On properties of local additive estimation based on the smooth backfitting estimator

Juhyun Park and Burkhardt Seifert
Lancaster University and University of Zürich

December 8, 2008

Abstract

We study properties of local additive estimation based on the smooth backfitting estimator by Mammen, Linton and Nielsen (1999). The local additive estimator defined as a restricted additive estimator and thus inherits locally properties of additive estimator. Our asymptotic analysis shows that this provides a new class of nonparametric regression estimators for high dimensional problem. Simulation studies are used to assess finite sample performance.

1 Local additive estimation

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

$$r(\mathbf{x}) = r_0 + r_1(x_1) + \cdots + r_d(x_d)\,. \tag{1.1}$$

This is a global assumption on the shape of the regression function.

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 $\hat{\mathbf{r}}_{ladd}(\mathbf{x}_0)$.

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 $[\mathbf{x}_0 \pm \mathbf{w}] = \{\mathbf{X}_i, \mathbf{X}_i \in [\mathbf{x}_0 - \mathbf{w}, \mathbf{x}_0 + \mathbf{w}]\}$. 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 $\hat{r}_{add}$. The SBE by Mammen et al. (1999) is known to be oracle optimal under general conditions and will be used as basis for local additive estimator in this report.

Consider a vector of functions $\mathbf{r}(\mathbf{x}) = (r^0(\mathbf{x}), \cdots, r^d(\mathbf{x}))$, where r^0 is additive and r^j depends only on $x_j, j = 1, \cdots, d$. In view of the local linear estimation, the first function $r^0(\mathbf{x})$ is the intercept and the others are slopes. The SBE is defined as the minimizer of

$$\int_{[-1,1]^d} \frac{1}{n} \sum_{i=1}^{n} \Big[Y_i - r^0(\mathbf{x}) - \sum_{j=1}^{d} r^j(x_j) \frac{X_{ij} - x_j}{h_j} \Big]^2 K_h(\mathbf{X}_i, \mathbf{x})\, d\mathbf{x}\,,$$

where $K_h(\mathbf{X}_i, \mathbf{x})$ is the kernel weight of the observation $(\mathbf{X}_i, Y_i)$ for the output point $\mathbf{x}$. Write $\hat{r}_{add}$ for the solution. Note that this is a *global* estimator, the additivity holding the whole support region.

The local additive estimator at $\mathbf{x}_0$, based on the SBE, is defined as a minimizer of the local norm with respect to $\mathbf{w} \in R^d$ of

$$\int_{\mathbf{x}_0 - \mathbf{w}}^{\mathbf{x}_0 + \mathbf{w}} \frac{1}{n} \sum_{i=1}^{n} \Big[Y_i - r^0(\mathbf{x}) - \sum_{j=1}^{d} r^j(x_j) \frac{X_{i,j} - x_j}{h_j} \Big]^2 \tilde{K}_h(\mathbf{X}_i, \mathbf{x})\, d\mathbf{x}\,, \tag{1.2}$$

where $\tilde{K}$ is a rescaled version of K, defined as $\tilde{K}_h(\mathbf{u}, \mathbf{v}) = K_h(\mathbf{u} - \mathbf{v}) / \int_{\mathbf{x}_0 - \mathbf{w}}^{\mathbf{x}_0 + \mathbf{w}} K_h(\mathbf{u} - \mathbf{v})\, d\mathbf{v}$. The solution to the minimization is denoted by $\hat{r}_{ladd}$. The local additive estimator at $\mathbf{x}_0$ is $\hat{r}_{ladd}(\mathbf{x}_0)$. Observe that the righthand side of (1.2) is equivalent to

$$\int_{[-1,1]^d} \frac{1}{\tilde{n}} \sum_{i=1}^{\tilde{n}} \Big[Y_i - \tilde{r}^0(\mathbf{u}) - \sum_{j=1}^{d} \tilde{r}^j(u_j) \frac{U_{i,j} - u_j}{\tilde{h}_j} \Big]^2 \mathbf{K}_{\tilde{h}}(\mathbf{U}_i, \mathbf{u})\, d\mathbf{u}\,,$$

where $\mathbf{U}_i$ and $\tilde{r}$ are given in (2) and (4), and

$$K_{\tilde{h}}(\mathbf{u}, \mathbf{v}) = \frac{K_{\tilde{h}}(\mathbf{u} - \mathbf{v})}{\int_{[-1,1]^d} K_{\tilde{h}}(\mathbf{u} - \mathbf{v})\, d\mathbf{v}}\,.$$

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

1.1 Preliminaries

Throughout the article, we will assume that

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

(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 $\tilde{n} \to \infty$.

The special case of uniform design will be separately dealt with later in this section. Denote the number of observations $\mathbf{X}_i$ in $[\mathbf{x}_0 \pm \mathbf{w}]$ by $\tilde{n}$ with

$$E[\tilde{n}] = n \int_{[\mathbf{x}_0 \pm \mathbf{w}]} f(\mathbf{x})\, d\mathbf{x} = n f(x_0)(2w)^d + O(nw^{d+3}) = O(nw^d)\,.$$

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

$$\mathbf{U} = \frac{\mathbf{X} - \mathbf{x}_0}{w} \tag{1.3}$$

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

$$\begin{aligned} \tilde{f}(\mathbf{u}) &= f(\mathbf{x}_0 + w\mathbf{u}) / \int_{[-1,1]^d} f(\mathbf{x}_0 + w\mathbf{u})\, d\mathbf{u} \\ &= \frac{f(\mathbf{x}_0 + w\mathbf{u})}{2^d f(\mathbf{x}_0)} + O(w^2)\,. \end{aligned} \tag{1.4}$$

The true regression function is substituted with

$$\tilde{r}(\mathbf{u}) = r(\mathbf{x}_0 + w\mathbf{u})\,. \tag{1.5}$$

In particular, $r(\mathbf{x}_0) = \tilde{r}(\mathbf{0})$. The transformed bandwidth becomes

$$\tilde{h}_j = h_j / w\,. \tag{1.6}$$

Then it can be shown using (1.8) below that the normal equations for the local additive estimator may be written as

$$\tilde{\mathcal{S}}_{add} \tilde{\mathbf{r}}_{add} = \mathcal{P}_{add} \tilde{\mathbf{r}}_L\,,$$

where appropriate transformations $\tilde{\mathcal{S}}_{add}$ and $\tilde{r}_L$ and $\mathcal{P}_{add}$ are given in Section 1.2. For completeness the derivation of the normal equation for the SBE is also reviewed there. Convergence of the operator $\tilde{\mathcal{S}}_{add}$ is studied in Section 1.3.

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 and MISE for the conditional expectation, bias, variance, mean squared error and integrated 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.

1.2 Normal equations for the SBE

Following Mammen et al. (1999), we begin with derivation of the normal equation of SBE on which our analysis is based, with additional notations and definitions.

Consider a Hilbert space $(\mathcal{F}, ||\cdot||_*)$ such that the local linear estimator corresponds to a projection of the response $\mathbf{Y}$ to some subspace $\mathcal{F}_{full}\subset\mathcal{F}$. The SBE is interpreted as a projection of $\mathbf{Y}$ to a subspace $\mathcal{F}_{add}\subset\mathcal{F}_{full}$ of additive functions. Formal definitions are given as follows.

Define the vector space of $n(d+1)$ functions

$$\mathcal{F}=\Big\{\mathbf{r}=(r^{i,j}\,,i=1,\cdots,n;\,j=0,\cdots,d)\big|r^{i,j}:[-1,1]^d\to\mathcal{R}\Big\}$$

and define a subspace $\mathcal{F}_{full}$ that restricts $r^{i,j}$ to $r^{0,j}$ as

$$\mathcal{F}_{full}=\Big\{\mathbf{r}=(r^0,\cdots,r^d)\big|r^j:[-1,1]^d\to\mathcal{R},j=0,\cdots,d\Big\}\,.$$

The observations $Y_i, i=1,\cdots,n$ lie in $\mathcal{F}$, coded by $\mathbf{r}_Y$. Define $r_Y^{i,j}(\mathbf{x})=Y_i$ if $j=0$, and 0 otherwise. $\mathbf{r}_Y$ is an equivalent representation of $\mathbf{Y}$ in $\mathcal{F}$. The semi-norm $||\cdot||_*$ on $\mathcal{F}$ is given by

$$||\mathbf{r}||_*^2=\int\frac{1}{n}\sum_{i=1}^n\Big[r^{i,0}(\mathbf{x})+\sum_{j=1}^d r^{i,j}(\mathbf{x})\frac{X_{i,j}-x_j}{h_j}\Big]^2K_h(\mathbf{X}_i,\mathbf{x})\,d\mathbf{x}\,,$$

The local linear estimator $\hat{r}_{ll}$ is defined as

$$\hat{\mathbf{r}}_{ll}=argmin_{\mathbf{r}\in\mathcal{F}_{full}}||\mathbf{r}_Y-\mathbf{r}||_*^2\,.$$

Now consider a subspace of additive functions $\mathcal{F}_{add}$:

$$\mathcal{F}_{add}=\Big\{\mathbf{r}\in\mathcal{F}_{full}\big|r^0(\mathbf{x})\text{ is additive and }r^j(\mathbf{x})\text{ depends only on }x_j\,,j=1,\cdots,d\Big\}\,.$$

The additive estimator $\hat{r}_{add}$ is defined as

$$\hat{\mathbf{r}}_{add} = argmin_{\mathbf{r}\in\mathcal{F}_{add}} ||\mathbf{r}_Y - \mathbf{r}||_*^2 .$$

The projection argument, originally introduced in Mammen et al. (2001), clarified *oracle optimality* of the estimator. Practical aspects of implementation were explored in Nielsen and Sperlich (2005), who introduced the term *smooth backfitting estimator*.

Let us introduce more notation. Define the L^2-norm on $\mathcal{F}$ by

$$||\mathbf{r}||_2^2 = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=0}^{d}\int \left[r^{i,j}(\mathbf{x})\right]^2 d\mathbf{x} .$$

Denote by $\mathcal{P}_{add}$ the $||\cdot||_2$-orthogonal projection from $\mathcal{F}_{full}$ into $\mathcal{F}_{add}$. Define the symmetric, continuous operator $\mathcal{S}_* : \mathcal{F}_{full} \to \mathcal{F}_{full}$ by

$$\begin{pmatrix} r^0(\mathbf{x}) \\ \vdots \\ r^d(\mathbf{x}) \end{pmatrix} \to \begin{pmatrix} S_{0,0}(\mathbf{x}) & \cdots & S_{0,d}(\mathbf{x}) \\ \vdots & & \vdots \\ S_{d,0}(\mathbf{x}) & \cdots & S_{d,d}(\mathbf{x}) \end{pmatrix} \begin{pmatrix} r^0(\mathbf{x}) \\ \vdots \\ r^d(\mathbf{x}) \end{pmatrix},$$

where

$$\begin{aligned} S_{0,0}(\mathbf{x}) &= \frac{1}{n}\sum_{i=1}^{n}\mathbf{K}_h(\mathbf{X}_i,\mathbf{x}) , \\ S_{0,j}(\mathbf{x}) &= S_{j,0}(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^{n}\mathbf{K}_h(\mathbf{X}_i,\mathbf{x})\frac{X_{i,j}-x_j}{h_j} , \\ S_{j,k}(\mathbf{x}) &= \frac{1}{n}\sum_{i=1}^{n}\mathbf{K}_h(\mathbf{X}_i,\mathbf{x})\frac{X_{i,j}-x_j}{h_j}\frac{X_{i,k}-x_k}{h_k} . \end{aligned}$$

By construction it holds that $||\mathbf{r}||_*^2 =< \mathbf{r}, \mathcal{S}_*\mathbf{r} >_2$. The normal equations for the local linear estimator, $\hat{\mathbf{r}}_{ll}$, are given by

$$\mathcal{S}_*\hat{\mathbf{r}}_{ll} = \mathbf{r}_L , \tag{1.7}$$

where

$$\begin{aligned} r_L^0(\mathbf{x}) &= \frac{1}{n}\sum_{i=1}^{n}\mathbf{K}_h(\mathbf{X}_i,\mathbf{x})Y_i , \\ r_L^j(\mathbf{x}) &= \frac{1}{n}\sum_{i=1}^{n}\mathbf{K}_h(\mathbf{X}_i,\mathbf{x})\frac{X_{i,j}-x_j}{h_j}Y_i , \quad j = 1,\cdots,d . \end{aligned}$$

The normal equations for the additive estimator, $\hat{r}_{add}$, may be written as

$$\mathcal{P}_{add}\mathcal{S}_*\mathcal{P}_{add}\hat{\mathbf{r}}_{add} = \mathcal{P}_{add}\mathcal{S}_*\mathbf{r}_{ll} .$$

Define $\mathcal{S}_{add} = \mathcal{P}_{add}\mathcal{S}_*\mathcal{P}_{add}$. Combined with (1.7), the normal equations are reduced to

$$\mathcal{S}_{add}\hat{\mathbf{r}}_{add} = \mathcal{P}_{add}\mathbf{r}_L\,. \tag{1.8}$$

1.3 Convergence of the operator $\tilde{\mathcal{S}}_{add}$

Note that $\tilde{\mathcal{S}}_{add}$ is a function of output point $\mathbf{x}_0$ as well as n. We first state a lemma for the projection operator $\mathcal{P}_{add}$ below.

Lemma 1. *For $\mathbf{r} \in \mathcal{F}_{full}$, let $\mathbf{r}_{add} = \mathcal{P}_{add}\mathbf{r}$. Then*

$$\begin{aligned} r^0_{add}(\mathbf{x}) &= \sum_{j=1}^{d} \frac{1}{2^{d-1}} \int r^0(\mathbf{x})\, d\mathbf{x}_{-j} - \frac{(d-1)}{2^d} \int r^0(\mathbf{x})\, d\mathbf{x} \\ r^j_{add}(\mathbf{x}) &= \frac{1}{2^{d-1}} \int r^j(\mathbf{x})\, d\mathbf{x}_{-j}\,, \end{aligned}$$

where $j = 1, \cdots, d$.

The definition of the projection leads to the above formulas easily and we omit the derivation.

Remark: We haven't imposed any identifiability condition for individual terms. If desired, $r^0_{add}(\mathbf{x})$ can be decomposed into

$$r^0_{add}(\mathbf{x}) = r_0 + \sum_{j=1}^{d} r_j(x_j)\,, \quad \int r_j(x_j) f_j(x_j)\, dx_j = 0\,,$$

where

$$\begin{aligned} r_0 &= \frac{1}{2^{d-1}} \sum_{j=1}^{d} \int r^0(\mathbf{x}) f_j(x_j)\, d\mathbf{x} - \frac{d-1}{2^d} \int r^0(\mathbf{x})\, d\mathbf{x}\,, \\ r_j(x_j) &= \frac{1}{2^{d-1}} \Big(\int r^0(\mathbf{x})\, d\mathbf{x}_{-j} - \int r^0(\mathbf{x}) f_j(x_j)\, d\mathbf{x} \Big)\,. \end{aligned}$$

Lemma 2. *Assume that $w \to 0$. Then $\tilde{\mathcal{S}}_{add}$ converges, with probability tending to one, as $n \to \infty$, to the limiting operator $\tilde{\mathcal{S}}_{add,\infty}$ defined by*

$$\begin{aligned} (\tilde{\mathcal{S}}_{add,\infty}\mathbf{r})^0(\mathbf{u}) &= \frac{1}{2^d} r^0_{add}(\mathbf{u})\,, \\ (\tilde{\mathcal{S}}_{add,\infty}\mathbf{r})^j(\mathbf{u}) &= \frac{1}{2^d} \mu_2(K) r^j_{add}(u_j)\,, \end{aligned}$$

where $\mathbf{r}_{add} = \mathcal{P}_{add}\mathbf{r}$. Moreover, the limiting operator $\tilde{\mathcal{S}}_{add,\infty}$ has a continuous inverse.

1.4 Properties of the local additive estimator

We study properties of the estimator in terms of bias and variance.

Lemma 3. *Variance of the local additive estimator is given as*

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

For the bias, we separate cases (1) when r is additive and (2) when it is general.

Lemma 4. *Suppose that the regression function is additive. Then bias of local additive estimator is the same as that of additive estimator.*

$$B[\hat{r}_{ladd}(\mathbf{x})] = \frac{\mu_2(K)}{2} \sum_{j=1}^{d} h_j^2 r_j''(x_j) + o(h^2)\,.$$

Now consider general regression function. Decompose $B_{ladd}(\mathbf{x})$ as

$$B_{ladd}(\mathbf{x}_0) = B^{(1)}_{ladd}(\mathbf{x}_0) + B^{(2)}_{ladd}(\mathbf{x}_0)\,,$$

where $B^{(1)}_{ladd}(\mathbf{x}_0)$ is associated with additive part of r and $B^{(2)}_{ladd}$ is associated with non–additive part of r. Lemma 4 may be applied to obtain the additive bias $B^{(1)}_{ladd}(\mathbf{x}_0)$. Consider the non–additive part $\tilde{r}^{(2)}(\mathbf{u})$ of $\tilde{r}(\mathbf{u})$. The bias of non–additive part depends crucially on the assumptions made on the regression function as well as the design density. Using Taylor approximation to 2.3, it is enough to focus on $b(\mathbf{u}) = u_j u_k$ only.

Lemma 5. *Suppose that*

$$b(\mathbf{u}) = u_j u_k$$

and let $\hat{b}_{add,w}(\mathbf{u})$ be the additive estimator based on the design density $\tilde{f}(\mathbf{u})$ given in (3).

$$\hat{b}_{add,w}(\mathbf{u}) = \frac{w}{3f_{j,k}(\mathbf{x}_0)}\Big(u_j \frac{\partial}{\partial u_k} f_{j,k}(\mathbf{x}_0) + u_k \frac{\partial}{\partial u_j} f_{j,k}(\mathbf{x}_0)\Big) + O(w^2) + O((nw^{d-1}h)^{-1/2})\,.$$

From (2.4), combined with Lemma 5, the following Corollary is easily derived.

Corollary 1. *Bias of local additive estimator is given as*

$$B[\hat{r}_{ladd}(\mathbf{x}_0)] = \frac{\mu_2(K)}{2} \sum_{j=1}^{d} h_j^2 r''_{j,j}(\mathbf{x}_0) + B^{(2)}_{ladd}(\mathbf{x}_0)\,,$$

where

$$B^{(2)}_{ladd}(\mathbf{x}_0) = O(w^4) + O((nw^{d-5}h)^{-1/2})\,.$$

1.5 Uniform design with very smooth regression function

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

$(A.1')$ 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*

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

Lemma 6. *Assume that $nh^{d-1}w \to 0$ and $(A.1')$ holds. Then, the non–additive bias can be expressed as*

$$B^{(2)}_{ladd}(\mathbf{x}_0) = -\frac{w^4}{4!\cdot 9}\sum_{j\neq k} r''''_{j,j,k,k}(\mathbf{x}_0) + o(w^4) .$$

Proposition 1 shows why higher order smoothness assumption would not help reduce bias further. Moreover, it can be deduced from the proof 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 $h = C_h w^2$ and $\tilde{n} = C_n n w^d$, where C_n is a constant, not dependent of C_h. The smoothing parameter w that minimizes asymptotic MSE is given by*

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

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 $ab < 0$.

2 Simulation studies

In order to assess the performance of the estimator in finite sample, we conducted some simulation studies. We are interested in investigating how the smoothing parameters are related to the performance of the estimators in terms of conditional MISE.

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 behaviour for non-additive regression function estimation. Likewise additive estimator is used to study the behaviour for additive regression function estimation. The main factor of consideration in our simulation studies is the regression function, ranging from additive to non-additive functions. Estimators are evaluated at an equidistant output grid of 21×21 points. Results are based on Monte-Carlo approximation of MISE.

2.1 Regression functions

The following functions are used for $d = 2$.

- Additive peaks (r_1):

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

- Approximately additive peaks (r_2): $r(\mathbf{x}) = r_2\begin{pmatrix} x_1+x_2 \\ -x_1+x_2 \end{pmatrix}$

- Superposed peaks(r_3): $r(\mathbf{x}) = 0.3\exp(-2\|\mathbf{x}+0.5\|^2) + 0.7\exp(-4\|\mathbf{x}-0.5\|^2) + 0.5\exp(-\frac{\|\mathbf{x}\|^2}{2})$

- Mixture of additive and non-additive polynomial (r_4): $r(\mathbf{x}) = \sum_{i=1}^{d} x_i^2 + 0.5x_1\sum_{j=2}^{2} x_j$

- Mixture of additive and periodic non-additive (r_5): $r(\mathbf{x}) = \cos(\pi||\mathbf{x}||) + \sum_{i=1}^{2}\sin(\pi x_i)$

- Periodic non-additive (r_6): $r(\mathbf{x}) = \cos(\pi||\mathbf{x}||)$

2.2 Design

Figure 1 about here.

A random uniform design on $[-1,1]^2$ was assumed with sample sizes 200, 400, and 1600. In addition, fixed uniform, fixed uniform jittered and linearly skewed fixed and jittered designs were considered, as shown in Figure 1.

2.3 MISE calculation

We briefly explain how to approximate MISE on output grids.

For a fixed point $\mathbf{x}$ on the grid, write $\hat{r}(\mathbf{x}) = \sum_i W_i(\mathbf{x})Y_i$ for a linear estimator. The variance can be approximated by Monte-Carlo simulation as follows.

$$V = V[\hat{r}(\mathbf{x})] = E\big[\sum_i W_i(\mathbf{x})\epsilon_i\big]^2 \approx \frac{\sigma^2}{n_{sim}} \sum_{k=1}^{n_{sim}} (\hat{r}^{(k)}(\mathbf{x}))^2 \,,$$

where $\hat{r}^{(k)}$ is an estimator for $r(\cdot) \equiv 0$ and $\sigma^2 = 1$. This formulation is useful to apply to all possible σ and regression functions r without additional computational cost. Results are based on 100 runs of simulation. For the bias, observe that

$$B = B[\hat{r}(\mathbf{x})] = E\big[\sum_i W_i(\mathbf{x})Y_i\big] - r(\mathbf{x}) = \sum_i W_i(\mathbf{x})r(\mathbf{X}_i) - r(\mathbf{x}) \,.$$

Using $\sigma = 0$, this can be calculated by one run of simulation and we obtain the $MSE = V + B^2$ for all values of σ. $MISE(\hat{r}) = \int V[\hat{r}(\mathbf{x})] + B^2[\hat{r}(\mathbf{x})]\, dx$ is then approximated by the mean over the output grid.

2.4 MISE performance

We illustrate with a sample of 400 observations from random uniform design on $[-1, 1]^2$, corresponding to R400 in Figure 1. We first consider a regression function

$$r(\mathbf{x}) = x_1^2 + x_2^2 + \frac{\alpha}{1-\alpha} x_1 x_2 \,, \tag{2.1}$$

where α controls the amount of non–additive structure in the function.

Figure 2 about here.

Figure 2 shows the performance of MISE when $\alpha = 0.4$ for various σs. In each panel, y-axis represents 21 bandwidths h ranging from 0.05 to 1 with an exponential increment. The first column is MISE for local linear estimator and the last column is MISE for additive estimator. The local additive estimator lies in between with different ratios of two smoothing parameters w/h ranging from 1 to 10 on a log scale. That is, for each h, local additive estimators were calculated with increasing w values until it covers the whole region and thus the upper triangular part was not calculated. The white left lower triangle corresponds to parameters where the estimator does not exist for all output grid points. The optimal choice for each estimator is marked by a circle. As σ increases,

optimal bandwidths become larger. Note that the optimal bandwidth for local additive estimator is in general smaller than that of local linear estimator and that the dark area around the optimal choice can be considered competitive. MISE ratios are given in the bottom, demonstrating a gain in MISE for the local additive estimator of more than one third compared to the local linear one and more than 85% compared to the additive estimator.

Tables 1-6 summarize MISE performance for regression functions $r_1 - r_6$ for increasing sample size. Numbers are multiplied by 1000. At each sample size, results are given at 3 different standard deviations. Tables are approximately ordered to follow trend from additive to non-additive structure. Thus, local linear estimator would be favorable for regression functions with strong non-additive structure while additive estimator would be favorable for regression functions close to additive structure. Results from local linear and additive estimators are in accordance with our expectation. Local additive estimator shows robust performance, adapting to the structure of regression function whenever possible. At each regression function, improvements with respect to increasing sample size are illustrated by reduction of MISE. Within each sample size, the amount of deterioration with increasing standard deviations is also illustrated with increasing MISE.

Tables 1-6- about here.

Tables 7-8 present results for additional designs shown in the bottom of Figure 1 for approximately additive (r_2) and non-additive (r_3) regression functions. We see that these are comparable to those from random uniform design.

Tables 7-8- about here.

d=3: We considered the regression function

$$r(\mathbf{x}) = cos(\pi||\mathbf{x}||^2)\,. \tag{2.2}$$

with sample sizes $n = 441, 625, 1089$. To maximize the utility of sample at each direction, we employed latin square designs of 3×7 ($n = 441$), 5×5 ($n = 625$) and 3×11 ($n = 1089$). We considered fixed designs and jittered versions, where a random error was added to each fixed point.

Figure 3 about here.

Figure 3 presents results for jittered design with $n = 1089$. The differences in performance from fixed designs are not dramatic but jittered designs produce slightly more stable estimators. Because

of dimensionality, the candidate regions of smoothing parameters are narrower but the behavior of the estimators is similar and thus the same conclusions apply.

References

[1] Gao, F. (2003). Moderate deviations and large deviations for kernel density estimators. *Journal of Theoretical Probability*, **16**, 401-418.

[2] 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.

[3] 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.

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

[5] Masry, E. (1996). Multivariate local polynomial regression for time series: uniform strong consistency and rates. *Journal of Time Series Analysis*, **17**, 571-599.

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

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

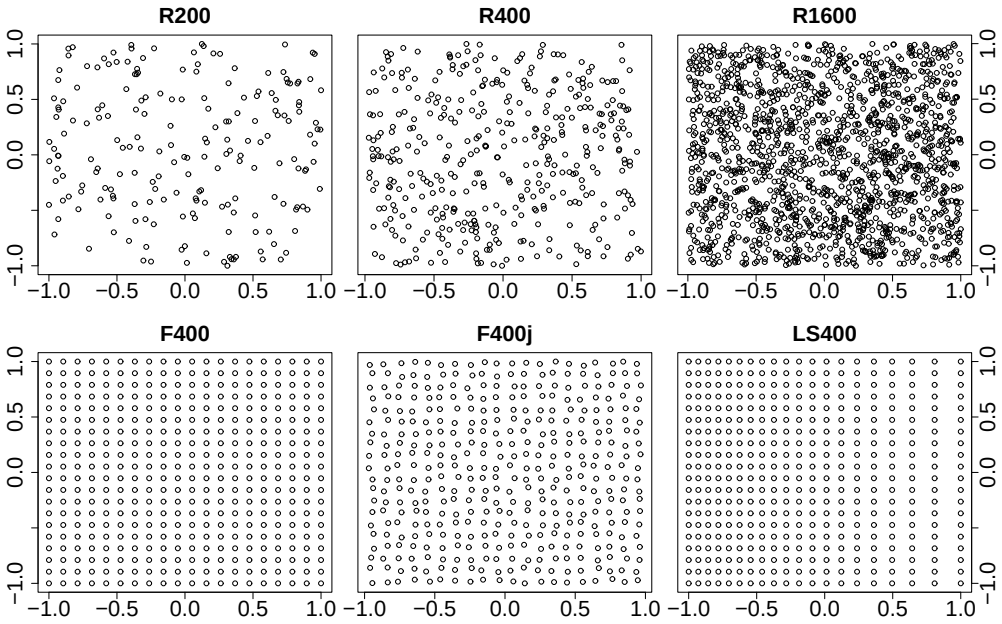

Figure 1: Simulation designs. Top row shows random uniform designs with increasing sample size. Bottom row shows additional fixed uniform, fixed uniform jittered and linearly skewed designs (slope=0.3) with sample size 400.

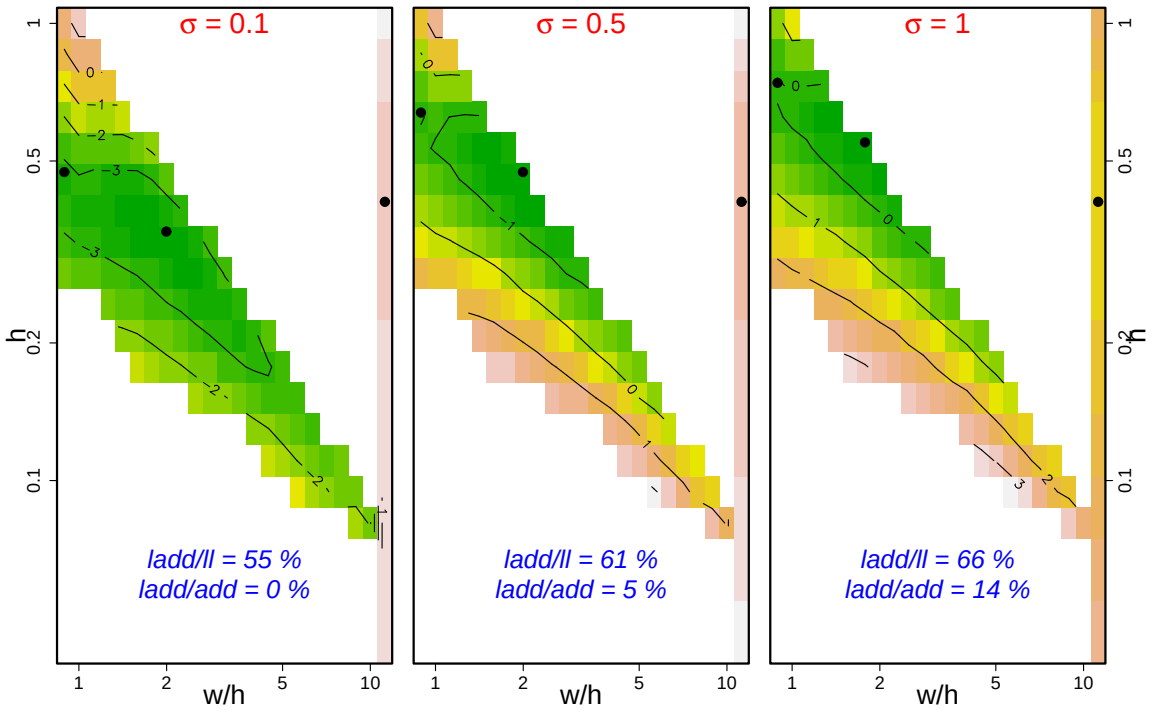

Figure 2: Comparison of MISE performance of estimators for 2–dimensional regression function (2.1) for different values of σ with 400 observations from random uniform design. Each panel contains local linear estimator at the first and additive estimator at the last column. Local additive estimator with increasing ratio of w/h lies in between. y-axis represents bandwidths h and circle is drawn at optimal choice for each estimator. Optimal parameters for local additive estimator moves to south east from optimal h for local linear, by reducing h (lower) and increasing w (right). Contour line indicates wide range of comparable selection. Gains and losses of local additive estimator in comparison to local linear and additive estimator were quantified as MISE ratios.

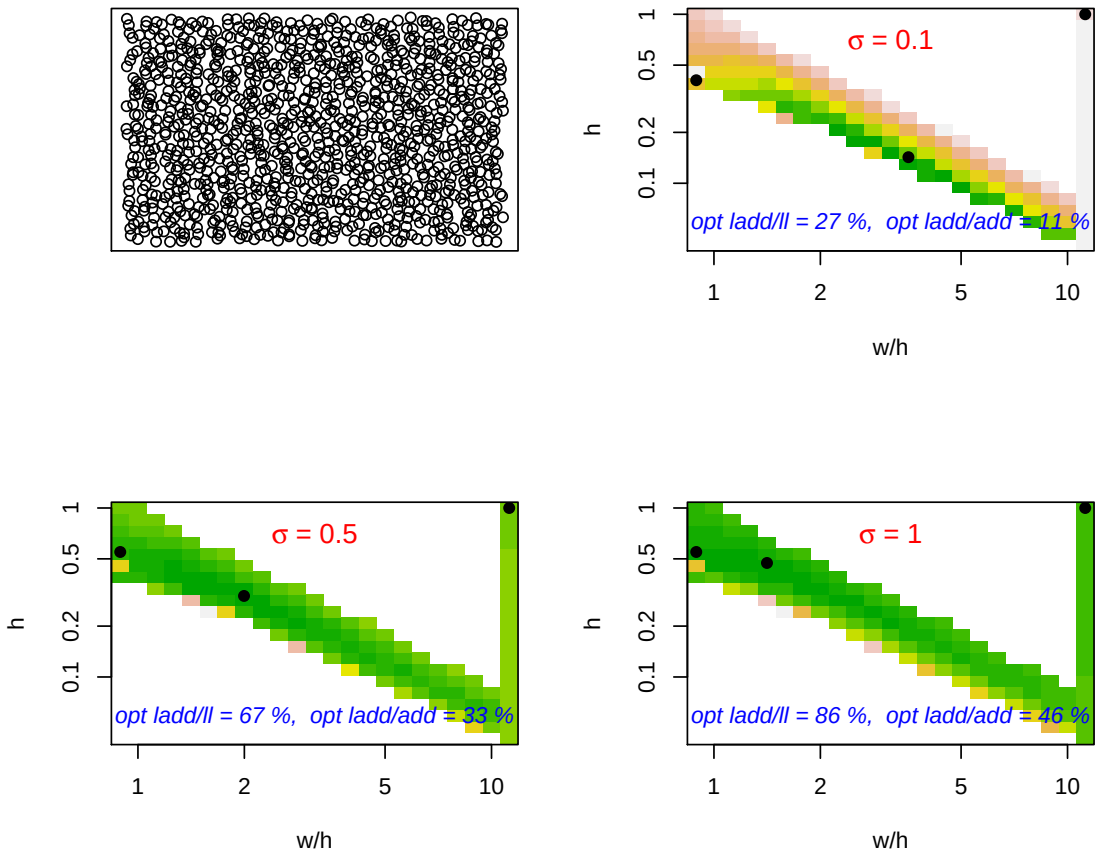

Figure 3: Comparison of MISE performance for 3–dimensional regression function (2.2). Latin square (3,11) jittered design was used to generate 1089 observations. Same explanation as Figure 2 applies except that the first panel is 2-dim projection view of latin square design.

R200			
σ	local linear (h_{opt})	local additive (h_{opt}, w_{opt})	additive (h_{opt})
0.1	6.2=558% (h=0.350)	1.1=100% (h=0.123, w=0.976)	1.6=145% (h=0.166)
0.5	20.7=180% (h=1.000)	11.5=100% (h=0.350, w=0.988)	14.7=128% (h=0.407)
1.0	26.1=105% (h=1.000)	24.9=100% (h=1.000, w=1.000)	21.3=85% (h=1.000)
R400			
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)
R1600			
σ	local linear (h_{opt})	local additive (h_{opt}, w_{opt})	additive (h_{opt})
0.1	0.9=362% (h=0.143)	0.2=100% (h=0.091, w=0.723)	0.2=67% (h=0.091)
0.5	7.3=287% (h=0.302)	2.6=100% (h=0.166, w=0.932)	2.8=108% (h=0.106)
1.0	14.8=189% (h=0.407)	7.8=100% (h=0.166, w=0.932)	9.9=126% (h=0.224)

Table 1: Comparison of MISE performance for additive regression function (r_1) for increasing sample size. At each sample size, results are given at 3 different standard devivations. Local additive estimator tries to mimic optimal additive estimator. Occasional outperformance by local additive estimator is due to slightly different approximation scheme at different output points.

R200			
σ	local linear (h_{opt})	local additive (h_{opt}, w_{opt})	additive (h_{opt})
0.1	2.9=94% (h=0.350)	3.1=100% (h=0.260, w=0.327)	3.9=126% (h=0.302)
0.5	10.5=135% (h=0.741)	7.8=100% (h=0.549, w=0.977)	8.6=111% (h=0.549)
1.0	19.0=121% (h=0.861)	15.7=100% (h=0.861, w=0.966)	15.0=95% (h=0.861)
R400			
σ	local linear (h_{opt})	local additive (h_{opt}, w_{opt})	additive (h_{opt})
0.1	2.6=108% (h=0.260)	2.4=100% (h=0.193, w=0.242)	3.9=160% (h=0.224)
0.5	13.4=124% (h=0.741)	10.8=100% (h=0.638, w=0.901)	10.7=99% (h=0.638)
1.0	33.0=137% (h=1.000)	24.1=100% (h=0.741, w=0.741)	22.7=94% (h=0.741)
R1600			
σ	local linear (h_{opt})	local additive (h_{opt}, w_{opt})	additive (h_{opt})
0.1	0.7=97% (h=0.166)	0.8=100% (h=0.143, w=0.160)	3.1=412% (h=0.143)
0.5	4.8=103% (h=0.407)	4.7=100% (h=0.350, w=0.441)	5.2=112% (h=0.224)
1.0	8.8=103% (h=0.638)	8.5=100% (h=0.473, w=0.668)	10.4=122% (h=0.473)

Table 2: Comparison of MISE performance for approximately additive regression function (r_2) for increasing sample size. At each sample size, results are given at 3 different standard deviations.

R200			
σ	local linear (h_{opt})	local additive (h_{opt}, w_{opt})	additive (h_{opt})
0.1	3.2=100% (h=0.350)	3.2=100% (h=0.260, w=0.518)	7.3=228% (h=0.350)
0.5	11.6=103% (h=0.861)	11.3=100% (h=0.549, w=0.977)	11.8=104% (h=0.741)
1.0	17.5=102% (h=1.000)	17.2=100% (h=0.861, w=0.966)	14.3=83% (h=1.000)
R400			
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)
R1600			
σ	local linear (h_{opt})	local additive (h_{opt}, w_{opt})	additive (h_{opt})
0.1	0.7=101% (h=0.166)	0.7=100% (h=0.143, w=0.180)	6.3=897% (h=0.166)
0.5	5.1=109% (h=0.407)	4.6=100% (h=0.260, w=0.581)	8.4=181% (h=0.224)
1.0	10.0=109% (h=0.549)	9.1=100% (h=0.407, w=0.575)	13.8=150% (h=0.638)

Table 3: Comparison of MISE performance for non-additive regression function with superposed peaks (r_3) for increasing sample size. At each sample size, results are given at 3 different standard deviations.

R200			
σ	local linear (h_{opt})	local additive (h_{opt}, w_{opt})	additive (h_{opt})
0.1	3.2=150% (h=0.350)	2.1=100% (h=0.260, w=0.367)	33.2=1547% (h=0.302)
0.5	21.9=147% (h=0.549)	14.9=100% (h=0.350, w=0.555)	41.9=281% (h=0.407)
1.0	44.3=111% (h=0.638)	40.0=100% (h=0.549, w=0.549)	56.8=142% (h=0.549)
R400			
σ	local linear (h_{opt})	local additive (h_{opt}, w_{opt})	additive (h_{opt})
0.1	1.8=112% (h=0.260)	1.6=100% (h=0.224, w=0.354)	30.4=1887% (h=0.193)
0.5	18.8=105% (h=0.473)	18.0=100% (h=0.407, w=0.512)	41.3=229% (h=0.407)
1.0	51.2=122% (h=0.549)	41.9=100% (h=0.549, w=0.616)	62.6=150% (h=0.549)
R1600			
σ	local linear (h_{opt})	local additive (h_{opt}, w_{opt})	additive (h_{opt})
0.1	0.5=131% (h=0.193)	0.4=100% (h=0.166, w=0.371)	28.3=7051% (h=0.193)
0.5	4.9=144% (h=0.350)	3.4=100% (h=0.224, w=0.562)	30.4=892% (h=0.224)
1.0	11.7=130% (h=0.473)	9.0=100% (h=0.302, w=0.602)	36.7=406% (h=0.224)

Table 4: Comparison of MISE performance for mixture of additive and non-additive polynomial regression function (r_4) for increasing sample size. At each sample size, results are given at 3 different standard deviations.

R200			
σ	local linear (h_{opt})	local additive (h_{opt}, w_{opt})	additive (h_{opt})
0.1	30.7=141% (h=0.350)	21.7=100% (h=0.260, w=0.327)	98.8=455% (h=0.193)
0.5	55.7=105% (h=0.350)	53.3=100% (h=0.260, w=0.367)	115.8=217% (h=0.260)
1.0	131.1=105% (h=0.407)	124.4=100% (h=0.350, w=0.350)	150.4=121% (h=0.350)
R400			
σ	local linear (h_{opt})	local additive (h_{opt}, w_{opt})	additive (h_{opt})
0.1	9.6=199% (h=0.260)	4.8=100% (h=0.166, w=0.263)	96.7=1997% (h=0.166)
0.5	38.3=100% (h=0.260)	38.4=100% (h=0.224, w=0.251)	113.9=296% (h=0.224)
1.0	113.4=98% (h=0.350)	116.2=100% (h=0.350, w=0.350)	155.4=134% (h=0.302)
R1600			
σ	local linear (h_{opt})	local additive (h_{opt}, w_{opt})	additive (h_{opt})
0.1	1.2=124% (h=0.123)	1.0=100% (h=0.106, w=0.188)	88.8=9094% (h=0.166)
0.5	10.8=114% (h=0.193)	9.5=100% (h=0.193, w=0.272)	91.2=958% (h=0.166)
1.0	28.4=112% (h=0.260)	25.3=100% (h=0.224, w=0.354)	98.3=389% (h=0.193)

Table 5: Comparison of MISE performance for mixture of additive and non-additive periodic regression function (r_5) for increasing sample size. At each sample size, results are given at 3 different standard deviations.

R200			
σ	local linear (h_{opt})	local additive (h_{opt}, w_{opt})	additive (h_{opt})
0.1	14.1=87% (h=0.350)	16.2=100% (h=0.260, w=0.327)	98.9=610% (h=0.224)
0.5	39.2=85% (h=0.350)	46.1=100% (h=0.350, w=0.350)	110.6=240% (h=0.350)
1.0	99.7=91% (h=0.473)	109.3=100% (h=0.350, w=0.350)	129.3=118% (h=0.473)
R400			
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)
R1600			
σ	local linear (h_{opt})	local additive (h_{opt}, w_{opt})	additive (h_{opt})
0.1	0.9=108% (h=0.143)	0.8=100% (h=0.123, w=0.195)	87.5=10688% (h=0.224)
0.5	8.2=104% (h=0.260)	7.9=100% (h=0.224, w=0.282)	89.5=1129% (h=0.224)
1.0	21.4=101% (h=0.302)	21.3=100% (h=0.260, w=0.327)	95.9=451% (h=0.224)

Table 6: Comparison of MISE performance for non-additive periodic regression function (r_6) for increasing sample size. At each sample size, results are given at 3 different standard deviations.

F400			
σ	local linear (h_{opt})	local additive (h_{opt}, w_{opt})	additive (h_{opt})
0.1	1.9=96% (h=0.224)	2.0=100% (h=0.224, w=0.251)	3.6=183% (h=0.224)
0.5	9.6=113% (h=0.638)	8.5=100% (h=0.473, w=0.668)	10.8=127% (h=0.473)
1.0	21.2=115% (h=0.741)	18.5=100% (h=0.638, w=0.716)	30.3=164% (h=0.549)
F400j			
σ	local linear (h_{opt})	local additive (h_{opt}, w_{opt})	additive (h_{opt})
0.1	1.6=106% (h=0.193)	1.5=100% (h=0.166, w=0.234)	3.6=233% (h=0.193)
0.5	6.8=109% (h=0.473)	6.2=100% (h=0.549, w=0.871)	8.5=138% (h=0.473)
1.0	13.6=128% (h=0.861)	10.6=100% (h=0.549, w=0.871)	18.9=178% (h=0.861)
LS400			
σ	local linear (h_{opt})	local additive (h_{opt}, w_{opt})	additive (h_{opt})
0.1	2.2=83% (h=0.260)	2.6=100% (h=0.302, w=0.339)	3.6=136% (h=0.224)
0.5	10.2=123% (h=0.638)	8.3=100% (h=0.549, w=0.692)	10.8=130% (h=0.473)
1.0	23.0=132% (h=0.741)	17.5=100% (h=0.638, w=0.716)	29.4=168% (h=0.638)
LS400j			
σ	local linear (h_{opt})	local additive (h_{opt}, w_{opt})	additive (h_{opt})
0.1	1.6=117% (h=0.260)	1.4=100% (h=0.193, w=0.272)	3.4=240% (h=0.193)
0.5	6.8=115% (h=0.473)	5.9=100% (h=0.473, w=0.944)	7.3=124% (h=0.407)
1.0	13.5=127% (h=0.861)	10.7=100% (h=0.549, w=0.871)	17.2=162% (h=0.473)

Table 7: Comparison of MISE performance for approximately additive regression function (r_2) with different designs. At each sample size, results are given at 3 different standard deviations.

F400			
σ	local linear (h_{opt})	local additive (h_{opt}, w_{opt})	additive (h_{opt})
0.1	1.7=92% (h=0.224)	1.8=100% (h=0.224, w=0.251)	6.7=365% (h=0.224)
0.5	10.9=116% (h=0.549)	9.5=100% (h=0.473, w=0.595)	14.3=151% (h=0.350)
1.0	22.7=112% (h=0.861)	20.3=100% (h=0.638, w=0.638)	33.6=165% (h=0.638)
F400j			
σ	local linear (h_{opt})	local additive (h_{opt}, w_{opt})	additive (h_{opt})
0.1	1.4=103% (h=0.193)	1.4=100% (h=0.193, w=0.242)	6.7=482% (h=0.193)
0.5	7.4=94% (h=0.407)	7.8=100% (h=0.407, w=0.575)	12.0=153% (h=0.473)
1.0	14.2=103% (h=1.000)	13.8=100% (h=0.861, w=0.861)	18.7=136% (h=1.000)
LS400			
σ	local linear (h_{opt})	local additive (h_{opt}, w_{opt})	additive (h_{opt})
0.1	2.0=72% (h=0.260)	2.8=100% (h=0.302, w=0.339)	7.0=247% (h=0.224)
0.5	11.9=125% (h=0.638)	9.5=100% (h=0.407, w=0.724)	14.5=153% (h=0.473)
1.0	24.3=125% (h=0.861)	19.4=100% (h=0.638, w=0.638)	32.7=169% (h=0.638)
LS400j			
σ	local linear (h_{opt})	local additive (h_{opt}, w_{opt})	additive (h_{opt})
0.1	1.6=114% (h=0.260)	1.4=100% (h=0.193, w=0.242)	6.7=494% (h=0.193)
0.5	7.4=105% (h=0.407)	7.0=100% (h=0.407, w=0.457)	10.8=154% (h=0.407)
1.0	14.3=105% (h=0.861)	13.7=100% (h=0.861, w=0.861)	20.7=152% (h=1.000)

Table 8: Comparison of MISE performance for non-additive regression function with superposed peaks (r_3) with different designs. At each sample size, results are given at 3 different standard deviations.

Appendix

Proof of Lemma 2 Recall that $\mathcal{S}_{add} = \mathcal{P}_{add}\mathcal{S}_*\mathcal{P}_{add}$. Put $\mathbf{s}(\mathbf{x}) = \mathcal{S}_*\mathbf{r}_{add}(\mathbf{x})$. The components are given as

$$\begin{aligned} s^0(\mathbf{x}) &= S_{0,0}(\mathbf{x})r^0_{add}(\mathbf{x}) + \sum_{j=1}^{d} S_{0,j}(\mathbf{x})r^j_{add}(x_j)\,, \\ s^k(\mathbf{x}) &= S_{k,0}(\mathbf{x})r^0_{add}(\mathbf{x}) + \sum_{j=1}^{d} S_{j,k}(\mathbf{x})r^j_{add}(x_j)\,. \end{aligned}$$

Write the additive function $r^0_{add}(\mathbf{x}) = \sum_{j=1}^{d} r_{add,j}(x_j)$. Then, the components of $\mathcal{S}_{add}\mathbf{r}_{add}(\mathbf{x})$ may be written as

$$\begin{aligned} (\mathcal{S}_{add}\mathbf{r}_{add})^0(\mathbf{x}) &= \frac{1}{2^{d-1}}\sum_{j=1}^{d}\int s^0(\mathbf{x})\,d\mathbf{x}_{-j} - \frac{d-1}{2^d}\int s^0(\mathbf{x})\,d\mathbf{x}\,, \\ (\mathcal{S}_{add}\mathbf{r}_{add})^k(\mathbf{x}) &= \frac{1}{2^{d-1}}\int s^k(\mathbf{x})\,d\mathbf{x}_{-k}\,. \end{aligned}$$

Further simplification yields

$$\begin{aligned} &\int s^0(\mathbf{x})d\mathbf{x}_{-k} \\ &= r_{add,k}(x_k)\int S_{0,0}(\mathbf{x})\,d\mathbf{x}_{-k} + \sum_{j\neq k}\int\Big(\int S_{0,0}(\mathbf{x})\,d\mathbf{x}_{-(j,k)}\Big)r_{add,j}(x_j)\,dx_j \\ &\quad + r^k_{add}(x_k)\int S_{0,j}(\mathbf{x})\,d\mathbf{x}_{-k} + \sum_{j\neq k}\int\Big(\int S_{0,j}(\mathbf{x})\,d\mathbf{x}_{-(j,k)}\Big)r^j_{add}(x_j)\,dx_j\,, \\ &\int s^0(\mathbf{x})d\mathbf{x} \\ &= \sum_{j=1}^{d}\int\Big(\int S_{0,0}(\mathbf{x})\,d\mathbf{x}_{-j}\Big)r_{add,j}(x_j)\,dx_j + \sum_{j=1}^{d}\int\Big(\int S_{0,j}(\mathbf{x})\,d\mathbf{x}_{-j}\Big)r^j_{add}(x_j)\,dx_j\,, \\ &\int s^k(\mathbf{x})\,d\mathbf{x}_{-k} \\ &= r_{add,k}(x_k)\int S_{k,0}(\mathbf{x})\,d\mathbf{x}_{-k} + \sum_{j\neq k}\int\Big(\int S_{k,0}(\mathbf{x})\,d\mathbf{x}_{-(j,k)}\Big)r_{add,j}(x_j)\,dx_j \\ &\quad r^k_{add}(x_k)\int S_{k,k}(\mathbf{x})\,d\mathbf{x}_{-k} + \sum_{j\neq k}\int\Big(\int S_{j,k}(\mathbf{x})\,d\mathbf{x}_{-(j,k)}\Big)r^j_{add}(x_j)\,dx_j\,. \end{aligned}$$

Hence the convergence of the operator is governed by the convergence of $\int S_{0,k}(\mathbf{x})\,d\mathbf{x}_{-j}$, $\int S_{0,k}(\mathbf{x})\,d\mathbf{x}_{-(j,k)}$, $\int S_{j,k}(\mathbf{x})\,d\mathbf{x}_{-(j,k)}$, $j = 1,\cdots,d, k = 0,\cdots,d$, which are simply one– and two–dimensional kernel estimators. Let $\hat{s}(\mathbf{x})$ be one of those estimators. The uniform convergence result can be applied to

derive

$$\sup_{\mathbf{x}} \left|\hat{s}(\mathbf{x}) - E\hat{s}(\mathbf{x})\right| = O\left(\sqrt{\frac{\log n}{nh^d}}\right)$$

(see e.g. Masry 1996). Without boundary correction, the range of $\mathbf{x}$ is understood as the set of interior points. Note that the expectation of E, as a function of $\mathbf{x}$, is used to describe unconditional expectation. When used as a function of $\mathbf{x}$ and $\mathbf{Y}$ elsewhere, it is conditional expectation. The pointwise rate of convergence is $O(\sqrt{nh^d})$. Similar argument applies to the locally additive estimator with $\mathbf{X}$ replaced by $\mathbf{U}$. As long as w is larger than h, the kernel estimators are not affected by the transformation and thus the same holds true for the estimator with $\mathbf{U}$. For example, the convergence of $\int S_{0,k}(\mathbf{x})\, d\mathbf{x}_{-j}$ can be formulated as, for $l \geq 0$,

$$\sup_{u_j} \Big| \frac{1}{\tilde{n}} \sum_{i=1}^{\tilde{n}} K_{\tilde{h}_j}(U_{i,j}, u_j) \Big(\frac{U_{i,j} - u_j}{\tilde{h}_j} \Big)^l - E\Big[\frac{1}{\tilde{n}} \sum_{i=1}^{\tilde{n}} K_{\tilde{h}_j}(U_{i,j}, u_j) \Big(\frac{U_{i,j} - u_j}{\tilde{h}_j} \Big)^l \Big] \Big| = O\left(\sqrt{\frac{\log n}{nh}} \right),$$

where $u_j = (x_j - x_{0j})/w$. Note that the convergence does not depend on the output point $\mathbf{x}_0$ and the local region w. Now assume that $w \to 0$ as $n \to \infty$. Then

$$\sup_{u_j} \Big| E\Big[\frac{1}{\tilde{n}} \sum_{i=1}^{\tilde{n}} K_{\tilde{h}_j}(U_{i,j}, u_j) \Big(\frac{U_{i,j} - u_j}{\tilde{h}_j} \Big)^l \Big] - \frac{1}{2} \int u^l K(u)\, du \Big| = O(w)\,.$$

Therefore, $\tilde{\mathcal{S}}_{add} = \tilde{\mathcal{S}}_{add,n}(\mathbf{x})$ converges uniformly in $\mathbf{x}$ as $n \to \infty$, provided that $nh/\log n \to \infty$, $h/w \to 0$ and $w \to 0$. The other estimators can be treated similarly. Further simplification leads to the expression of the limiting operator defined above. Now we show that the finite operator has also continuous inverse. As $\tilde{\mathcal{S}}_{add,\infty}$ has an inverse operator, it is enough to show that

$$||\tilde{\mathcal{S}}_{add,n} - \tilde{\mathcal{S}}_{add,\infty}|| < \frac{1}{||\tilde{\mathcal{S}}^{-1}_{add,\infty}||}\,.$$

Because $\tilde{\mathcal{S}}_{add,n}$ converges to $\tilde{\mathcal{S}}_{add,\infty}$, the lefthand side can be made arbitrarily small for large n with probability tending to one. Therefore, $\tilde{\mathcal{S}}^{-1}_{add,n}$ exists and has continuous inverse with probability tending to one as $n \to \infty$. □.

Sketch of proof for Lemma 3 From the normal equation (1.8) and Lemma 2, the asymptotic variance can be obtained from

$$V[\hat{r}_{ladd}(\mathbf{x})] = \frac{1}{4^d} V[(\mathcal{P}_{add}\tilde{\mathbf{r}}_L)^0(\mathbf{x})]\big(1 + o(1)\big)\,,$$

and

$$V[(\mathcal{P}_{add}\tilde{r}_L)^0(\mathbf{0})] = \frac{\sigma^2}{4^{d-1}\tilde{n}} \sum_i \Big\{ \sum_j K_{\tilde{h}_j}(U_{i,j}, 0) - \frac{d-1}{2} \Big\}^2\,.$$

Lemma 3 now can be obtained from standard calculation. □.

Sketch of proof for Lemma 4 From the normal equation (1.8) and Lemma 2, the asymptotic bias can be calculated from

$$\begin{aligned} B[\hat{r}_{ladd}(\mathbf{x})] &= \tilde{\mathcal{S}}^{-1}_{add,\infty}\{E[(\mathcal{P}_{add}\tilde{\mathbf{r}}_L)^0(\mathbf{x})] - r(\mathbf{x})\}\big(1+o_p(1)\big) \\ &= \Big(\frac{1}{2^d}E[(\mathcal{P}_{add}\tilde{\mathbf{r}}_L)^0(\mathbf{x})] - r(\mathbf{x})\Big)\big(1+o(1)\big)\,. \end{aligned}$$

Then it can be deduced that

$$E[(\mathcal{P}_{add}\tilde{r}_L)^0(\mathbf{0})] = \frac{1}{2^{d-1}\tilde{n}}\sum_i\Big\{\sum_j \tilde{r}(\mathbf{U}_i)K_{\tilde{h}_j}(U_{i,j},0) - \frac{d-1}{2}\tilde{r}(\mathbf{U}_i)\Big\}\,. \tag{2.3}$$

Lemma 4 now can be obtained from standard calculation. □.

Proof of Lemma 5 From (2.3) together with Lemma 2, we may write

$$E[\hat{b}_{add,w}(\mathbf{u})] = \Big\{2\sum_{l=1}^{d}\Big(\frac{1}{\tilde{n}}\sum_{i=1}^{\tilde{n}}U_{i,j}U_{i,k}K_{\tilde{h}_l}(U_{i,l},u_l)\Big) - (d-1)\Big(\frac{1}{\tilde{n}}\sum_{i=1}^{\tilde{n}}U_{i,j}U_{i,k}\Big)\Big\}(1+o_p(1))\,. \tag{2.4}$$

Then, if f is twice continuously differentiable

$$\begin{aligned} (a) &= E[U_jU_kK_{\tilde{h}_l}(U_l,x_l)] = \int u_ju_kK_{\tilde{h}_l}(u_l,x_l)\tilde{f}_{j,k,l}(u_j,u_k,u_l)\,du_jdu_kdu_l \\ &= \int u_ju_k\Big\{\int K(u_l)\frac{1}{2^3f(\mathbf{x}_0)}\Big(f(\mathbf{x}_0) + w\big(f'_j(\mathbf{x}_0)u_j + f'_k(\mathbf{x}_0)u_k \\ &\quad + f'_l(\mathbf{x}_0)(x_l+\tilde{h}_lu_l)\big) + O(w^2)\Big)\,du_l\Big\}\,du_jdu_k = O(w^2) \end{aligned}$$

$$\begin{aligned} (b) &= E[U_jU_kK_{\tilde{h}_k}(U_k,x_k)] \\ &= \int u_j\Big\{\int (x_k+\tilde{h}_lu_k)K(u_k)\frac{1}{4f(\mathbf{x}_0)}\Big(f(\mathbf{x}_0) + w\big(f'_j(\mathbf{x}_0)u_j \\ &\quad + f'_k(\mathbf{x}_0)(x_k+\tilde{h}_ku_k)\big) + O(w^2)\Big)\,du_k\Big\}\,du_j \\ &= \frac{wf'_j(\mathbf{x}_0)}{4f(\mathbf{x}_0)}\int u_j^2\,du_j + O(w^2) = \frac{wx_k}{6}\frac{\frac{\partial}{\partial u_j}f''_{j,k}(\mathbf{x}_0)}{f_{j,k}(\mathbf{x}_0)} + O(w^2) \\ (c) &= E[U_jU_k] = O(w^2) \end{aligned}$$

Therefore, the result follows. □.

Proof of Proposition 1 Under the assumption $(A.1')$, the non–additive part $\tilde{r}^{(2)}(\mathbf{u})$ of $\tilde{r}(\mathbf{u})$ can be represented as

$$\begin{aligned} &\tilde{r}^{(2)}(\mathbf{u}) \\ &= \frac{w^2}{2}\sum_{j\neq k}r''_{j,k}(\mathbf{0})u_ju_k + \frac{w^3}{3!}\sum_{j,k,l}r'''_{j,k,l}(\mathbf{0})u_ju_ku_l + \frac{w^4}{4!}\sum_{j,k,l,m}r''''_{j,k,l,m}(\mathbf{0})u_ju_ku_lu_m + o(w^4)\,. \end{aligned}$$

Note that $\tilde{f}$ is uniform on $[-1,1]^d$ when f is uniform. This implies that

$$E[U_jU_k]=0\,,E[U_jU_kU_l]=0\,,E[U_j^2U_k]=0$$
$$E[U_jU_kU_lU_m]=0\,,E[U_j^2U_kU_l]=0\,,E[U_j^3U_k]=0\,.$$

It turns out that the same holds true with the kernel function included when evaluated at $\mathbf{u}=\mathbf{0}$. Hence, bias is dominated by the term $u_j^2u_k^2$ with order of w^4. From (2.4), the following Lemma combined with Lemma 4 proves Proposition 1. □

Proof of Lemma 6

$$\begin{aligned}
E[U_j^\delta U_k K_{\tilde{h}_l}(U_l,u_l)] &= 0\,, \quad l\neq j\neq k \quad \text{or} \quad (l=j)\neq k\,,\,\delta=1,2,3\\
E[U_j^3 U_k K_{\tilde{h}_k}(U_k,u_k)] &= 0\,, \quad j\neq k\\
E[U_jU_kU_lK_{\tilde{h}_m}(U_m,u_m)] &= 0\,, \quad m\neq j\neq k\neq l \quad \text{or} \quad (m=j)\neq k\neq l\\
E[U_j^2U_kU_lK_{\tilde{h}_m}(U_m,u_m)] &= 0\,, \quad (m=j)\neq k\neq l \quad \text{or} \quad j\neq(m=k)\neq l\\
E[U_jU_kU_lU_mK_{\tilde{h}_n}(U_n,u_n)] &= 0\,, \quad n\neq j\neq k\neq l\neq m \quad \text{or} \quad (n=j)\neq k\neq l\neq m\,,
\end{aligned}$$

and

$$E[U_j^2U_kK_{\tilde{h}_k}(U_k,u_k)]=\frac{1}{6}u_k\,,\quad j\neq k\,.$$

Hence, when $\mathbf{u}=\mathbf{0}$, the average of those terms are of order $O((\tilde{h}\sqrt{\tilde{n}})^{-1})$. Thus bias is dominated by the term $u_j^2u_k^2$, provided that $r''''_{j,j,k,k}$ is bounded. For $\tilde{r}(\mathbf{u})=\tilde{r}^{(2)}(\mathbf{u})$, we get

$$\begin{aligned}
&\sum_{l=1}^{d}\Big(2E\big[\sum_{j\neq k}r''''_{j,j,k,k}(\mathbf{x}_0)U_j^2U_k^2K_{\tilde{h}_l}(U_l,0)\big]\Big)-(d-1)E\big[\sum_{j\neq k}r''''_{j,j,k,k}(\mathbf{x}_0)U_j^2U_k^2\big]\\
&\quad=\sum_{j\neq k}r''''_{j,j,k,k}(\mathbf{x}_0)\Big\{E\big[U_j^2U_k^2\big](d-2)+2E\big[U_j^2U_k^2\big(K_{\tilde{h}_j}(U_j,0)+K_{\tilde{h}_k}(U_k,0)\big)\big]\Big\}\\
&\qquad-(d-1)\sum_{j\neq k}r''''_{j,j,k,k}(\mathbf{x}_0)E\big[U_j^2U_k^2\big]\\
&\quad=\sum_{j\neq k}r''''_{j,j,k,k}(\mathbf{x}_0)\Big\{2E\big[U_j^2U_k^2\big(K_{\tilde{h}_j}(U_j,0)+K_{\tilde{h}_k}(U_k,0)\big)\big]-E\big[U_j^2U_k^2\big]\Big\}\\
&\quad=\sum_{j\neq k}r''''_{j,j,k,k}(\mathbf{x}_0)\Big\{\mu_2(K)\frac{h_j^2+h_k^2}{3w^2}-\frac{1}{9}\Big\}\,.
\end{aligned}$$

$$B^{(2)}_{ladd}(\mathbf{x}_0)=\Big\{-\frac{w^4}{4!\cdot 9}\sum_{j\neq k}r''''_{j,j,k,k}(\mathbf{x}_0)+\frac{w^2}{4!\cdot 3}\mu_2(K)\sum_{j\neq k}r''''_{j,j,k,k}(\mathbf{x}_0)(h_j^2+h_k^2)\Big\}+o(w^4)$$

Assuming that $h=o(w)$, the leading term is $O(w^4)$. This proves Proposition 1. □

Proof of Proposition 2 From Proposition 1, the leading terms of asymptotic MSE (AMSE) can be expressed as

$$\begin{aligned} AMSE(w) &= w^8\Big(\frac{C_h^2\mu_2(K)}{2}\sum_{j=1}^{d} r''_{j,j}(\mathbf{x}_0) - \frac{1}{4!\cdot 9}\sum_{j\neq k} r''''_{j,j,k,k}(\mathbf{x}_0)\Big)^2 + \frac{2\mu_0(K^2)\sigma^2 d}{C_h C_n n w^{d+1}} \\ &= w^8(aC_h^2 - b)^2 + \frac{c}{C_h C_n n w^{d+1}}\,. \end{aligned}$$

Then,

$$AMSE'(w) = 8(aC_h^2 - b)^2 w^7 - \frac{(d+1)c}{C_h C_n n w^{d+2}} = 0$$

leads to

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

As $AMSE''(w) \geq 0$, the extremal point is minimal, which proves Proposition 2. □

Proof of Proposition 3 From Proposition 2, the optimal AMSE can be expressed as

$$AMSE = (aC_h^2 - b)^2 w^8 + \frac{8(aC_h^2 - b)^2 C_n}{d+1} w^8 = \frac{d+1+8C_n}{d+1}(aC_h^2 - b)^2 w^8\,.$$

Substituting w with the optimal one and writing $x = C_h$ give

$$\begin{aligned} AMSE &= \frac{d+1+8C_n}{d+1}(ax^2 - b)^2\Big(\frac{(d+1)c}{8x(ax^2-b)^2 C_n}\Big)^{\frac{8}{d+9}} n^{-\frac{8}{n+9}}\,, \\ \{AMSE\}^{d+9} &= \Big(\frac{d+1+8C_n}{d+1}\Big)^{d+1+8C_n} n^{-8}(ax^2-b)^{2(d+9)}\Big(\frac{(d+1)c}{8x(ax^2-b)^2 C_n}\Big)^8 \\ &= \Big(\frac{d+1+8C_n}{d+1}\Big)^{d+9} n^{-8}\Big(\frac{(d+1)c}{8}\Big)^8 \frac{(ax^2-b)^{2(d+1)}}{x^8}\frac{1}{C_n^8}\,. \end{aligned}$$

Let

$$g(x) = \frac{(ax^2 - b)^{2(d+1)}}{x^8}\,.$$

It can be shown that all other cases except for $ab < 0$ produce degenerated solutions. When $ab < 0$, the minimizer can be found from the solution of $g'(x) = 0$.

$$\begin{aligned} g'(x) &= \frac{2ax2(d+1)(ax^2-b)^{2d+1}x^8 - 8x^7(ax^2-b)^{2(d+1)}}{x^{16}} \\ &= \frac{(ax^2-b)^{2d+1}\big(4a(d-1)x^2 + 8b\big)}{x^9}\,. \end{aligned}$$

Thus, $g'(x) = 0$ leads to

$$ax^2 - b = 0\,,\ \text{or}\ a(d-1)x^2 + 2b = 0\,.$$

Assuming that $ab < 0$,

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