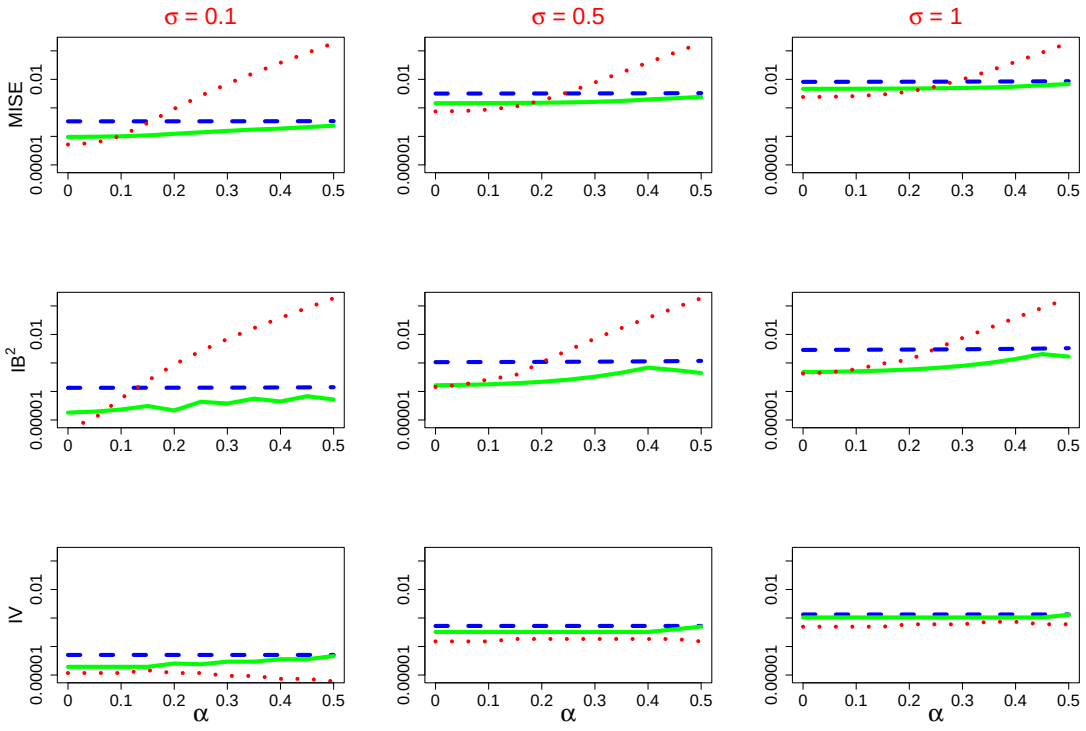

Figure 3: Effect of nonadditive regression function on the MISE performance for MISE-optimal parameters. MISE (first row), integrated squared bias (second row) and variance (third row) as functions of α in (12) is plotted on a log scale for increasing σ. Local linear estimator (dashed line) is not affected but additive estimator (dotted line) dramatically deteriorates. Local additive estimator (solid line) shows relatively robust performance.

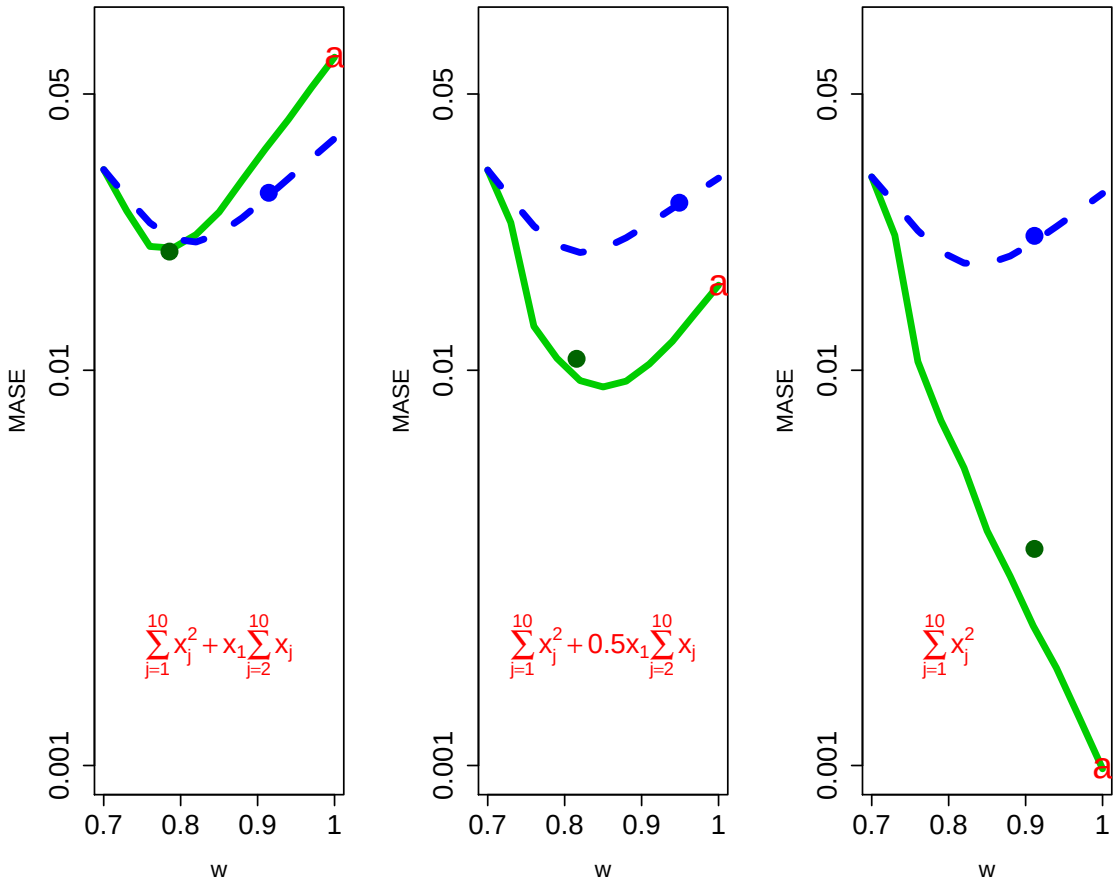

Figure 4: Comparison of unconditional MASE performance of a 10–dimensional regression function (13) for local linear estimator (- -), local additive estimator (–) and additive estimator ("a"). x-axis represents bandwidths for local linear estimator and w for local additive estimator with an internal choice by *gam* of h at given w. Dots and "a" show mean MASE at GCV-optimal smoothing parameters vs. mean GCV-optimal w.

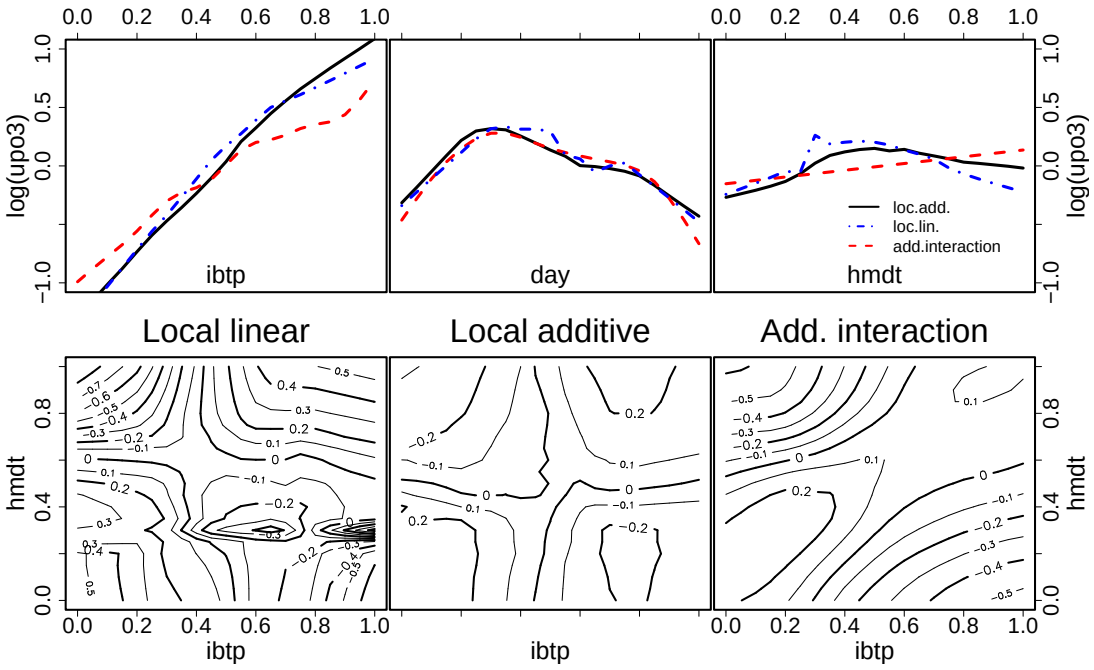

Figure 5: Comparison of local linear, local additive, and additive spline with interaction estimators. Top row shows univariate additive components and bottom row shows bivariate components of ibtp and hmdt for each estimator.

Appendix

Proof of Proposition 4 We use the following fact for additive regression functions

$$tr(H) \to K(0) \sum_j 1/h_j := tr(H)_\infty \,,$$

which can be deduced from (6.11a) or (6.11) in Mammen and Park (2005). Thus, PLS (p. 1269, Mammen and Park 2005) is defined for additive regression functions as

$$PLS = \hat{\sigma}^2 \left(1 + 2\frac{tr(H)_\infty}{n} \right) .$$

In this form, it can be generalized to nonadditive functions. Firstly from the definition of AIC_T, it can be written as

$$AIC_T + 1 = \frac{1}{\sigma^2} \left(PLS + 2(\sigma^2 - \hat{\sigma}^2)\frac{tr(H)}{n} + 2\hat{\sigma}^2 \frac{tr(H) - tr(H)_\infty}{n} \right) .$$

Then observe that

$$2(\sigma^2 - \hat{\sigma}^2)\frac{tr(H)}{n} + 2\hat{\sigma}^2 \frac{tr(H) - tr(H)_\infty}{n} = o\left(\frac{tr(H)}{n} \right) .$$

Therefore, it follows that

$$AIC_T + 1 = \frac{1}{\sigma^2} \left(PLS + o\left(\frac{tr(H)}{n} \right) \right) ,$$

as long as $\hat{\sigma}^2$ is consistent and $\hat{r}$ is stable. □

Proof of Proposition 5 AIC_T can be written as

$$\begin{aligned} \mathrm{AIC}_T &= \frac{1}{n\sigma^2}\mathbf{r}'(I-H)'(I-H)\mathbf{r} + \frac{1}{n\sigma^2}\varepsilon'(I-H)'(I-H)\varepsilon \\ &\quad + \frac{1}{n\sigma^2} 2\varepsilon'(I-H)'(I-H)\mathbf{r} - 1 + \frac{2tr(H)}{n} \,. \end{aligned}$$

Observe that

$$\varepsilon'(I-H)'(I-H)\varepsilon = \varepsilon'\varepsilon - 2E[\varepsilon' H\varepsilon] + O_p(\sqrt{V[\varepsilon' H\varepsilon]}) + E[\varepsilon' H'H\varepsilon] + O_p(\sqrt{V[\varepsilon' H'H\varepsilon]}) \,.$$

Since $E[\varepsilon' H \varepsilon] = tr(H)\sigma^2$, we have

$$\begin{aligned}
&\mathrm{AIC}_T - \Big(\frac{1}{n\sigma^2}\varepsilon'\varepsilon - 1\Big) \\
&= \frac{1}{n\sigma^2}||(I-H)\mathbf{r}||^2 + \frac{1}{n\sigma^2}E[\varepsilon' H' H\varepsilon] + \frac{1}{n\sigma^2}2\varepsilon'(I-H)'(I-H)\mathbf{r} \\
&\quad + \frac{1}{n\sigma^2}O_p(\sqrt{V[\varepsilon' H\varepsilon]}) + \frac{1}{n\sigma^2}O_p(\sqrt{V[\varepsilon' H' H\varepsilon]})\Big)\,.
\end{aligned}$$

The rest follows from a series of lemmas below.

Lemma 1.

$$tr((H'H)'(H'H)) = O\big(tr(H'H)\big) = O\big(1/(w^{d-1}h)\big)$$

Proof: Denote by H_i the hat matrix of the additive estimator used for local additive estimation at $x_0 = X_i$. Then, inflating the matrix to an $n \times n$ matrix, the ith line of H is the ith line of $H_i := H_{i,i}$. Now, considering the form of the estimator $\hat{r}_i = \hat{r}_0 + \hat{r}_1 + \ldots + \hat{r}_d$, where all components are oracle, we have

$$H_{i,j} = \begin{cases} O(\frac{1}{\tilde{n}\tilde{h}}) & \text{if for all } k: |X_{i_k} - X_{j_k}| \le w \text{ and for some } k: |X_{i_k} - X_{j_k}| \le h \\ O(\frac{1}{\tilde{n}}) & \text{if for all } k: |X_{i_k} - X_{j_k}| \le w \text{ and for all } k: |X_{i_k} - X_{j_k}| > h \\ 0 & \text{otherwise} \end{cases}$$

Note, that these $O()$s are uniform over X because of (A.5), using Gao (2003) as in Studer et al. (2005). Let's first look at $tr(H'H)$. We have

$$H'_{i,i}H_{i,i} = O\left(O(\frac{1}{\tilde{n}\tilde{h}})^2 O(\tilde{n}\tilde{h}) + O(\frac{1}{\tilde{n}})^2 O(\tilde{n})\right) = O\left(\frac{1}{\tilde{n}\tilde{h}}\right) = O\left(\frac{1}{nw^{d-1}h}\right).$$

Consequently,

$$tr(H'H) = O\left(\frac{1}{w^{d-1}h}\right).$$

Now, look at the general elements of $H'H$. With a slight abuse of notation,

$$\begin{aligned}
H'_{i,i}H_{j,j} &= O\left(O(\frac{1}{\tilde{n}\tilde{h}})^2O(\tilde{n}\tilde{h}) + O(\frac{1}{\tilde{n}\tilde{h}})O(\frac{1}{\tilde{n}})O(\tilde{n}\tilde{h}) + O(\frac{1}{\tilde{n}})^2O(\tilde{n})\right) \\
&\quad \text{if for all } k : (X_{i_k} \pm w) \cap (X_{j_k} \pm w) \neq \varnothing \\
&\quad \text{and for some } k : (X_{i_k} \pm h) \cap (X_{j_k} \pm h) \neq \varnothing \\
&\quad O\left(O(\frac{1}{\tilde{n}\tilde{h}})O(\frac{1}{\tilde{n}})O(\tilde{n}\tilde{h}) + O(\frac{1}{\tilde{n}})^2O(\tilde{n})\right) \\
&\quad \text{if for all } k : (X_{i_k} \pm w) \cap (X_{j_k} \pm w) \neq \varnothing \\
&\quad \text{and for some } k : (X_{i_k} \pm w) \cap (X_{j_k} \pm h) \neq \varnothing \\
&\quad \text{or } (X_{i_k} \pm h) \cap (X_{j_k} \pm w) \neq \varnothing \\
&\quad O\left(O(\frac{1}{\tilde{n}})\right) \qquad \text{if for all } k : (X_{i_k} \pm w) \cap (X_{j_k} \pm w) \neq \varnothing \\
&\quad \text{and } (X_{i_k} \pm w) \cap (X_{j_k} \pm h) = \varnothing \text{ and } (X_{i_k} \pm h) \cap (X_{j_k} \pm w) = \varnothing \\
&\quad 0 \quad \text{otherwise}.
\end{aligned}$$

Finally,

$$\begin{aligned}
H'_{i,i}H_{j,j} &= O(\frac{1}{\tilde{n}\tilde{h}}) \text{ if for all } k : (X_{i_k} \pm w) \cap (X_{j_k} \pm w) \neq \varnothing \\
&\qquad \text{and for some } k : (X_{i_k} \pm h) \cap (X_{j_k} \pm h) \neq \varnothing \\
&\quad O(\frac{1}{\tilde{n}}) \text{ if for all } k : (X_{i_k} \pm w) \cap (X_{j_k} \pm w) \neq \varnothing \\
&\qquad \text{and for all } k : (X_{i_k} \pm h) \cap (X_{j_k} \pm h) = \varnothing \\
&\quad 0 \text{ otherwise}.
\end{aligned}$$

Thus, $H'H$ has the same structure as H, of course with different constants and larger non-zero regions, but all of the same order. Therefore,

$$tr(H'HH'H) = O(tr(H'H)) = O\left(\frac{1}{w^{d-1}h}\right).$$

Lemma 2.

$$\frac{1}{n\sigma^2}\varepsilon'(I-H)'(I-H)\mathbf{r} = \frac{1}{n\sigma^2} < (I-H)\varepsilon, (I-H)\mathbf{r} >= O_p(\frac{h^2+w^4}{\sqrt{n}})$$

Proof: First note that $(I-H)\mathbf{r} = O((h^2+w^4)\mathbf{1})$.

$$\begin{aligned}
\frac{1}{n\sigma^2} &< (I-H)\varepsilon, (I-H)\mathbf{r} > \\
&\leq \frac{1}{n\sigma^2}||(I-H)\varepsilon||||(I-H)\mathbf{r}|| \\
&\leq \frac{1}{n\sigma^2}O_p(\sqrt{n})O((h^2+w^4)\mathbf{1}) = O_p\big(\frac{h^2+w^4}{\sqrt{n}}\big)
\end{aligned}$$

where the last inequality follows from

$$||(I-H)\varepsilon||^2 = O_p(tr(V[||(I-H)\varepsilon||^2)) = O_p(tr((I-H)(I-H)'\sigma^2) = O_p(n)\,.$$

Lemma 3.

$$V[\varepsilon' H\varepsilon] = V[<\varepsilon, H\varepsilon>] = O(E[||H\varepsilon||^2])$$

If $tr((H'H)(H'H)) = O(tr(H'H)$*, then*

$$V[\varepsilon' H'H\varepsilon] = V[||H\varepsilon||^2] = O(E[||H\varepsilon||^2])$$

Proof: Similar to the proof of lemma 5 in the appendix of Studer et al. (2005), we use the following fact: For symmetric matrices B and C and $E[\varepsilon^4] = (3+\kappa)\sigma^4$,

$$Cov(\varepsilon' B\varepsilon, \varepsilon' C\varepsilon) = 2\sigma^4 tr(BC) + \kappa\sigma^4 tr(B \cdot diag(C))$$

Putting $B = C = \frac{1}{2}(H+H')$ gives

$$\begin{aligned}
V[\frac{1}{2}\varepsilon'(H+H')\varepsilon] &= 2\sigma^4(\frac{1}{4}tr(HH+2H'H+H'H') + \kappa\sigma^4 tr(\frac{1}{4}(H'+H)diag(H+H')) \\
V[\varepsilon' H\varepsilon] &= \sigma^4(tr(HH+H'H) + \kappa\, tr(diag(H)^2) \\
&\leq \sigma^4(tr(HH) + tr(H'H) + |\kappa| tr(H'H)
\end{aligned}$$

Using the equivalence of the trace to Hilbert-Schmidt norm,

$$||H||_{HS}^2 = tr(H'H)$$

it follows that

$$tr(HH) =< H', H >_{HS} \leq ||H'||_{HS}||H||_{HS} = ||H||^2_{HS} = tr(H'H)\,.$$

Hence,

$$\begin{aligned} V[\varepsilon' H\varepsilon] &\leq \sigma^4(2tr(H'H) + |\kappa|tr(H'H)) \\ &= \sigma^4(2+|\kappa|)tr(H'H) \\ &= \sigma^2(2+|\kappa|)E[||H\varepsilon||^2] \end{aligned}$$

Thus, $V[< \varepsilon, H\varepsilon >] = O(E[||H\varepsilon||^2])$. Moreover, replacing H by $H'H$ in the above leads to

$$V[\varepsilon' H'H\varepsilon] \leq \sigma^2(2+|\kappa|)tr((H'H)'(H'H))\,.$$

Hence, if $tr((H'H)'(H'H)) = O(tr(H'H))$, then $V[||H\varepsilon||^2] = O(tr(H'H)) = O(E[||H\varepsilon||^2])$. □