Journal of Multivariate Analysis

Simple and Efficient Improvements of Multivariate Local Linear Regression

Ming-Yen Cheng[1] **and Liang Peng**[2]

Abstract

This paper studies improvements of multivariate local linear regression. Two intuitively appealing variance reduction techniques are proposed. They both yield estimators that retain the same asymptotic conditional bias as the multivariate local linear estimator and have smaller asymptotic conditional variances. The estimators are further examined in aspects of bandwidth selection, asymptotic relative efficiency and implementation. Their asymptotic relative efficiencies with respect to the multivariate local linear estimator are very attractive and increase exponentially as the number of covariates increases. Data-driven bandwidth selection procedures for the new estimators are straightforward given those for local linear regression. Since the proposed estimators each has a simple form, implementation is easy and requires much less or about the same amount of effort. In addition, boundary corrections are automatic as in the usual multivariate local linear regression.

AMS 1991 subject classification: 62G08; 62G05; 62G20

Keywords. Bandwidth selection, kernel smoothing, local linear regression, multiple regression, nonparametric regression, variance reduction.

[1]Department of Mathematics, National Taiwan University, Taipei 106, Taiwan. Email: cheng@math.ntu.edu.tw

[2]School of Mathematics, Georgia Institute of Technology, Atlanta GA 30332-0160, USA. Email: peng@math.gatech.edu

1 Introduction

Nonparametric regression methods are useful for exploratory data analysis and for representing underlying features that can not be well described by parametric regression models. In the recent two decades, many attentions have been paid to local polynomial modeling for nonparametric regression which was first suggested by Stone (1977) and Cleveland (1979). Fan (1993) and many others investigated the theoretical and numerical properties. Ruppert and Wand (1994) established theoretical results for local polynomial regression with multiple covariates. Wand and Jones (1995), Fan and Gijbels (1996) and Simonoff (1996) provided excellent reviews. We consider reducing variance in multivariate local linear regression. This is of fundamental interests since local linear techniques are very useful and efficient in a wide range of fields including survival analysis, longitudinal data analysis, time series modeling and so on.

The nonparametric regression model with multiple covariates is as follows

$$Y_i = m(X_i) + \nu^{1/2}(X_i)\,\varepsilon_i, \tag{1.1}$$

where $\{X_i = (X_{i1}, \cdots, X_{id})^T\}_{i=1}^n$ are i.i.d. random vectors with density function f and independent of $\varepsilon_1, \cdots, \varepsilon_n$, which are i.i.d. random variables with mean zero and variance one. The local linear estimator of the conditional mean function $m(\cdot)$ at $x = (x_1, \cdots, x_d)^T$ is $\widehat{\alpha}$, the solution for α to the following locally kernel weighted least squares problem

$$\min_{\alpha,\beta} \sum_{i=1}^{n} \left\{Y_i - \alpha - \beta^T(X_i - x)\right\}^2 \Pi_{j=1}^{d} K\Big(\frac{X_{ij} - x_j}{b_j h}\Big), \tag{1.2}$$

where $K(\cdot)$ is a one-dimensional kernel function, $h > 0$, and $b_i > 0$, $i = 1, \cdots, d$, are constants. Here $b_1, \cdots, b_d$ are tuning parameters which allow us to choose different bandwidths for each direction: in (1.2) the bandwidth for kernel smoothing along the i-th covariate is $b_i h$, $i = 1, \cdots, d$. The kernel weight function in (1.2) is taken as a product kernel and the bandwidth matrix $H^{1/2} = diag\{hb_1, \cdots, hb_d\}$ is diagonal. From standard weighted least squares theory, the local linear estimator is given by

$$\widehat{m}(x) = e^T (X_x^T W_x X_x)^{-1} X_x^T W_x Y, \tag{1.3}$$

where $e = (1, 0, \cdots, 0)^T$ is a $(d+1)$-vector, $Y = (Y_1, \cdots, Y_n)^T$,

$$W_x = diag\Big\{\Pi_{j=1}^d K\big(\frac{X_{1j} - x_j}{b_j h}\big), \cdots, \Pi_{j=1}^d K\big(\frac{X_{nj} - x_j}{b_j h}\big)\Big\}, \quad X_x = \begin{pmatrix} 1 & (X_1 - x)^T \\ \vdots & \vdots \\ 1 & (X_n - x)^T \end{pmatrix}.$$

Define

$$B = diag\{b_1^2, \cdots, b_d^2\}, \quad \mu_2(K) = \int s^2 K(s)\, ds, \quad R(K) = \int K^2(s)\, ds,$$

$$M_2(x) = \begin{pmatrix} \frac{\partial^2}{\partial x_1 \partial x_1} m(x), & \cdots, & \frac{\partial^2}{\partial x_1 \partial x_d} m(x) \\ \vdots & \vdots & \vdots \\ \frac{\partial^2}{\partial x_d \partial x_1} m(x), & \cdots, & \frac{\partial^2}{\partial x_d \partial x_d} m(x) \end{pmatrix}.$$

If x is an interior point, Ruppert and Wand (1994) showed that, under regularity conditions,

$$\mathrm{E}\big\{\widehat{m}(x) - m(x)\big|X_1, \cdots, X_n\big\} = \frac{1}{2} h^2 \mu_2(K)\, tr\big\{B M_2(x)\big\} + o_p\big(h^2\big), \tag{1.4}$$

$$\mathrm{Var}\big\{\widehat{m}(x)\big|X_1, \cdots, X_n\big\} = \frac{\nu(x)}{n h^d f(x) \Pi_{i=1}^d b_i} R(K)^d \big\{1 + o_p(1)\big\}. \tag{1.5}$$

Here, that x is an interior point means that the set $S_{x,K} = \big\{(z_1, \cdots, z_d)^T : \Pi_{j=1}^d K\big((z_j - x_j)/(b_j h)\big) > 0\big\}$, i.e. support of the local kernel weight function in the local least squares problem (1.2), is entirely contained in the support of the design density f. Expressions (1.4) and (1.5) reveal behaviors of $\widehat{m}(x)$. The conditional variance has a slower rate of convergence as the number of covariates d increases and the conditional bias is of the same order h^2 for any value of d. Performance of $\widehat{m}(x)$ can be measured by the asymptotically optimal conditional mean squared error, i.e. the asymptotic conditional mean squared error minimized over all bandwidths. It has the order $n^{-4/(d+4)}$ and deteriorates for larger values of d. This is known as the curse of dimensionality problem. It occurs naturally because, with the same sample size, the design points $X_1, \cdots, X_n$ are much less dense in higher dimensions so the variance inflates to a slower rate. Therefore, reducing variance of multivariate local linear regression becomes very important and it is investigated in the subsequent sections.

We propose two types of estimators of $m(x)$ that improve the multivariate local linear regression estimator $\widehat{m}(x)$ in terms of reducing the asymptotic conditional variance while keeping the same asymptotic conditional bias. The first variance reducing estimator is

introduced in Section 2.1. It has a very appealing property of achieving variance reduction while requiring even much less computational effort, by a factor decreasing exponentially in d, than the original local linear estimator. Our second method, proposed in Section 2.2, is even more effective in the sense that its pointwise relative efficiency with respect to $\widehat{m}(x)$ is uniform and is the best that the first method can achieve at only certain points. The way it is constructed can be easily explained by the first method.

Section 2 introduces the variance reducing techniques and investigates the asymptotic conditional biases and variances. Bandwidth selection, the most crucial problem in nonparametric smoothing, is discussed in Section 3. Section 4 studies the asymptotic relative efficiencies and issues such as implementation and boundary corrections. A simulation study and a real application are presented in Section 5. All proofs are given in Section 6.

2 Methodology

2.1 Method I – Fixed Local Linear Constraints

Let $G = (G_1, \cdots, G_d)^T$ be a vector of odd integers, and for each $i = 1, \cdots, d$ let $\{\alpha_{i,j} : j = 1, \cdots, G_i\}$ be an equally spaced grid of points with bin width

$$\delta_i b_i h = \alpha_{i,j+1} - \alpha_{i,j} \quad \text{for} \quad j = 1, \cdots, G_i - 1,$$

where $\delta_i > 0$, $i = 1, \cdots, d$, are given tuning parameters. In practice, choosing $\delta_i \in [0.5, 1.5]$, $i = 1, \cdots, d$, for moderate sample sizes is preferred. Then $\Lambda = \left\{(\alpha_{1,u_1}, \cdots, \alpha_{d,u_d})^T : u_i = 1, \cdots, G_i \text{ for each } i = 1, \cdots, d\right\}$ is a collection of grid points in the range $D = [\alpha_{1,1}, \alpha_{1,G_1}] \times \cdots \times [\alpha_{d,1}, \alpha_{d,G_d}] \subset R^d$. Denote

$$D_v = [\alpha_{1,2v_1}, \alpha_{1,2v_1+2}] \times \cdots \times [\alpha_{d,2v_d}, \alpha_{d,2v_d+2}],$$

where $2v_i \in \{1, 3, \cdots, G_i - 2\}$ for $i = 1, \cdots, d$. Then the D_v's form a partition of D. And for any fixed point $x = (x_1, \cdots, x_d)^T \in D$, there exist two vectors $v = (v_1, \cdots, v_d)^T$ and $r = (r_1, \cdots, r_d)^T$, where $r_i \in [-1, 1]$ for $i = 1, \cdots, d$, such that x is expressed as

$$x_i = \alpha_{i,2v_i+1} + r_i \delta_i b_i h \quad \text{for each } i = 1, \cdots, d. \tag{2.1}$$

So the vector v indicates the subset D_v of D that x belongs to and the vector r marks the location of x relative to the set of grid points that fall within D_v, i.e.,

$$\begin{aligned}\Lambda_v &= \Lambda \cap D_v = \Lambda \cap [\alpha_{1,2v_1}, \alpha_{1,2v_1+2}] \times \cdots \times [\alpha_{d,2v_d}, \alpha_{d,2v_d+2}] \\ &= \big\{x^*(k_1, \cdots, k_d) = (\alpha_{1,2v_1+k_1}, \cdots, \alpha_{d,2v_d+k_d})^T : (k_1, \cdots, k_d)^T \in \{0,1,2\}^d\big\}. \quad (2.2)\end{aligned}$$

The local linear estimator $\widehat{m}(x)$ involves an inverse operation associated with the local design matrix in which only a few design points have positive weights, see (1.2) and (1.3). That contributes much instability to $\widehat{m}(x)$. Therefore, our idea of variance reduction in local linear estimation of $m(x)$ at any $x \in D_v \subset D$ is the following. Given $\{\widehat{m}(\alpha) : \alpha \in \Lambda\}$, i.e. the local linear estimates evaluated over Λ, we form a linear combination of the values $\widehat{m}(\alpha), \alpha \in \Lambda_v \subset \Lambda$, to be a new estimate of $m(x)$ instead of recomputing $\widehat{m}(x)$ at x as in (1.3). In this way the resultant estimate is not allowed to differ too much from the values $\widehat{m}(\alpha), \alpha \in \Lambda_v$, where Λ_v is a degenerate subset of D_v, and its source of variability is restricted to their variances and covariances. In other words, our new estimator at any $x \in D_v$ is constrained by $\widehat{m}(\alpha), \alpha \in \Lambda_v$, and in general will have a smaller variance than $\widehat{m}(x)$. Meanwhile, to ensure the asymptotic conditional bias unchanged, the new estimator has to be subject to certain moment conditions. This can be accomplished by forcing the coefficients in the linear combination to fulfill the corresponding requirements.

Formally, put $\delta = (\delta_1, \cdots, \delta_d)^T$ and let

$$A_0(s) = \frac{s(s-1)}{2}, \quad A_1(s) = 1 - s^2, \quad A_2(s) = \frac{s(s+1)}{2}. \quad (2.3)$$

Our first variance reduced estimator is defined as

$$\begin{aligned}\widetilde{m}(x; r, \delta) &= \sum_{(k_1,\cdots,k_d)^T \in \{0,1,2\}^d} \Big\{\Pi_{i=1}^d A_{k_i}(r_i)\Big\} \widehat{m}\big(x^*(k_1, \cdots, k_d)\big) \\ &= \sum_{x^*(k_1,\cdots,k_d) \in \Lambda_v} \Big\{\Pi_{i=1}^d A_{k_i}(r_i)\Big\} \widehat{m}\big(x^*(k_1, \cdots, k_d)\big). \quad (2.4)\end{aligned}$$

That is, $\widetilde{m}(x; r, \delta)$ is a linear combination of $\widehat{m}(\alpha)$, $\alpha \in \Lambda_v \subset \Lambda$, the original local linear estimates at the 3^d grid points in $\Lambda_v \subset D_v$ where $x \in D_v$.

Since the functions $A_0(s)$, $A_1(s)$ and $A_2(s)$ satisfy $A_0(s) + A_1(s) + A_2(s) = 1$ for all $s \in [-1, 1]$, it is clear from (2.3) and (2.4) that $\widetilde{m}(x; r, \delta) = \widehat{m}(x)$ for all $x \in \Lambda$ and of

course $\widetilde{m}(x;r,\delta)$ and $\widehat{m}(x)$ have the exactly same finite and large sample behaviors over Λ. So the interesting part is $D \setminus \Lambda$ where $\widetilde{m}(x;r,\delta)$ and $\widehat{m}(x)$ are not equal. The fact that $\widetilde{m}(x;r,\delta)$ has a smaller variance than $\widehat{m}(x)$ for all $x \in D \setminus \Lambda$ can be explained in two ways. First, compared to $\widehat{m}(x)$, $\widetilde{m}(x;r,\delta)$ is constructed using more data points as the collection $\{\widehat{m}(\alpha) : \alpha \in \Lambda_v\}$ is based on observations with their X-values falling in a larger neighborhood of x. Another reason is that $\widetilde{m}(x;r,\delta)$ is constrained by the local linear estimates $\widehat{m}(\alpha)$, $\alpha \in \Lambda_v$, instead of being built from a separate local linear fitting or in any modified way, so its source of variation is restricted to the local linear fittings at $\alpha \in \Lambda_v$.

Concerning the bias of $\widetilde{m}(x;r,\delta)$ as an estimator of $m(x)$, the expected values of $\widehat{m}(\alpha)$, $\alpha \in \Lambda_v$, differ from $m(x)$ by more than the usual h^2-order bias of $\widehat{m}(x)$. For $\widetilde{m}(x;r,\delta)$ to have the same asymptotic bias as $\widehat{m}(x)$, noting that points in Λ_v are all distant from x at the order h, the coefficients in the linear combination defining $\widetilde{m}(x;r,\delta)$ have to sum up to one and cancel the extra order-h and order-h^2 biases contributed by $\widehat{m}(\alpha)$, $\alpha \in \Lambda_v$. Since $A_0(s)$, $A_1(s)$ and $A_2(s)$ defined in (2.3) satisfy conditions (6.2), the coefficients $\Pi_{i=1}^{d} A_{k_i}(r_i)$ in the linear combination defining $\widetilde{m}(x;r,\delta)$ fulfill these requirements.

Throughout this paper we assume the following regularity conditions:

(A1) The kernel K is a compactly supported, bounded kernel such that $\mu_2(K) \in (0,\infty)$;

(A2) The point x is in the support of f. At x, ν is continuous, f is continuously differentiable and all second-order derivatives of m are continuous. Also, $f(x) > 0$ and $\nu(x) > 0$;

(A3) The bandwidth h satisfies $h = h(n) \to 0$ and $nh^d \to \infty$ as $n \to \infty$.

Our main result is as follows.

Theorem 1. Suppose that x is any point in D with the corresponding vectors v and r as in (2.1). Assume that every element of Λ_v is an interior point. Then, under conditions (A1)–(A3), as $n \to \infty$,

$$\mathrm{E}\big\{\widetilde{m}(x;r,\delta) - m(x)\big|X_1,\cdots,X_n\big\} = \frac{1}{2}h^2\mu_2(K)\,tr\big\{BM_2(x)\big\} + o_p\big(h^2\big), \qquad (2.5)$$

$$\mathrm{Var}\big\{\widetilde{m}(x;r,\delta)\big|X_1,\cdots,X_n\big\} = \frac{\nu(x)}{nh^d f(x)\Pi_{i=1}^d b_i}\,\Pi_{i=1}^d\big\{R(K)-r_i^2(1-r_i^2)C(\delta_i)\big\} + o_p\big\{(nh^d)^{-1}\big\}, \quad (2.6)$$

where

$$C(s) = \frac{3}{2}C(0,s) - 2C(\frac{1}{2},s) + \frac{1}{2}C(1,s)\,,$$
$$C(s,t) = \int K(u-st)K(u+st)\,du.$$

Hence, from (1.4), (1.5), (2.5) and (2.6), $\widehat{m}(x)$ and $\widetilde{m}(x;r,\delta)$ have the same asymptotic conditional bias and their asymptotic conditional variances differ only by the constant factors $\Pi_{i=1}^d R(K)$ and $\Pi_{i=1}^d\big\{R(K)-r_i^2(1-r_i^2)C(\delta_i)\big\}$. Therefore, comparison between the asymptotic conditional variances lies on the vector δ of the bin widths for Λ and the vector r indicating the location of x in $D_v \subset D$, but not the vector v. The two quantities $\Pi_{i=1}^d R(K)$ and $\Pi_{i=1}^d\big\{R(K)-r_i^2(1-r_i^2)C(\delta_i)\big\}$ are equal when $r_i^2(1-r_i^2)C(\delta_i)=0$ for $i=1,\cdots,d$. Concerning δ, $C(\delta_1)=\cdots=C(\delta_d)=0$ if and only if $\delta_1=\cdots=\delta_d=0$, which corresponds to $\widetilde{m}(x;r,\delta)=\widehat{m}(x)$ for all $x\in D$ and is not meaningful at all. The case that some δ_i are zero is not of any interest either, since in that case Λ is degenerate in the sense that it does not span D. So we are only interested in the case where all the bin widths are positive:

$$\delta_i > 0,\ i=1,\cdots,d.$$

This condition is assumed throughout this paper. As for r, note that $r_i^2(1-r_i^2)=0$ if and only if $r_i\in\{-1,0,1\}$. Under the assumption that $\delta_i>0$, $i=1,\cdots,d$, the two asymptotic conditional variances are equal if and only if $r_i\in\{-1,0,1\}$ for all $i=1,\cdots,d$ and that corresponds to $x\in\Lambda$, which coincide with the fact that $\widetilde{m}(x;r,\delta)=\widehat{m}(x)$ for $x\in\Lambda$. On the other hand, $r_i^2(1-r_i^2)>0$ for all $r_i\in[-1,1]\setminus\{-1,0,1\}$ and, for commonly used kernels, such as the Epanechnikov kernel $K(u)=0.75(1-u^2)I(-1<u<1)$ and the Normal kernel $K(u)=\exp(-u^2/2)/\sqrt{2\pi}$, $C(s)>0$ for all $s>0$; see Cheng, Peng and Wu (2005). Hence we have, for all $x\in D\setminus\Lambda$,

$$\mathrm{Var}\big\{\widetilde{m}(x;r,\delta)\big|X_1,\cdots,X_n\big\} < \mathrm{Var}\big\{\widehat{m}(x)\big|X_1,\cdots,X_n\big\} \text{ asymptotically.}$$

Ratio of the asymptotic conditional variance of $\widetilde{m}(x;r,\delta)$ to that of $\widehat{m}(x)$ is

$$\Pi_{i=1}^d\Big\{\frac{R(K)}{R(K)-r_i^2(1-r_i^2)C(\delta_i)}\Big\}.$$

Given B and δ, this ratio, as well as the pointwise asymptotic relative efficiency of $\widetilde{m}(x;r,\delta)$ with respect to $\widehat{m}(x)$, differs as x varies in D_v. And, irrelevant to the vector v, $\widetilde{m}(x;r,\delta)$ attains the most relative variance reduction when x has its associated vector $r=(r_1,\cdots,r_d)^T$ taking the values $r_i=\pm\sqrt{1/2}$ for all $i=1,\cdots,d$. That is, within each D_v, $\widetilde{m}(x;r,\delta)$ is asymptotically most efficient relative to $\widehat{m}(x)$ at the 2^d points

$$x=(\alpha_{1,2v_1},\cdots,\alpha_{d,2v_d})^T+h\big((1+r_1)\delta_1 b_1,\cdots,(1+r_d)\delta_d b_d\big)^T,\ r\in\{-1/\sqrt{2},1/\sqrt{2}\}^d. \quad (2.7)$$

And the maximum is uniform, hence unique, across all such points and over all subsets D_v of D. Asymptotic relative efficiency of $\widetilde{m}(x;r,\delta)$ with respect to $\widehat{m}(x)$ is investigated in further details in Section 4.1.

2.2 Method II – Varying Local Linear Constraints

As observed in Section 2.1, our first variance reduced estimator $\widetilde{m}(x;r,\delta)$ improves $\widehat{m}(x)$ in a non-uniform manner as x varies in D. And the same best pointwise relative variance reduction occurs at the 2^d points given in (2.7) in each subset D_v of D. Our second variance reducing estimator is then constructed to achieve this best relative efficiency everywhere. The approach is that, fixing at any vector $r\in\big\{-1/\sqrt{2},1/\sqrt{2}\big\}^d$ and for each x, evaluate the usual local linear estimates at 3^d points surrounding x and then linearly combine these estimates to form a new estimator in the same way as in Section 2.1. But now these 3^d neighboring points are determined by x and r and hence differ as x changes.

Consider that given both B and δ being positive vectors. Fix any $r=(r_1,\cdots,r_d)^T\in\{-1/\sqrt{2},1/\sqrt{2}\}^d$. Then, for every x where $m(x)$ is to be estimated, let

$$\alpha_{x,r}=x-h\left((1+r_1)\delta_1 b_1,\cdots,(1+r_d)\delta_d b_d\right)^T,$$

$$\Lambda_{x,r}=\left\{x^*_{k_1,\cdots,k_d}(x;r)=\alpha_{x,r}+h\left(k_1\delta_1 b_1,\cdots,k_d\delta_d b_d\right)^T:(k_1,\cdots,k_d)^T\in\{0,1,2\}^d\right\}.$$

Define a variance reduced estimator of $m(x)$ as

$$\begin{aligned}\widetilde{m}_r(x;\delta) &= \sum_{(k_1,\cdots,k_d)^T\in\{0,1,2\}^d}\left\{\Pi_{i=1}^d A_{k_i}(r_i)\right\}\widehat{m}\big(x^*_{k_1,\cdots,k_d}(x;r)\big)\\ &= \sum_{x^*_{k_1,\cdots,k_d}(x;r)\in\Lambda_{x,r}}\left\{\Pi_{i=1}^d A_{k_i}(r_i)\right\}\widehat{m}\big(x^*_{k_1,\cdots,k_d}(x;r)\big).\end{aligned}$$

Thus $\widetilde{m}_r(x;\delta)$ is a linear combination of the local linear estimates over $\Lambda_{x,r}$ and $\widetilde{m}(x;r,\delta)$ is a linear combination of the local linear estimates at Λ_v. The coefficients in the linear combinations are parallel. These facts explain clearly that $\widetilde{m}_r(x;\delta)$ enjoys the same variance reducing property as $\widetilde{m}(x;r,\delta)$. The main difference between $\widetilde{m}_r(x;\delta)$ and $\widetilde{m}(x;r,\delta)$ is that the set $\Lambda_{x,r}$ in the definition of $\widetilde{m}_r(x;\delta)$ varies as x changes and $r \in \{-1/\sqrt{2}, 1/\sqrt{2}\}^d$ is fixed, while the grid Λ_v for defining $\widetilde{m}(x;r,\delta)$ is fixed for all $x \in D_v = [\alpha_{1,2v_1}, \alpha_{1,2v_1+2}] \times \cdots \times [\alpha_{d,2v_d}, \alpha_{d,2v_d+2}]$ and r depends on x. See also (2.7). Again, $\delta_1, \cdots, \delta_d$ are given tuning parameters and, for moderate sample sizes, choosing $\delta_i \in [0.5, 1.5]$, $i = 1, \cdots, d$, is preferred. If support of X is bounded, say $Supp(X) = [0,1]^d$, then to keep $\Lambda_{x,r}$ within $Supp(X)$, in practice we take $\delta_i(x_i) = \min\big\{\delta_i, x_i/[(1+\sqrt{1/2})h], (1-x_i)/[(1+\sqrt{1/2})h]\big\}$, $i = 1, \cdots, d$, for given $\delta_1, \cdots, \delta_d$.

The following theorem follows immediately from Theorem 1.

Theorem 2. Suppose that r is any given vector in $\{-1/\sqrt{2}, 1/\sqrt{2}\}^d$ and every element of $\Lambda_{x,r}$ is an interior point. Then, under conditions (A1)–(A3), as $n \to \infty$,

$$\mathrm{E}\big\{\widetilde{m}_r(x;\delta) - m(x)\big|X_1, \cdots, X_n\big\} = \frac{1}{2}h^2\mu_2(K)\, tr\big\{BM_2(x)\big\} + o_p\big(h^2\big)\,, \tag{2.8}$$

$$\mathrm{Var}\big\{\widetilde{m}_r(x;\delta)\big|X_1, \cdots, X_n\big\} = \frac{\nu(x)}{nh^d f(x)\Pi_{i=1}^d b_i}\,\Pi_{i=1}^d\big\{R(K) - C(\delta_i)/4\big\}\big\{1 + o_p(1)\big\}\,. \tag{2.9}$$

Therefore the asymptotic conditional biases of $\widetilde{m}_r(x;\delta)$ and $\widehat{m}(x)$ are again the same. And the ratio of the asymptotic conditional variances is constant over all values of x satisfying the conditions in Theorem 2.

There are 2^d such estimators indexed by $r \in \big\{-1/\sqrt{2}, 1/\sqrt{2}\big\}^d$. For a particular value of r, since the set $\Lambda_{x,r}$ is skewed around x, i.e. the 3^d points in $\Lambda_{x,r}$ are asymmetrically distributed about x, finite sample bias of $\widetilde{m}_r(x;\delta)$ may be more than the asymptotic prediction. A way to avoid potential finite sample biases arising from this skewness is to take an average of all the 2^d estimates. That is, given B and δ, the averaged variance reduced estimator is defined as

$$\bar{m}(x;\delta) = 2^{-d} \sum_{r \in \{-1/\sqrt{2}, 1/\sqrt{2}\}^d} \widetilde{m}_r(x;\delta)\,,$$

for every x. The following theorem is proved in Section 6.

Theorem 3. Suppose that every element of $\Lambda_{x,r}$ is an interior point for every $r \in \{-1/\sqrt{2}, 1/\sqrt{2}\}^d$. Then, under conditions (A1)–(A3), as $n \to \infty$,

$$\mathrm{E}\big\{\bar{m}(x;\delta) - m(x)\big|X_1, \cdots, X_n\big\} = \frac{1}{2}h^2\mu_2(K)\, tr\big\{BM_2(x)\big\} + o_p\big(h^2\big), \tag{2.10}$$

$$\begin{aligned}\mathrm{Var}\big\{\bar{m}(x;\delta)\big|X_1, \cdots, X_n\big\} &= \frac{\nu(x)}{nh^d f(x)\Pi_{i=1}^d b_i}\Pi_{i=1}^d\Big\{R(K) - \frac{C(\delta_i)}{4} - \frac{D(\delta_i)}{2}\Big\} \\ &\quad + o_p\big\{(nh^d)^{-1}\big\}, \end{aligned} \tag{2.11}$$

where

$$\begin{aligned} D(s) &= C(0,s) - \frac{1}{4}C(s) - \frac{1+\sqrt{2}}{4}C\Big(\frac{\sqrt{2}-1}{2}, s\Big) - \frac{3+2\sqrt{2}}{16}C\Big(\frac{2-\sqrt{2}}{2}, s\Big) \\ &\quad -\frac{1}{8}C\Big(\frac{\sqrt{2}}{2}, s\Big) - \frac{1-\sqrt{2}}{4}C\Big(\frac{\sqrt{2}+1}{2}, s\Big) - \frac{3-2\sqrt{2}}{16}C\Big(\frac{\sqrt{2}+2}{2}, s\Big). \end{aligned}$$

The quantity $D(\delta_i)$ in (2.11) is always nonnegative for $\delta_i \geq 0$, see Cheng, Peng and Wu (2005). Therefore, from (2.8)–(2.11), besides being equally biased asymptotically as $\widetilde{m}_r(x;\delta)$, $\bar{m}(x;\delta)$ has an even smaller asymptotic conditional variance.

3 Bandwidth Selection

The asymptotically optimal local bandwidth that minimizes the asymptotic conditional mean squared error of $\widehat{m}(x)$ is

$$h_0(x) = \Big[\frac{\nu(x)R(K)^d}{nf(x)\mu_2(K)^2\, tr\{BM_2(x)\}^2\Pi_{i=1}^d b_i}\Big]^{1/(d+4)}, \tag{3.1}$$

and those for $\widetilde{m}(x;r,\delta)$, $\widetilde{m}_r(x;\delta)$ and $\bar{m}(x;\delta)$ are respectively

$$h_1(x) = B_1(x,r;\delta)\, h_0(x)\,, \quad h_2(x) = B_2(\delta)\, h_0(x)\,, \quad h_3(x) = B_3(\delta)\, h_0(x)\,, \tag{3.2}$$

where

$$\begin{aligned} B_1(x,r;\delta) &= \Pi_{i=1}^d\Big[\big\{R(K) - r_i^2(1-r_i^2)C(\delta_i)\big\}/R(K)\Big]^{1/(d+4)}, \\ B_2(\delta) &= \Pi_{i=1}^d\Big[\big\{R(K) - C(\delta_i)/4\big\}/R(K)\Big]^{1/(d+4)}, \\ B_3(\delta) &= \Pi_{i=1}^d\Big[\big\{R(K) - C(\delta_i)/4 - D(\delta_i)/2\big\}/R(K)\Big]^{1/(d+4)}. \end{aligned}$$

Many popular and reliable data-driven bandwidth selection rules for kernel smoothing are constructed based on the asymptotically optimal bandwidth expressions. Note that $C(\delta_i)$'s and $D(\delta_i)$'s in (3.2), relating the asymptotically optimal local bandwidths, are all constants determined by δ and the kernel K. Then an important implication of (3.2) is that data-based local bandwidth selection for any of the proposed estimators $\widetilde{m}(x; r, \delta)$, $\widetilde{m}_r(x; \delta)$ and $\bar{m}(x; \delta)$ is simply a matter of adjusting any local bandwidth selector for $\widehat{m}(x)$ by multiplying the constant factors accordingly.

Asymptotically optimal global bandwidths for a kernel estimator $\bar{m}(x; h)$ of $m(x)$ based on bandwidth h are usually derived from global measures of discrepancy such as

$$\mathrm{IAMSE}\big(\bar{m}; h\big) = \int_D \mathrm{AMSE}\big\{\bar{m}(x; h)\big\}\, f(x)\, w(x)\, dx\,, \tag{3.3}$$

where $\mathrm{AMSE}\big\{\bar{m}(x; h)\big\}$ is the asymptotic conditional mean squared error of $\bar{m}(x; h)$ and $w(x)$ is a known weight function. Asymptotically optimal global bandwidths for $\widetilde{m}(x; r, \delta)$ and $\widehat{m}(x)$ that minimize this IAMSE measure with respect to h usually do not admit the simple relation the local counterparts have in (3.2). The reason is that the relative variance reduction achieved by $\widetilde{m}(x; r, \delta)$ is non-uniform over every D_v. However, suppose that the conditional variance function $\nu(x)$ has a low curvature within each D_v. Then, since the bin widths $\delta_i b_i h$, $i = 1, \cdots, d$, of Λ are all of order h, a sensible data-driven global bandwidth for $\widetilde{m}(x; r, \delta)$ can be obtained from multiplying one for $\widehat{m}(x)$ by the constant factor

$$\left[\int_D \Pi_{i=1}^d \big\{R(K) - r_i^2(1 - r_i^2)C(\delta_i)\big\}\, w(x)\, dx \,/ \int_D R(K)^d\, w(x)\, dx\right]^{1/(d+4)}, \tag{3.4}$$

see (1.4), (1.5), (2.5) and (2.6). Hence, the task of automatically choosing a global bandwidth for the proposed estimator $\widetilde{m}(x; r, \delta)$ can be done analogously, or at least with not much more difficulty, as that for $\widehat{m}(x)$.

Denote as h_0, h_2 and h_3 the asymptotically optimal global bandwidths of $\widehat{m}(\cdot)$, $\widetilde{m}_r(\cdot; \delta)$ and $\bar{m}(\cdot; \delta)$. They are defined as the bandwidths minimizing the global measure in (3.3) for the respective estimators, with D being the set of points where the estimators are all defined. Then, from (2.5), (2.6), (2.8)–(2.11),

$$h_2 = B_2(\delta)\, h_0, \quad h_3 = B_3(\delta)\, h_0\,, \tag{3.5}$$

where $B_2(\delta)$ and $B_3(\delta)$ are exactly as given in (3.2). The constants $B_2(\delta)$ and $B_3(\delta)$ depend only on the known kernel function K and the given bin width vector δ. Hence,

automatic global bandwidth selection procedures for both $\widetilde{m}_r(x;\delta)$ and $\bar{m}(x;\delta)$ can be easily established from adjusting those for the usual multivariate local linear estimator $\widehat{m}(x)$ by the appropriate constants. For example, Ruppert (1997) and Yang and Tschernig (1999) established automatic bandwidth procedures for multivariate local linear regression. Thus the new estimators $\widetilde{m}_r(x;\delta)$ and $\bar{m}(x;\delta)$ both enjoy the appealing advantage that they achieve a great extent of variance reduction and improvement of efficiency without introducing any further unknown factors to bandwidth selection, the most important issue in nonparametric smoothing.

All the above adjustments are rather easy compared to bandwidth selection problems created by other modifications of local linear estimation, which usually change both the asymptotic conditional bias and variance. The reason for this major advantage is that the pointwise asymptotic conditional bias is unaffected everywhere no matter which of the variance reduction methods is applied. An even more promising feature of both $\widetilde{m}_r(x;\delta)$ and $\bar{m}(x;\delta)$ is that they each has its pointwise asymptotic variance as the same constant multiple of that of $\widehat{m}(x)$ at all x.

4 Comparisons

This section is devoted to examine in details impacts of the new estimators on several essential issues in nonparametric smoothing, including relative efficiency, implementation and boundary effects.

4.1 Relative Efficiencies

The pointwise asymptotically optimal conditional mean squared error of $\widehat{m}(x)$, achieved by $h_0(x)$ in (3.1), is

$$\mathrm{AMSE}\big\{\widehat{m}(x);h_0(x)\big\} = \frac{5}{4}\Big\{\frac{\nu(x)R(K)^d}{nf(x)\,\Pi_{i=1}^d b_i}\Big\}^{4/(d+4)}\Big[\mu_2(K)^2\,tr\{BM_2(x)\}^2\Big]^{d/(d+4)}. \qquad (4.1)$$

The asymptotically optimal local bandwidths in (3.2) respectively yield the pointwise asymptotically optimal mean squared errors of $\widetilde{m}(x;r,,\delta)$, $\widetilde{m}_r(x;\delta)$ and $\bar{m}(x;\delta)$ as

$$\mathrm{AMSE}\big\{\widetilde{m}(x;r,\delta);h_1(x)\big\} = B_1(x;r,\delta)^4\,\mathrm{AMSE}\big\{\widehat{m}(x);h_0(x)\big\}\,,$$

$$\mathrm{AMSE}\big\{\widetilde{m}_r(x;\delta);h_2(x)\big\} = B_2(\delta)^4\,\mathrm{AMSE}\big\{\widehat{m}(x);h_0(x)\big\}\,,$$

$$\mathrm{AMSE}\big\{\bar{m}(x;\delta);h_3(x)\big\} = B_3(\delta)^4\,\mathrm{AMSE}\big\{\widehat{m}(x);h_0(x)\big\}\,.$$

Thus, given B and δ, the pointwise asymptotic relative efficiencies of the new estimators with respect to $\widehat{m}(x)$ are

$$\mathrm{Eff}\big\{\widetilde{m}(x;r,\delta),\widehat{m}(x)\big\} = B_1(x;r,\delta)^{-4}, \tag{4.2}$$

$$\mathrm{Eff}\big\{\widetilde{m}_r(x;\delta),\widehat{m}(x)\big\} = B_2(\delta)^{-4}, \quad \mathrm{Eff}\big\{\bar{m}(x;\delta),\widehat{m}(x)\big\} = B_3(\delta)^{-4}\,, \tag{4.3}$$

for every x. Similarly, taking the asymptotically optimal global bandwidths in (3.5) yields that the global asymptotic relative efficiencies of $\widetilde{m}_r(x;\delta)$ and $\bar{m}(x;\delta)$ with respect to $\widehat{m}(x)$ are

$$\mathrm{Eff}\big\{\widetilde{m}_r(\cdot;\delta),\widehat{m}(\cdot)\big\} = B_2(\delta)^{-4}\,, \quad \mathrm{Eff}\big\{\bar{m}(\cdot;\delta),\widehat{m}(\cdot)\big\} = B_3(\delta)^{-4}\,. \tag{4.4}$$

Therefore, for each of $\widetilde{m}_r(x;\delta)$ and $\bar{m}(x;\delta)$, local and global asymptotic relative efficiencies compared to $\widehat{m}(x)$ are the same and depend only on δ and K. Global asymptotic relative efficiency of $\widetilde{m}(x;r,\delta)$ with respect to $\widehat{m}(x)$ has a more complex form since the pointwise relative variance reduction is non-uniform. However, it is well approximated by a simple expression, not elaborated here, arising from the constant adjustment (3.4) to the global bandwidth.

Rewrite the pointwise and global asymptotic relative efficiencies $B_2(\delta)^{-4}$ and $B_3(\delta)^{-4}$, in (4.3) and (4.4), as

$$B_2(\delta)^{-4} = \Pi_{i=1}^d\, S(\delta_i)^{-4/(d+4)}\,, \quad B_3(\delta)^{-4} = \Pi_{i=1}^d\, T(\delta_i)^{-4/(d+4)}\,,$$

where $S(u) = \big\{R(K) - C(u)/4\big\}/R(K)$ and $T(u) = \big\{R(K) - C(u)/4 - D(u)/2\big\}/R(K)$. Consider the simplest case that $\delta_1 = \cdots = \delta_d = \delta_0$. Then

$$\begin{aligned}
\mathrm{Eff}\big\{\widetilde{m}_r(x;\delta),\widehat{m}(x)\big\} &= \mathrm{Eff}\big\{\widetilde{m}_r(\cdot;\delta),\widehat{m}(\cdot)\big\} = S(\delta_0)^{-4d/(d+4)}\,,\\
\mathrm{Eff}\big\{\bar{m}(x;\delta),\widehat{m}(x)\big\} &= \mathrm{Eff}\big\{\bar{m}(\cdot;\delta),\widehat{m}(\cdot)\big\} = T(\delta_0)^{-4d/(d+4)}\,.
\end{aligned} \tag{4.5}$$

For commonly used kernels, $S(\delta_0)^{-1}$ is greater than one for all $\delta_0 > 0$, increases in δ_0, and has an upper limit 8/5. Also, $T(\delta_0)^{-1}$ is greater than one for all $\delta_0 > 0$, increases in δ_0, and has an upper limit 16/5. If K is supported on [-1,1], the respective upper limits occur when $\delta_0 = 2$ and $\delta_0 = 2/(\sqrt{2}-1)$ in which case variances of the proposed estimators consist of only variances and no covariance of the local linear estimates in the linear combinations. Then, from (4.5), an important property of our variance reduction techniques is that the relative efficiency improvements increase as the dimensionality d of the covariates $X = (X_1, \cdots, X_d)^T$ grows. Table 1 contains some values of $S(\delta_0)^{-4d/(d+4)}$ and $T(\delta_0)^{-4d/(d+4)}$ when K is the Epanechnikov kernel.

Table 1: Relative efficiencies $\mathrm{Eff}\{\widetilde{m}_r(\cdot;\delta), \widehat{m}(\cdot)\}$ (upper half) and $\mathrm{Eff}\{\bar{m}(\cdot;\delta), \widehat{m}(\cdot)\}$ (lower half) when $\delta_1 = \cdots = \delta_d = \delta_0$ and when K is the Epanechnikov kernel.

δ_0	d=1	d=2	d=3	d=4	d=5
0.6	1.064	1.110	1.143	1.169	1.189
0.8	1.088	1.151	1.198	1.235	1.264
1.0	1.113	1.195	1.257	1.306	1.345
1.2	1.166	1.292	1.391	1.469	1.533
1.6	1.293	1.535	1.735	1.902	2.043
2.0	1.456	1.871	2.238	2.560	2.842
0.6	1.089	1.153	1.201	1.238	1.268
0.8	1.121	1.210	1.278	1.332	1.375
1.0	1.168	1.295	1.394	1.473	1.538
1.2	1.237	1.425	1.576	1.700	1.804
1.6	1.393	1.737	2.033	2.288	2.509
2.0	1.580	2.142	2.663	3.136	3.560
$2/(\sqrt{2}-1)$	2.536	4.716	7.345	10.240	13.260

4.2 Implementation

Since there is no parametric structural assumptions on the unknown regression function $m(\cdot)$ in nonparametric smoothing, nonparametric regression estimators are usually evaluated over a range of x-values in practice. Consider computing estimators of $m(x)$ over a fine grid Λ_0 to provide sensible comprehension of the regression function $m(\cdot)$ over a range of interest. Then the local linear estimator $\widehat{m}(x)$ requires to perform the local least

squares fitting (1.2) at every $x \in \Lambda_0$.

To compute our first variance reduced estimator $\widetilde{m}(x; r, \delta)$ for all $x \in \Lambda_0$ one can proceed as follows. Form D and Λ as in Section 2.1 such that D covers Λ_0 and Λ is coarser than Λ_0, and then compute the estimates $\widetilde{m}(x; r, \delta)$, $x \in \Lambda_0$, using $\widehat{m}(\alpha), \alpha \in \Lambda$, as described in Section 2.1. One appealing feature of $\widetilde{m}(x; r, \delta)$ is that it is very easy to implement since the coefficients in the linear combination are just products of the simple one dimensional functions A_0, A_1 and A_2. Therefore, implemented in the above mentioned way, $\widetilde{m}(x; r, \delta)$ amounts to a considerable saving of computational time compared to $\widehat{m}(x)$. This is a particularly important advantage in the multivariate case. For example, if the number of grid points at each dimension in Λ is a fixed proportion of that in Λ_0, then the number of evaluations of the local linear estimator is reduced exponentially as the dimension d increases. Hence, the better asymptotic conditional mean squared error performance of $\widetilde{m}(x; r, \delta)$ compared to $\widehat{m}(x)$ is true at virtually little cost but a great saving of computational effort.

The estimators $\widetilde{m}_r(x, \delta)$ and $\bar{m}(x, \delta)$ are more computationally involved. For example, in order to evaluate $\widetilde{m}_r(x, \delta)$ at each x, one needs to calculate the 3^d local linear estimates $\widehat{m}(\alpha)$, $\alpha \in \lambda_{x,r}$. In a naive way, that requires 3^d times the effort compared to what $\widehat{m}(x)$ needs. Fortunately, there are ways to avoid such an increase of computational effort. Suppose that again $\widetilde{m}_r(x, \delta)$ is to be evaluated over a grid Λ_0. One approach is to take the bin widths of the grid Λ_0 to be in proportions to the bin widths $\delta_i\, b_i\, h$, $i = 1, \cdots, d$, of $\Lambda_{x,r}$ so that $\widehat{m}(\alpha)$, $\alpha \in \Lambda_{x,r}$, can be reused for other values of $x \in \Lambda_0$. Also, the set of coefficients in the linear combination is the same for all $x \in \Lambda_0$ so needs to be evaluated once only. Hence, in this manner, $\widetilde{m}_r(x, \delta)$ is computed with about the same amount of effort as the local linear estimator $\widehat{m}(x)$. The estimator $\bar{m}(x, \delta)$ can be constructed in a similar way to alleviate computational effort.

4.3 Behaviors at Boundary Regions

One reason that the local linear technique is very popular in practice and in many contexts is that it does boundary correction automatically. For instance, when x is a boundary

point, the conditional bias and variance of $\widehat{m}(x)$ are both kept at the same orders as in the interior. Theorem 2.2 of Ruppert and Wand (1994) formally defines a boundary point in multivariate local linear regression and provides asymptotic expressions of the conditional bias and variance. The theorem shows that only the constant factors involving K and x are changed and the constant factors depend on how far, relative to the bandwidth matrix, x is away from the boundary.

Clearly from the form of $\widetilde{m}(x; r, \delta)$, Theorem 1 can be extended to include the case where Λ_v is not entirely contained in the interior, so that asymptotic results of the conditional bias and variance of $\widetilde{m}(x; r, \delta)$ are given for every $x \in D$. This is not elaborated here because it is straightforward but the notation becomes much more complicated. One can show that the conditional bias and variance is of the same orders at all $x \in D$ as long as Λ_v is in the support of f, which is of course true in general. Therefore, $\widetilde{m}(x; r, \delta)$ also achieves automatic boundary corrections.

For a boundary point x, comparison between the conditional variances of $\widehat{m}(x)$ and $\widetilde{m}(x; r, \delta)$ becomes tedious as the constant coefficients are both very complex. However, we can argue that $\mathrm{Var}\big\{\widetilde{m}(x; r, \delta)\big|X_1, \cdots, X_n\big\}$ is again asymptotically smaller than $\mathrm{Var}\big\{\widehat{m}(x)\big|X_1, \cdots, X_n\big\}$ at boundary regions in the following way. Our estimator $\widetilde{m}(x; r, \delta)$ is a linear combination of $\widehat{m}(\alpha), \alpha \in \Lambda_v$. It is well known that $\mathrm{Var}\big\{\widehat{m}(\alpha)\big|X_1, \cdots, X_n\big\}$ is much smaller than $\mathrm{Var}\big\{\widehat{m}(x)\big|X_1, \cdots, X_n\big\}$ for those $\alpha \in \Lambda_v$ that are more away from the boundary than x. The weight in $\widetilde{m}(x; r, \delta)$ put on any $\widehat{m}(\alpha)$ with α closer to the boundary than x becomes close to one only when x is right nearby it. Otherwise, $\widetilde{m}(x; r, \delta)$ spreads its weights on $\widehat{m}(\alpha)$ for those $\alpha \in \Lambda_v$ more away from the boundary than x.

The constant multiplier in the asymptotic expression of $\mathrm{E}^2\big\{\widehat{m}(x) - m(x)\big|X_1, \cdots, X_n\big\}$ changes and it generally becomes smaller as x moves from the interior to the boundary region. Therefore, $\widetilde{m}(x; r, \delta)$ no longer has the same asymptotic conditional bias as $\widehat{m}(x)$ at a boundary point x. However, it is generally true that, as x moves more toward the boundary, the decrease in $\mathrm{E}^2\big\{\widehat{m}(x) - m(x)\big|X_1, \cdots, X_n\big\}$ is much less than the increase in $\mathrm{Var}\big\{\widehat{m}(x)\big|X_1, \cdots, X_n\big\}$. Then the difference between $\mathrm{E}^2\big\{\widetilde{m}(x; r, \delta) - m(x)\big|X_1, \cdots, X_n\big\}$ and $\mathrm{E}^2\big\{\widehat{m}(x) - m(x)\big|X_1, \cdots, X_n\big\}$ is small compared to the variance reduction yielded by $\widetilde{m}(x; r, \delta)$. This implies that the asymptotic conditional mean squared error of $\widetilde{m}(x; r, \delta)$

is again smaller than that of $\widehat{m}(x)$ when x is a boundary point.

Consider behaviors of $\widetilde{m}_r(x;\delta)$ and $\bar{m}(x;\delta)$ in boundary regions. Suppose that $\Lambda_{x,r}$ is entirely contained in the support of the design density f. Then the above arguments can be used to demonstrate that $\widetilde{m}_r(x;\delta)$ and $\bar{m}(x;\delta)$ also have the automatic boundary correction property, and that they have smaller asymptotic mean squared errors than $\widehat{m}(x)$ in boundary regions.

Asymptotic behaviors of the conditional bias and variance of the local linear estimator $\widehat{m}(x)$ when x is near the boundary are illustrated and discussed in details by Fan and Gijbels (1992), Ruppert and Wand (1994) and Fan and Gijbels (1996), among others.

5 Simulation study and real application

5.1 Simulation Study

Consider model (1.1) with $d = 2$, $\nu(x) = 1$, $m(x_1, x_2) = \sin(2\pi x_1) + \sin(2\pi x_2)$. Let X_{i1} and X_{i2} be independent with each being uniformly distributed on $[0, 1]$, and ε_i has a standard Normal distribution. We drew samples from the above model with sample size $n = 400$ and $n = 1000$. In computing $\hat{m}(x_1, x_2)$ and $\bar{m}(x_1, x_2; \delta)$, we employed the biweight kernel $K(u) = \frac{15}{16}(1 - u^2)^2 I(|u| \leq 1)$, $h = 0.15$ and 0.2, $b_1 = b_2 = 1$, $\delta_i = min\{1, x_i/[(1 + 1\sqrt{2})h], (1 - x_i)/[(1 + 1/\sqrt{2})h]\}$ for $i = 1, 2$. In Figures 1-4, the natural logarithm of ratio of the mean squared error of $\bar{m}(x_1, x_2; \delta)$ to that of $\hat{m}(x_1, x_2)$ is plotted against different $x_1 = 0, 0.05, \cdots, 1$ and $x_2 = 0, 0.05, \cdots, 1$, and also against $x_2 = 0, 0.5, \cdots, 1$, but with fixed $x_1 = 0.2, 0.5, 0.8$. These figures show that, in all cases considered, $\bar{m}(x_1, x_2; \delta)$ has a significantly smaller mean squared error than $\hat{m}(x_1, x_2)$ when (x_1, x_2) is an interior point. Occasionally, the mean squared error of $\bar{m}(x_1, x_2; \delta)$ is slightly larger than that of $\hat{m}(x_1, x_2)$ for some boundary points. Boundary behaviours is not a main issue in nonparametric multivariate regression since the data contain very little information about the regression surface there.

5.2 Real application

We applied the local linear estimator $\hat{m}(\cdot)$ and our variance reduced estimate $\bar{m}(\cdot;\delta)$ to the Boston housing price data set. This data set consists of the median value of owner-occupied homes in 506 U.S. census tracts in the Boston area in 1970, together with several variables which might explain the variation of housing value, see Harrison and Rubinfeld (1978). Here we fit model (1.1) to the median values of homes with two covariates, $x_1 = LSTAT$ (lower status of the population) and $x_2 = PTRATIO$ (pupil-teacher ratio by town). We computed estimators $\hat{m}(x_1, x_2)$ and $\bar{m}(x_1, x_2; \delta)$ by taking the same set of tuning parameters δ_i, $i = 1, 2$, and the same kernel as in Section 5.1 and $h = 1$, and $b_1 = 1.5, b_2 = 0.5$ were employed here. These estimators are plotted in Figure 5. Close to the boundary $PTRATIO = 0$, mainly $PTRATIO$ less than 1.8, the two estimators behave quite differently since the regression surface changes drastically in that boundary region. The spikes in $\bar{m}(x_1, x_2; \delta)$ will disappear if smaller values of δ_i, $i = 1, 2$, are used. For $PTRATIO > 3$ or $LSTAT > 10$, our estimator $\bar{m}(x_1, x_2; \delta)$ effectively smoothes out the spurious bumps produced by $\hat{m}(x_1, x_2)$.

6 Proofs

Proof of Theorem 1. Put $\Delta = ((k_1 - 1 - r_1)\delta_1 b_1, \cdots, (k_d - 1 - r_d)\delta_d b_d)^T$. Then $x^*(k_1, \cdots, k_d) - x = h\Delta$. Further

$$m(x^*(k_1, \cdots, k_d)) - m(x) = h\Delta^T M_1(x) + \frac{1}{2}h^2\Delta^T M_2(x)\Delta + o(h^2), \tag{6.1}$$

where $M_1(x) = \left(\frac{\partial}{\partial x_1}m(x), \cdots, \frac{\partial}{\partial x_d}m(x)\right)^T$. Since, for any $s \in [-1, 1]$,

$$\sum_{j\in\{0,1,2\}} A_j(s) = 1\,, \quad \sum_{j\in\{0,1,2\}} A_j(s)(j - 1 - s)^l = 0\,,\ l = 1,\ 2\,, \tag{6.2}$$

we have $\displaystyle\sum_{(k_1,\cdots,k_d)^T\in\{0,1,2\}^d}\Big\{\Pi_{i=1}^d A_{k_i}(r_i)\Big\}=1,$

$$
\begin{aligned}
&\mathrm{E}\big\{\widetilde{m}(x;r,\delta)-m(x)\big|X_1,\cdots,X_n\big\}\\
&= \sum_{(k_1,\cdots,k_d)^T\in\{0,1,2\}^d}\Big\{\Pi_{i=1}^d A_{k_i}(r_i)\Big\}\mathrm{E}\big\{\widehat{m}\big(x^*(k_1,\cdots,k_d)\big)-m\big(x^*(k_1,\cdots,k_d)\big)\big|X_1,\cdots,X_n\big\}\\
&\quad+\sum_{(k_1,\cdots,k_d)^T\in\{0,1,2\}^d}\Big\{\Pi_{i=1}^d A_{k_i}(r_i)\Big\}\Big\{m\big(x^*(k_1,\cdots,k_d)\big)-m(x)\Big\}\\
&= \sum_{(k_1,\cdots,k_d)^T\in\{0,1,2\}^d}\Big\{\Pi_{i=1}^d A_{k_i}(r_i)\Big\}\Big[\frac{1}{2}\mu_2(K)h^2\,tr\big\{BM_2(x)\big\}+o_p\big(h^2\big)\Big]\\
&\quad+\sum_{(k_1,\cdots,k_d)^T\in\{0,1,2\}^d}\Big\{\Pi_{i=1}^d A_{k_i}(r_i)\Big\}\Big\{h\Delta^T M_1(x)+\frac{1}{2}h^2\Delta^T M_2(x)\Delta+o\big(h^2\big)\Big\}\\
&= \frac{1}{2}\mu_2(K)\,h^2\,tr\big\{BM_2(x)\big\}+h\,E_1+\frac{1}{2}\,h^2\,E_2+o_p\big(h^2\big)\,.
\end{aligned}
$$

where

$$
\begin{aligned}
E_1 &= \sum_{(k_1,\cdots,k_d)^T\in\{0,1,2\}^d}\Big\{\Pi_{i=1}^d A_{k_i}(r_i)\Big\}\Delta^T M_1(x)\,,\\
E_2 &= \sum_{(k_1,\cdots,k_d)^T\in\{0,1,2\}^d}\Big\{\Pi_{i=1}^d A_{k_i}(r_i)\Big\}\Delta^T M_2(x)\Delta\,.
\end{aligned}
$$

Then (2.5) follows if $E_1=E_2=0$ which can be validated by showing that

$$
\begin{aligned}
E_1 &= \sum_{(k_1,\cdots,k_d)^T\in\{0,1,2\}^d}\Big\{\Pi_{i=1}^d A_{k_i}(r_i)\Big\}\sum_{j=1}^d(k_j-1-r_j)\delta_j b_j\frac{\partial}{\partial x_j}m(x)\\
&= \sum_{(k_1,\cdots,k_{d-1})^T\in\{0,1,2\}^{d-1}}\Big\{\Pi_{i=1}^{d-1} A_{k_i}(r_i)\Big\}\Big\{\sum_{k_d\in\{0,1,2\}}A_{k_d}(r_d)\sum_{j=1}^d(k_j-1-r_j)\delta_j b_j\frac{\partial}{\partial x_j}m(x)\Big\}\\
&= \sum_{(k_1,\cdots,k_{d-1})^T\in\{0,1,2\}^{d-1}}\Big\{\Pi_{i=1}^{d-1} A_{k_i}(r_i)\Big\}\Big\{\sum_{j=1}^{d-1}(k_j-1-r_j)\delta_j b_j\frac{\partial}{\partial x_j}m(x)\Big\}\,,
\end{aligned}
$$

$$
\begin{aligned}
E_2 &= \sum_{(k_1,\cdots,k_d)^T\in\{0,1,2\}^d} \Big\{\Pi_{i=1}^d A_{k_i}(r_i)\Big\} \\
&\qquad \times\Big\{\sum_{j=1}^d\sum_{l=1}^d (k_j-1-r_j)\delta_j b_j (k_l-1-r_l)\delta_l b_l \frac{\partial^2}{\partial x_j\partial x_l} m(x)\Big\} \\
&= \sum_{(k_1,\cdots,k_{d-1})^T\in\{0,1,2\}^{d-1}} \Big\{\Pi_{i=1}^{d-1} A_{k_i}(r_i)\Big\}\Big\{\sum_{k_d\in\{0,1,2\}} A_{k_d}(r_d)(k_d-1-r_d)^2\delta_d^2 b_d^2 \frac{\partial^2}{\partial x_d^2} m(x) \\
&\quad + \sum_{k_d\in\{0,1,2\}} A_{k_d}(r_d)\sum_{j=1}^{d-1}\sum_{l=1}^{d-1}(k_j-1-r_j)\delta_j b_j(k_l-1-r_l)\delta_l b_l\frac{\partial^2}{\partial x_j\partial x_l}m(x) \\
&\quad + 2\sum_{k_d\in\{0,1,2\}} A_{k_d}(r_d)\sum_{l=1}^{d-1}(k_d-1-r_d)\delta_d b_d(k_l-1-r_l)\delta_l b_l\frac{\partial^2}{\partial x_d\partial x_l}m(x)\Big\} \\
&= \sum_{(k_1,\cdots,k_{d-1})^T\in\{0,1,2\}^{d-1}} \Big\{\Pi_{i=1}^{d-1} A_{k_i}(r_i)\Big\} \\
&\qquad \times\Big\{\sum_{j=1}^{d-1}\sum_{l=1}^{d-1}(k_j-1-r_j)\delta_j b_j(k_l-1-r_l)\delta_l b_l\frac{\partial^2}{\partial x_j\partial x_l}m(x)\Big\},
\end{aligned}
$$

and by induction.

To deal with the conditional variance, let $C^*(a,b)=\int K(s+a)K(s+b)\,ds$, $\Sigma = diag\big\{\nu(X_1),\cdots,\nu(X_n)\big\}$, $z=(\alpha_{1,2\nu_1+k_1},\cdots,\alpha_{d,2\nu_d+k_d})^T$ and $y=(\alpha_{1,2\nu_1+l_1},\cdots,\alpha_{d,2\nu_d+l_d})^T$, where $k_i,l_i\in\{0,1,2\}$ for $i=1,\cdots,d$. The covariance of $\widehat{m}(z)$ and $\widehat{m}(y)$ conditional on $X_1,\cdots,X_n$ is

$$
\begin{aligned}
&\mathrm{E}\Big[\big\{\widehat{m}(z)-\mathrm{E}\big(\widehat{m}(z)\big|X_1,\cdots,X_n\big)\big\}\big\{\widehat{m}(y)-\mathrm{E}\big(\widehat{m}(y)\big|X_1,\cdots,X_n\big)\big\}\Big|X_1,\cdots,X_n\Big] \\
=\;& e^T\big(X_z^TW_zX_z\big)^{-1}X_z^TW_z\Sigma W_yX_y\big(X_y^TW_yX_y\big)^{-1}e\,.
\end{aligned}
$$

Note that

$$
\begin{aligned}
&X_z^TW_z\Sigma W_yX_y \\
&= \sum_{i=1}^n \Pi_{j=1}^d\Big\{K\big(\frac{X_{ij}-z_j}{b_jh}\big)K\big(\frac{X_{ij}-y_j}{b_jh}\big)\Big\}\nu(X_i)\begin{pmatrix} 1 & (X_i-y)^T \\ (X_i-z) & (X_i-z)(X_i-y)^T\end{pmatrix} \\
&= \begin{pmatrix} I & II \\ III & IV\end{pmatrix},
\end{aligned}
$$

where

$$
\begin{aligned}
&\{\nu(x)f(x)\}^{-1}I \\
&= n\,\Pi_{j=1}^{d}\Big\{\int K\Big(\frac{s-\alpha_{j,2\nu_j+k_j}}{b_jh}\Big)K\Big(\frac{s-\alpha_{j,2\nu_j+l_j}}{b_jh}\Big)\,ds\Big\}\big\{1+o_p(1)\big\} \\
&= n\,\Pi_{j=1}^{d}\Big\{\int K\Big(\frac{s-x_j}{b_jh}+\frac{x_j-\alpha_{j,2\nu_j+k_j}}{b_jh}\Big)K\Big(\frac{s-x_j}{b_jh}+\frac{x_j-\alpha_{j,2\nu_j+l_j}}{b_jh}\Big)\,ds\Big\}\big\{1+o_p(1)\big\} \\
&= n\,\Pi_{j=1}^{d}\Big\{b_jh\int K\big(u+(1-k_j+r_j)\delta_j\big)K\big(u+(1-l_j+r_j)\delta_j\big)\,du\Big\}\big\{1+o_p(1)\big\} \\
&= nh^d\,\Pi_{j=1}^{d}\Big\{b_jC^*\big((1-k_j+r_j)\delta_j,(1-l_j+r_j)\delta_j\big)\Big\}\{1+o_p(1)\},
\end{aligned}
$$

every element in II and III is $O_p(nh^{d+1})$ and every element in IV is $O_p(nh^{d+2})$. Also, similar to Ruppert and Wand (1994), we can show that

$$
(X_z^TW_zX_z)^{-1}=\frac{1}{nh^d\Pi_{j=1}^{d}b_j}\begin{pmatrix}\big\{f(x)\big\}^{-1}+o_p(1) & -M_1(x)^T\big\{f(x)\big\}^{-2}+o_p(1)\\ -M_1(x)\big\{f(x)\big\}^{-2}+o_p(1) & \big\{\mu_2(K)f(x)H\big\}^{-1}+o_p(H^{-1})\end{pmatrix},
$$

$$
(X_y^TW_yX_y)^{-1}=\frac{1}{nh^d\Pi_{j=1}^{d}b_j}\begin{pmatrix}\big\{f(x)\big\}^{-1}+o_p(1) & -M_1(x)^T\big\{f(x)\big\}^{-2}+o_p(1)\\ -M_1(x)\big\{f(x)\big\}^{-2}+o_p(1) & \big\{\mu_2(K)f(x)H\big\}^{-1}+o_p(H^{-1})\end{pmatrix}.
$$

So

$$
\begin{aligned}
&\mathrm{E}\Big[\big\{\widehat{m}(z)-\mathrm{E}\big(\widehat{m}(z)\big|X_1,\cdots,X_n\big)\big\}\big\{\widehat{m}(y)-\mathrm{E}\big(\widehat{m}(y)\big|X_1,\cdots,X_n\big)\big\}\Big|X_1,\cdots,X_n\Big] \\
&= \frac{\nu(x)}{nh^df(x)\Pi_{j=1}^{d}b_j}\Pi_{j=1}^{d}\Big\{C^*\big((1-k_j+r_j)\delta_j,(1-l_j+r_j)\delta_j\big)\Big\}\big\{1+o_p(1)\big\}. \qquad (6.3)
\end{aligned}
$$

Further

$$
\begin{aligned}
&\mathrm{Var}\big\{\widetilde{m}(x;r,\delta)|X_1,\cdots,X_n\big\} \\
&= \frac{\nu(x)}{nh^df(x)\Pi_{j=1}^{d}b_j}\sum_{(k_1,\cdots,k_d)^T\in\{0,1,2\}^d}\ \sum_{(l_1,\cdots,l_d)^T\in\{0,1,2\}^d}\Big\{\Pi_{i=1}^{d}A_{k_i}(r_i)\Big\}\Big\{\Pi_{i=1}^{d}A_{l_i}(r_i)\Big\} \\
&\qquad\qquad\times\Big\{\Pi_{j=1}^{d}C^*\big((1-k_j+r_j)\delta_j,(1-l_j+r_j)\delta_j\big)\Big\}\big\{1+o_p(1)\big\} \\
&= \frac{\nu(x)}{nh^df(x)\Pi_{j=1}^{d}b_j}\sum_{(k_1,\cdots,k_{d-1})^T\in\{0,1,2\}^{d-1}}\ \sum_{(l_1,\cdots,l_{d-1})^T\in\{0,1,2\}^{d-1}}\Big\{\Pi_{i=1}^{d-1}A_{k_i}(r_i)\Big\}\Big\{\Pi_{i=1}^{d-1}A_{l_i}(r_i)\Big\} \\
&\qquad\qquad\times\Big\{\Pi_{j=1}^{d-1}C^*\big((1-k_j+r_j)\delta_j,(1-l_j+r_j)\delta_j\big)\Big\}V_1\big\{1+o_p(1)\big\},
\end{aligned}
$$

where

$$
\begin{aligned}
V_1 &= \sum_{k_d\in\{0,1,2\}} A_{k_d}(r_d) \sum_{l_d\in\{0,1,2\}} A_{l_d}(r_d)\, C^*\big((1-k_d+r_d)\delta_d, (1-l_d+r_d)\delta_d\big) \\
&= \frac{r_d^2(r_d-1)^2}{2^2}C(0,\delta_d) + (1-r_d^2)\frac{r_d(r_d-1)}{2}C\big(\frac{1}{2},\delta_d\big) + \frac{r_d(r_d+1)}{2}\frac{r_d(r_d-1)}{2}C(1,\delta_d) \\
&\quad + \frac{r_d(r_d-1)}{2}(1-r_d^2)C(\frac{1}{2},\delta_d) + (1-r_d^2)^2C(0,\delta_d) + \frac{r_d(r_d+1)}{2}(1-r_d^2)C(\frac{1}{2},\delta_d) \\
&\quad + \frac{r_d(r_d-1)}{2}\frac{r_d(r_d+1)}{2}C(1,\delta_d) + (1-r_d^2)\frac{r_d(r_d+1)}{2}C\big(\frac{1}{2},\delta_d\big) + \frac{r_d^2(r_d+1)^2}{2^2}C(0,\delta_d) \\
&= \frac{3r_d^4-3r_d^2+2}{2}C(0,\delta_d) + 2r_d^2(1-r_d^2)C(\frac{1}{2},\delta_d) + \frac{r_d^2(r_d^2-1)}{2}C(1,\delta_d) \\
&= R(K) - r_d^2(1-r_d^2)C(\delta_d)\,.
\end{aligned}
$$

Thus, by induction, we have

$$
\mathrm{Var}\big\{\widetilde{m}(x;r,\delta)\big|X_1,\cdots,X_n\big\} = \frac{\nu(x)}{nh^d f(x)\Pi_{j=1}^d b_j}\Pi_{j=1}^d\big\{R(K)-r_j^2(1-r_j^2)C(\delta_j)\big\}\big\{1+o_p(1)\big\}\,.
$$

Proof of Theorem 3. Note that x and $x^*_{k_1,\cdots,k_d}(x;r)$ in the definition of $\widetilde{m}_r(x;\delta)$ follows the relation $x^*_{k_1,\cdots,k_d}(x;r) = x+\big((k_1-1-r_1)\delta_1 b_1 h,\cdots,(k_d-1-r_d)\delta_d b_d h\big)^T = x+h\Delta$. In the definition of $\widetilde{m}(x;r,\delta)$, x and $x^*(k_1,\cdots,k_d)$ following a parallel relation $x^*(k_1,\cdots,k_d) = x+h\Delta$. Therefore the conditional bias result (2.10) follows from (2.5) and we only need to show (2.11). Write

$$
\begin{aligned}
\bar{m}(x;\delta) &= 2^{-d} \sum_{(r_1,\cdots,r_d)^T\in\{-1/\sqrt{2},1/\sqrt{2}\}^d} \sum_{(k_1,\cdots,k_d)^T\in\{0,1,2\}^d} \Big\{\Pi_{i=1}^d A_{k_i}(r_i)\Big\} \\
&\qquad \times\widehat{m}\big((x_1+(k_1-1-r_1)\delta_1 b_1 h,\cdots,x_d+(k_d-1-r_d)\delta_d b_d h)^T\big).
\end{aligned}
$$

Hence, using (6.3) we have

$$\begin{aligned}
\mathrm{Var}\big\{\bar{m}(x;\delta)|X_1,\cdots,X_n\big\} &= \frac{\nu(x)}{2^{2d}nh^d f(x)\Pi_{i=1}^d b_i} \sum_{(r_1,\cdots,r_d)^T\in\{-1/\sqrt{2},1/\sqrt{2}\}^d} \\
&\quad \sum_{(s_1,\cdots,s_d)^T\in\{-1/\sqrt{2},1/\sqrt{2}\}^d} \sum_{(k_1,\cdots,k_d)^T\in\{0,1,2\}^d} \sum_{(l_1,\cdots,l_d)^T\in\{0,1,2\}^d} \Big\{\Pi_{i=1}^d A_{k_i}(r_i)\Big\}\Big\{\Pi_{i=1}^d A_{l_i}(s_i)\Big\} \\
&\quad \times\Big\{\Pi_{i=1}^d C^*\big((1-k_i+r_i)\delta_i,(1-l_i+s_i)\delta_i\big)\Big\}\big\{1+o_p(1)\big\} \\
&= \frac{\nu(x)}{2^{2d}nh^d f(x)\Pi_{i=1}^d b_i} \sum_{(r_1,\cdots,r_{d-1})^T\in\{-1/\sqrt{2},1/\sqrt{2}\}^{d-1}} \sum_{(s_1,\cdots,s_{d-1})^T\in\{-1/\sqrt{2},1/\sqrt{2}\}^{d-1}} \\
&\quad \sum_{(k_1,\cdots,k_{d-1})^T\in\{0,1,2\}^{d-1}} \sum_{(l_1,\cdots,l_{d-1})^T\in\{0,1,2\}^{d-1}} \Big\{\Pi_{i=1}^{d-1} A_{k_i}(r_i)\Big\}\Big\{\Pi_{i=1}^{d-1} A_{l_i}(s_i)\Big\} \\
&\quad \times\Big\{\Pi_{i=1}^{d-1} C^*\big((1-k_i+r_i)\delta_i,(1-l_i+s_i)\delta_i\big)\Big\} V_2\big\{1+o_p(1)\big\}.
\end{aligned}$$

where

$$\begin{aligned}
V_2 &= \sum_{k_d\in\{0,1,2\}}\sum_{l_d\in\{0,1,2\}} A_{k_d}(-1/\sqrt{2})A_{l_d}(-1/\sqrt{2})C^*\big((1-k_d-1/\sqrt{2})\delta_d,(1-l_d-1/\sqrt{2})\delta_d\big) \\
&\quad + \sum_{k_d\in\{0,1,2\}}\sum_{l_d\in\{0,1,2\}} A_{k_d}(-1/\sqrt{2})A_{l_d}(1/\sqrt{2})C^*\big((1-k_d-1/\sqrt{2})\delta_d,(1-l_d+1/\sqrt{2})\delta_d\big) \\
&\quad + \sum_{k_d\in\{0,1,2\}}\sum_{l_d\in\{0,1,2\}} A_{k_d}(1/\sqrt{2})A_{l_d}(-1/\sqrt{2})C^*\big((1-k_d+1/\sqrt{2})\delta_d,(1-l_d-1/\sqrt{2})\delta_d\big) \\
&\quad + \sum_{k_d\in\{0,1,2\}}\sum_{l_d\in\{0,1,2\}} A_{k_d}(1/\sqrt{2})A_{l_d}(1/\sqrt{2})C^*\big((1-k_d+1/\sqrt{2})\delta_d,(1-l_d+1/\sqrt{2})\delta\big) \\
&= 4\big\{R(K)-C(\delta_d)/4-D(\delta_d)/2\big\}.
\end{aligned}$$

By induction, we can show (2.11).

Acknowledgments

We thank two referees and an associate editor for their helpful comments. Ming-Yen Cheng's research was partially supported by NSC grant NSC-92-2118-M-002-012 and Mathematics Division, National Center for Theoretical Sciences (Taipei). Liang Peng's research was supported by NSF grant DMS-04-03443.

References

[1] Cheng, M.-Y., Peng, L. and Wu, J.-S. (2005). Reducing variance in smoothing. *Technical Report.*

[2] Cleveland, W.S. (1979). Robust locally weighted regression and smoothing scatterplots. *J. Amer. Statist. Assoc.*, **74**, 829–836.

[3] Fan, J. (1993). Local linear regression smoothers and their minimax efficiencies. *Ann. Statist.*, **21**, 196–216.

[4] Fan, J. and Gijbels, I. (1992). Variable bandwidth and local linear regression smoothers. *Ann. Statist.*, **20**, 2008–2036.

[5] Fan, J. and Gijbels, I. (1996). *Local polynomial modelling and its applications*. Chapman & Hall: London.

[6] Harrison, D. and Rubinfeld, D.L. (1978). Hedonic prices and the demand for clean air. *J. Environ. Economics and Management,* **5**, 81 - 102.

[7] Ruppert, D. (1997) Empirical-bias bandwidths for local polynomial nonparametric regression and density estimation. *J. Amer. Statist. Assoc.*, **92**, 1049–1062.

[8] Ruppert, D. and Wand, M.P. (1994). Multivariate locally weighted least squares regression. *Ann. Statist.* **22**, 1346–1370.

[9] Simonoff, J.S. (1996). *Smoothing methods in statistics*. Springer-Verlag: New York.

[10] Stone, C.J. (1977). Consistent nonparametric regression. *Ann. Statist.*, **5**, 595–645.

[11] Yang, L. and Tschernig, R. (1999). Multivariate bandwidth selection for local linear regression. *J. Roy. Statist. Soc. Ser. B*, **61**, 793–815.

[12] Wand, M.P. and Jones, M.C. (1995). *Kernel Smoothing*. Chapman & Hall: London.

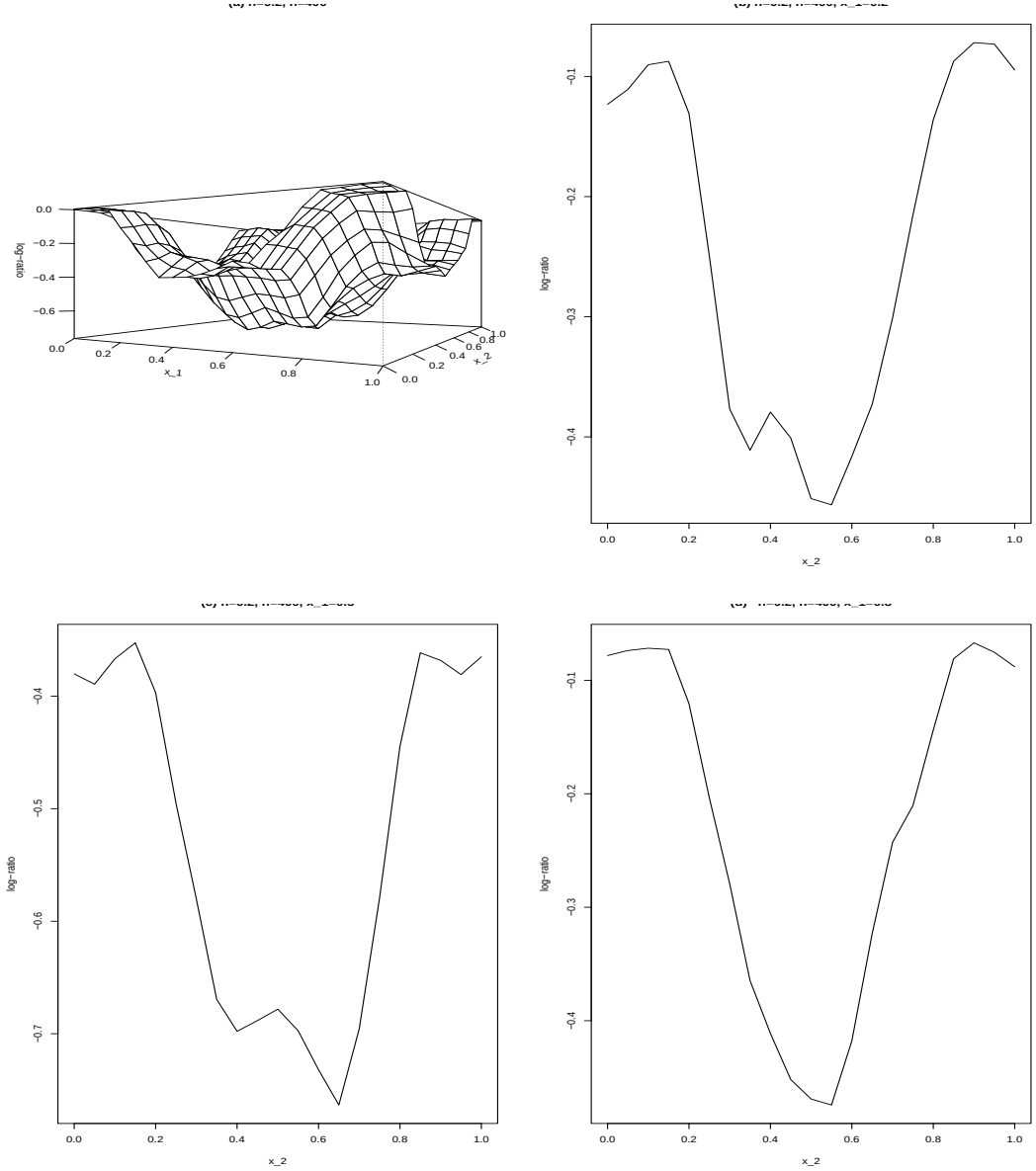

Figure 1: *Log-ratios with* $n = 400$ *and* $h = 0.2$. The natural logarithm of ratio of the mean squared error of $\bar{m}(x_1, x_2; \delta)$ to that of $\hat{m}(x_1, x_2)$ is plotted against $x_1 = 0, 0.05, \cdots, 1$ and $x_2 = 0, 0.05, \cdots, 1$, and also against $x_2 = 0, 0.05, \cdots, 1$, but with fixed $x_1 = 0.2, 0.5, 0.8$.

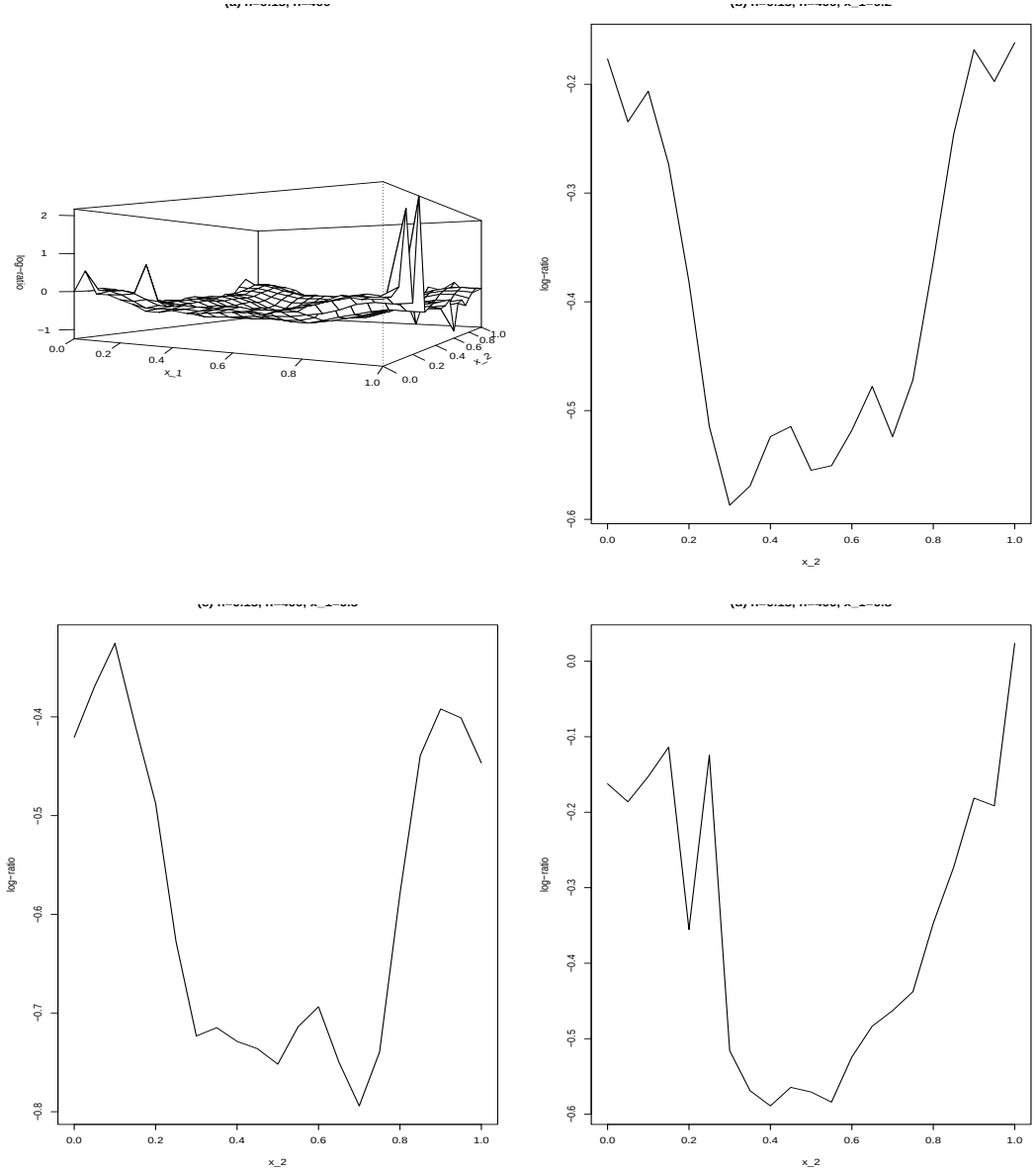

Figure 2: *Log-ratios with* $n = 400$ *and* $h = 0.15$. The natural logarithm of ratio of the mean squared error of $\bar{m}(x_1, x_2; \delta)$ to that of $\hat{m}(x_1, x_2)$ is plotted against $x_1 = 0, 0.05, \cdots, 1$ and $x_2 = 0, 0.05, \cdots, 1$, and also against $x_2 = 0, 0.05, \cdots, 1$, but with fixed $x_1 = 0.2, 0.5, 0.8$.

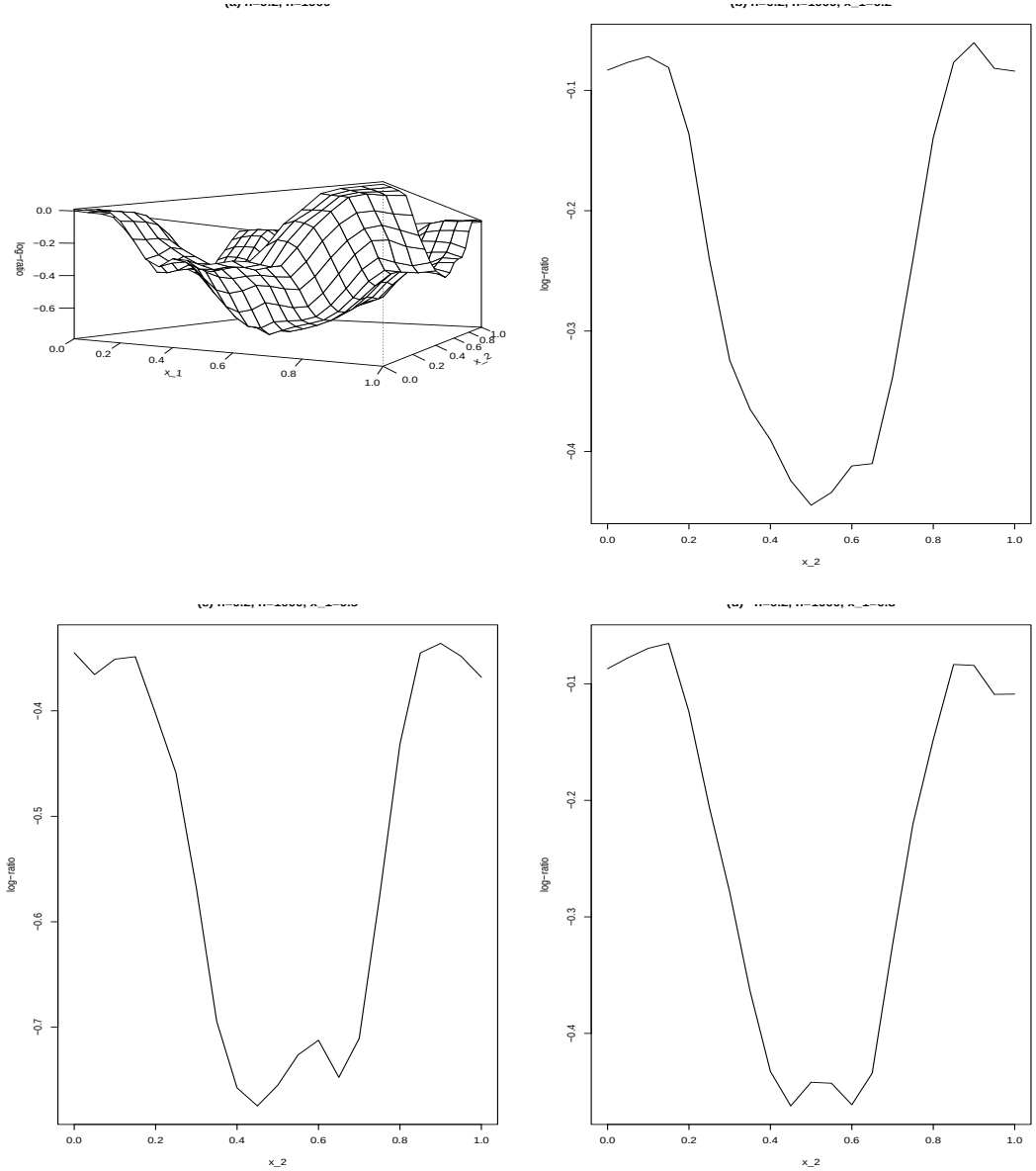

Figure 3: *Log-ratios with* $n = 1000$ *and* $h = 0.2$. The natural logarithm of ratio of the mean squared error of $\bar{m}(x_1, x_2; \delta)$ to that of $\hat{m}(x_1, x_2)$ is plotted against $x_1 = 0, 0.05, \cdots, 1$ and $x_2 = 0, 0.05, \cdots, 1$, and also against $x_2 = 0, 0.05, \cdots, 1$, but with fixed $x_1 = 0.2, 0.5, 0.8$.

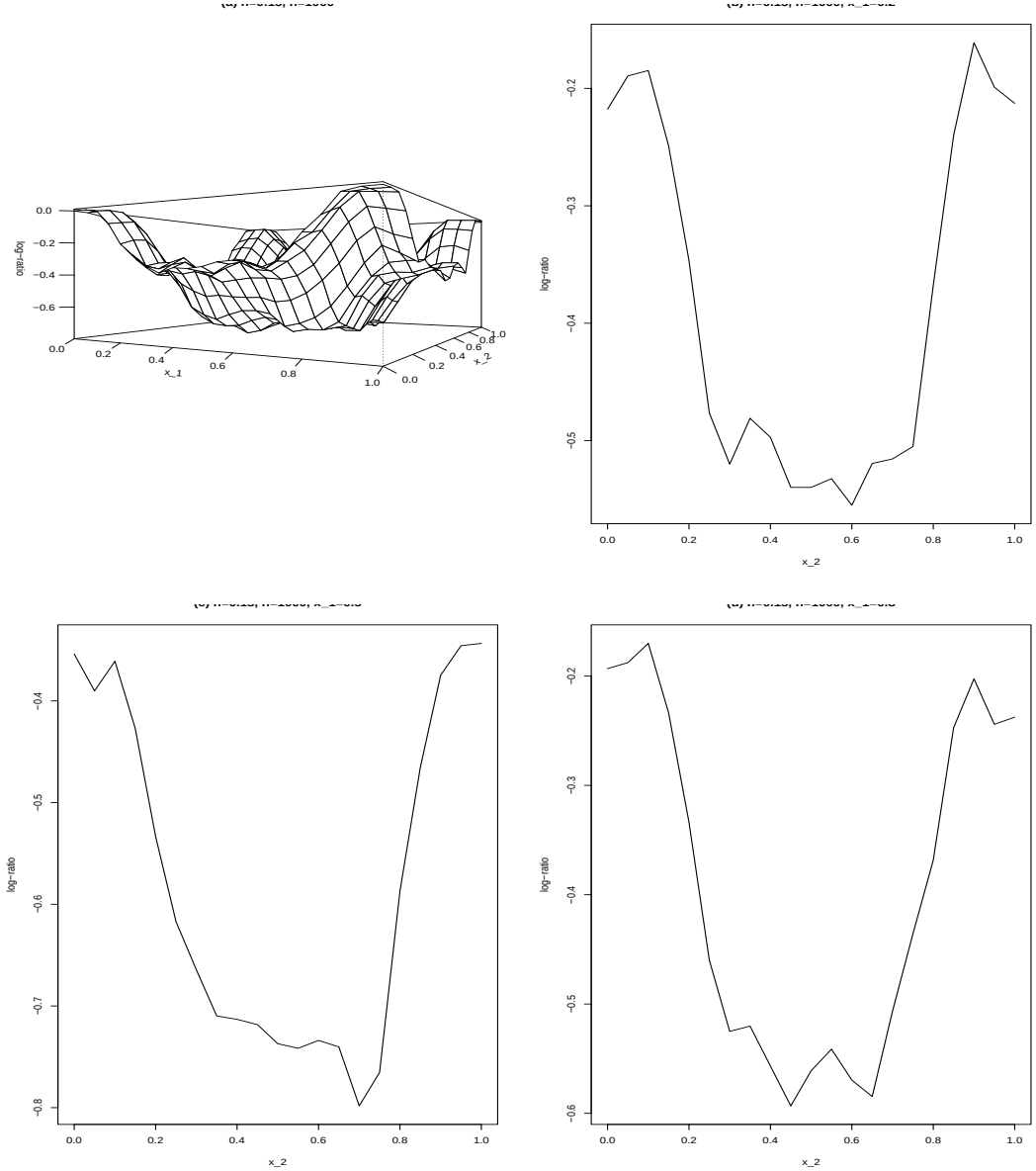

Figure 4: *Log-ratios with* $n = 1000$ *and* $h = 0.15$. The natural logarithm of ratio of the mean squared error of $\bar{m}(x_1, x_2; \delta)$ to that of $\hat{m}(x_1, x_2)$ is plotted against $x_1 = 0, 0.05, \cdots, 1$ and $x_2 = 0, 0.05, \cdots, 1$, and also against $x_2 = 0, 0.05, \cdots, 1$, but with fixed $x_1 = 0.2, 0.5, 0.8$.

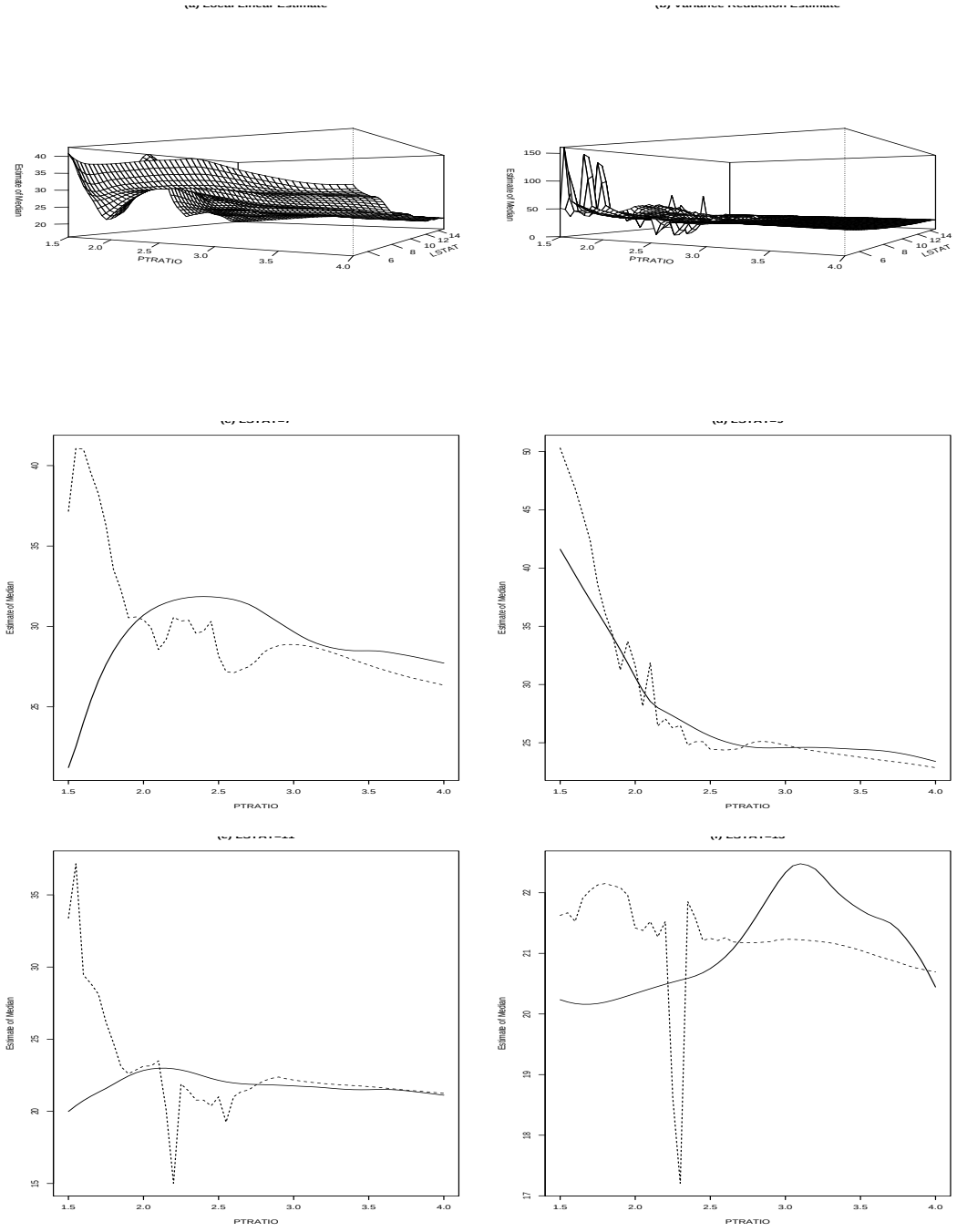

Figure 5: *Analysis of Boston housing price data set.* Panels (a) and (b) are respectively the local linear estimate $\hat{m}(\cdot)$ and the variance reduction estimate $\bar{m}(\cdot;\delta)$. Panels (c)–(f) plot $\hat{m}(x_1, x_2)$ (solid line) and $\bar{m}(x_1, x_2; \delta)$ (dotted line) against x_2 when $x_1 = 7, 9, 11, 13$, respectively.