TAIWANESE JOURNAL OF MATHEMATICS
Vol. 10, No. 6, pp. 1465-1483, December 2006
This paper is available online at http://www.math.nthu.edu.tw/tjm/

ON STUDY OF A MODIFIED LOCAL CONSTANT M-SMOOTHER

Jung-Huei Lin, Tzu-Kuei Chang and Chih-Kang Chu

Abstract. In the case of the equally spaced fixed design nonparametric regression, the local constant M-smoother (LCM) with local maximizing is proposed by Chu, Glad, Godtliebsen, and Marron (1998) to correct for the effect of discontinuity on the kernel regression estimator. It has the interesting property of jump-preserving. However, in the jump region, it is inconsistent when the magnitude of the noise is larger than the size of the jump in the regression. To adjust for this drawback to the ordinary LCM, we propose to construct the LCM with global maximizing instead of local maximizing as well as with binned data instead of original data. Our proposed estimator is analyzed by the asymptotic mean square error. Both binning and global maximizing have no effect on the asymptotic mean square error of the ordinary LCM in the smooth region, but have an effect on improving the inconsistency of the ordinary LCM in the jump region. Simulation studies demonstrate that the regression function estimate produced by our modified LCM is better than those by alternatives, in the sense of yielding smaller sample mean integrated square error, showing more accurately the location of jump point, and having smoother appearance.

1. Introduction

Nonparametric regression is a smoothing method for recovering the regression function from noisy data. Due to simplicity of computation and explanation, the kernel regression estimator is one of the most widely used smoothers. For a detailed introduction and asymptotic properties of the kernel regression estimator, see, for example, the monographs by Eubank (1988), Müller (1988), Härdle (1990, 1991), Scott (1992), Wand and Jones (1995), Fan and Gijbels (1996), and Simonoff (1996).

However, in application of the kernel regression estimator, the underlying regression function may or may not have discontinuity points. For example, consider

Received January 12, 2005; revised June 16, 2005.
Communicated by Yuh-Jia Lee.
2000 *Mathematics Subject Classification*: Primary 62G05; Secondary 62G20.
Key words and phrases: Asymptotic mean square error, Binning, Jump-preserving smoothing, Kernel regression estimator, Local constant M-smoother, Nonparametric regression.

the cases of studying the impact of advertising, the effect of medicine, and the influence of sudden changes in government policies and international relationships. If the regression function is affected instantly by such actions, then the resulting regression function has discontinuity points; otherwise it does not. See, for example, Shiau (1985) for many interesting examples where the regression functions are discontinuous. If the regression function has discontinuity points, then the kernel regression estimator is inadvisable, because smoothing tends to blur the jumps. For a detailed introduction of the effect of discontinuity points on the kernel regression estimator, see, for example, Wu and Chu (1993a, 1993b) and the references therein.

To improve the adverse effect of discontinuity points on the kernel regression estimator, several remedies are available. For example, it can be improved by using locally varying bandwidths (Fan and Gijbels 1995), by employing robust kernel estimators (Härdle and Gasser 1984; Tsybakov 1986; Besl, Birch, and Watson 1989; Fan, Hu, and Truong 1994; Fan and Jiang 1999), and by applying diagnostic kernel regression estimators (McDonald and Owen 1986; Hall and Titterington 1992; Qiu 2003). To perform these kernel regression estimators dealing with discontinuity, we do not need to know in advance that the regression function under study is discontinuous. However, due to the essence of local average and the possibility that the magnitude of the noise is large relative to the jump size of the regression function, estimators corresponding to the first two remedies still might be inconsistent in the jump region. Also, by the randomness of the diagnostic result, regression function estimates produced by estimators associated to the third remedy always have rough appearance. For these facts, see, for example, Figures 2-5 in Section 4. On the other hand, when the regression function is known to be discontinuous, kernel smoothing methods that produce discontinuous output have been proposed by Müller (1992) and Wu and Chu (1993b) for the one dimensional regression function, and by Qiu (1998) for the two dimensional regression function. To perform these kernel smoothing methods, the numbers of jump points and jump curves of the one and the two dimensional regression functions, respectively, need further to be known. However, in practice, it is not easy to obtain the correct information about these numbers.

Without assuming the discontinuity of the regression function, Chu, Glad, Godtliebsen, and Marron (1998) propose the local constant M-smoother (LCM). It is constructed by applying the idea of redescending M-estimation to local constant fits, and by exploiting the controversial local maximizer. For introduction of robust M-estimation, see, for example, the monographs by Huber (1981) and Hampel, Ronchetti, Rousseeuw, and Stahel (1986). Such LCM has the interesting property of jump-preserving. But it is weak in terms of efficiency of noise reduction, and suffers from inconsistency in the jump region when the magnitude of the noise is larger than the size of the jump in the regression. There are twofold distinctions between the LCM and the robust kernel estimator. Firstly, the former method is constructed by choosing the local maximizer; but the latter approach is developed by finding the global maximizer. Secondly, the redescending function employed by the former method is usually taken as the Gaussian function with mean 0 and standard deviation $g = g_n$ depending on the sample size n (see for example, Chapter 4 of Silverman 1986 for discussion and references concerning such choice); but that used by the latter approach is the Gaussian function with mean 0 and standard deviation 1 not depending on the sample size n.

The purpose of this article is to propose a modified LCM to improve the inconsistency of the ordinary LCM in the jump region. Our modified LCM is constructed by two stages. In the first stage, in order to reduce the magnitude of the noise of the data to deal with, the original data are transformed into the binned data (Härdle 1990). In the second stage, the LCM with global maximizing instead of local maximizing is applied to smooth the binned data obtained in the first stage. It is shown that both binning and global maximizing have no effect on the asymptotic mean square error (AMSE) of the ordinary LCM in the smooth region, but has an effect on improving the inconsistency of the ordinary LCM in the jump region. The latter effect is caused by the fact that the relative magnitude of the noise to the jump size in the binned data decreases as the bin size increases. Simulation studies demonstrate that the regression function estimate produced by our modified LCM is better than those by alternatives, in the sense of yielding smaller sample mean integrated square error, showing more accurately the location of jump point, and having smoother appearance.

This article is organized as follows. Section 2 describes the regression settings and the precise formulation of the modified LCM. The asymptotic behavior of our suggested estimator is studied in Section 3. Simulation results which show the finite sample behavior of the proposed estimator are contained in Section 4. Finally, proofs of the main theoretical results are given in Section 5.

2. Regression Settings and Proposed Estimators

In this paper, the equally spaced fixed design nonparametric regression model is considered. The regression model is given by

$$Y_i = m(x_i) + \epsilon_i, \tag{2.1}$$

for $i = 1, \cdots, n$. Here m is an unknown regression function defined on the closed interval [0,1], x_i are equally spaced fixed design points, that is, $x_i = i/n$, ϵ_i are independent and identically distributed regression errors with mean 0 and variance σ^2, $0 < \sigma^2 < \infty$, and Y_i are noisy observations of m at x_i.

The regression function m in (2.1) is defined by

$$m(x) = \mu(x) + d \cdot I_{[t,1]}(x), \tag{2.2}$$

for $x \in [0,1]$, where μ is a continuous function defined on [0,1], t is the location of jump point, d is the jump size of m at t, and I is an indicator function. To avoid confounding the issues studied here with boundary problems, assume that $t \in [\delta, 1-\delta]$. Here and throughout this paper, δ is an arbitrarily small positive constant. For simplicity of presentation, assume that the number of jump point of m is 1. An important special case is $d = 0$, that is, m is a continuous regression function.

The purpose of this paper is to use the observations Y_i to discover the value of $m(x)$, for $x \in [0,1]$. For this, a modified LCM is considered. It is constructed by two stages. Firstly, in the pre-binning stage, the original data (x_i, Y_i) are transformed into the binned data on equally spaced partitioned points of [0,1]. For this, give the bin size $2b$, where $b = b_n \to 0$ with $nb \to \infty$ as $n \to \infty$, and let $p_j = (2j-1)b$ denote the center of the j-th bin, the interval $[p_j - b, p_j + b)$, and $\bar{Y}_j$ the binned data obtained at p_j by averaging the observations Y_i occurred in the j-th bin, for each $j = 1, \cdots, q$. Here $q = [(2b)^{-1}]$. The notation $[x]$ denotes the integer part of x. Specifically, $\bar{Y}_j$ is defined by

$$\bar{Y}_j = [\sum_{i=1}^{n} W\{(p_j - x_i)/b\}Y_i]/[\sum_{i=1}^{n} W\{(p_j - x_i)/b\}],$$

where W is the uniform kernel function defined by $W(z) = (1/2)I_{[-1,1]}(z)$. Under some regularity conditions, the variance of our binned data is roughly equal to $(2nb)^{-1}\sigma^2$, and is of smaller order in magnitude than that of the original data. Hence binning has an effect on reducing the magnitude of noise of the data. If 1 is not a multiple of $2b$, then the observations Y_i with design points $x_i \in [0, 2bq]$ are used to construct the binned data, and the rest of the observations are dropped out. For simplicity of notation, assume that 1 is a multiple of $2b$.

Secondly, in the smoothing stage, the modified LCM is employed to smooth the binned data $(p_j, \bar{Y}_j)$. Give the kernel function K as a symmetric probability density function supported on the interval [-1,1], L as the Gaussian function, and both bandwidths $h = h_n$ and $g = g_n$ tending to 0 as $n \to \infty$, but the bandwidth h is of larger order than the bin size $2b$ in magnitude. Our modified LCM $\hat{m}_{MOD}(x)$ for $m(x)$ is constructed by applying both the local constant fit for the binned data to the "inside" of the kernel function L and the local linear weighting scheme to the "outside" of L, and by taking the global maximizer of the resulting M function

$$S_{MOD}(\xi; x) = \sum_{j=1}^{q} \alpha_j(x) L_g(\bar{Y}_j - \xi) \tag{2.3}$$

over ξ, for each $x \in [0,1]$. Here $L_g(\cdot) = g^{-1}L(\cdot/g)$, the Gaussian function with mean 0 and standard deviation g. For the reason of taking L as the Gaussian function, see Chu et al. (1998). To avoid boundary effects (Müller 1988) on the ordinary LCM, the weights $\alpha_j(x)$ in (2.3) are generated by applying the local linear fit to the "outside" of the kernel function L, as Hwang (2002) suggests. These weights $\alpha_j(x)$ are taken as those assigned to the observations by the local linear smoother in Fan (1992, 1993). Specifically,

$$\alpha_j(x) = \{S_2(x) - (x - p_j)S_1(x)\}K_h(x - p_j)/\{S_0(x)S_2(x) - S_1^2(x)\},$$

where $K_h(\cdot) = h^{-1}K(\cdot/h)$ and $S_k(x) = \sum_{j=1}^{q}(x - p_j)^k K_h(x - p_j)$, for $k \geq 0$. The asymptotic behavior of the proposed estimator $\hat{m}_{MOD}(x)$ will be studied in Section 3.

If $\alpha_j(x)$, $(p_j, \bar{Y}_j)$, and q in (2.3) are replaced respectively with $K_h(x_i - x_j)$, (x_j, Y_j), and n, then the local maximizer of the corresponding M function closest to the observation Y_i is the ordinary LCM for $m(x_i)$. Further, if the value of g in (2.3) is taken as $g = 1$, then the global maximizer of the corresponding M function is the robust kernel estimator (Härdle and Gasser 1984) for $m(x_i)$. It is known that these two estimators suffer from boundary effects, and the drawback can be simply improved by using the local linear weighting scheme as we do in (2.3). Such resulting estimators free from boundary effects are the boundary modified LCM in Hwang (2002) and the boundary modified robust kernel estimator, and are denoted by $\hat{m}_{LCM}(x_i)$ and $\hat{m}_{RBT}(x_i)$, respectively, in each case. They might be inconsistent in the jump region, and their finite sample performance will be compared with that of $\hat{m}_{MOD}(x_i)$ in Section 4. On the other hand, for improving boundary effects on the ordinary LCM, Rue, Chu, Godtliebsen, and Marron (2002) propose the local linear M-smoother by applying the local linear fitting inside of the kernel function L. However, the local linear M-smoother still suffers from the inconsistency in the jump region. Due to the fact of using an extra tuning parameter, it also suffers from the sensitivity to random fluctuations (Hwang 2002).

The motivation of our modified LCM is now given. It is illustrated in the following Figure 1. By the formulation of $S_{MOD}(\xi; x)$ in (2.3), it is essentially a kernel density estimate using only the binned data $(p_j, \bar{Y}_j)$ occurred in the compact window $[x - h, x + h]$. See, for example, the monograph by Silverman (1986) for a detailed introduction of the kernel density estimator. If x is in the smooth region, then, by some regularity conditions, the observations $\bar{Y}_j$ employed by $S_{MOD}(\xi; x)$ have expectations close to $m(x)$. Hence, the corresponding $S_{MOD}(\xi; x)$ has a single global maximizer located around $m(x)$, asymptotically. The smaller the magnitude of the noise in the binned data, the sharper the peak located around $m(x)$ shown by $S_{MOD}(\xi; x)$. On the other hand, if x is in the jump region, then, by similar arguments, $S_{MOD}(\xi; x)$ is a mixture density estimate having two local maximizers located around $\mu(x)$ and $\mu(x) + d$, asymptotically. In this case, if $x < t$, then the global maximizer of $S_{MOD}(\xi; x)$ is located around $\mu(x)$ since the number of the binned data used by $S_{MOD}(\xi; x)$ having expectations approaching $\mu(x)$ is larger

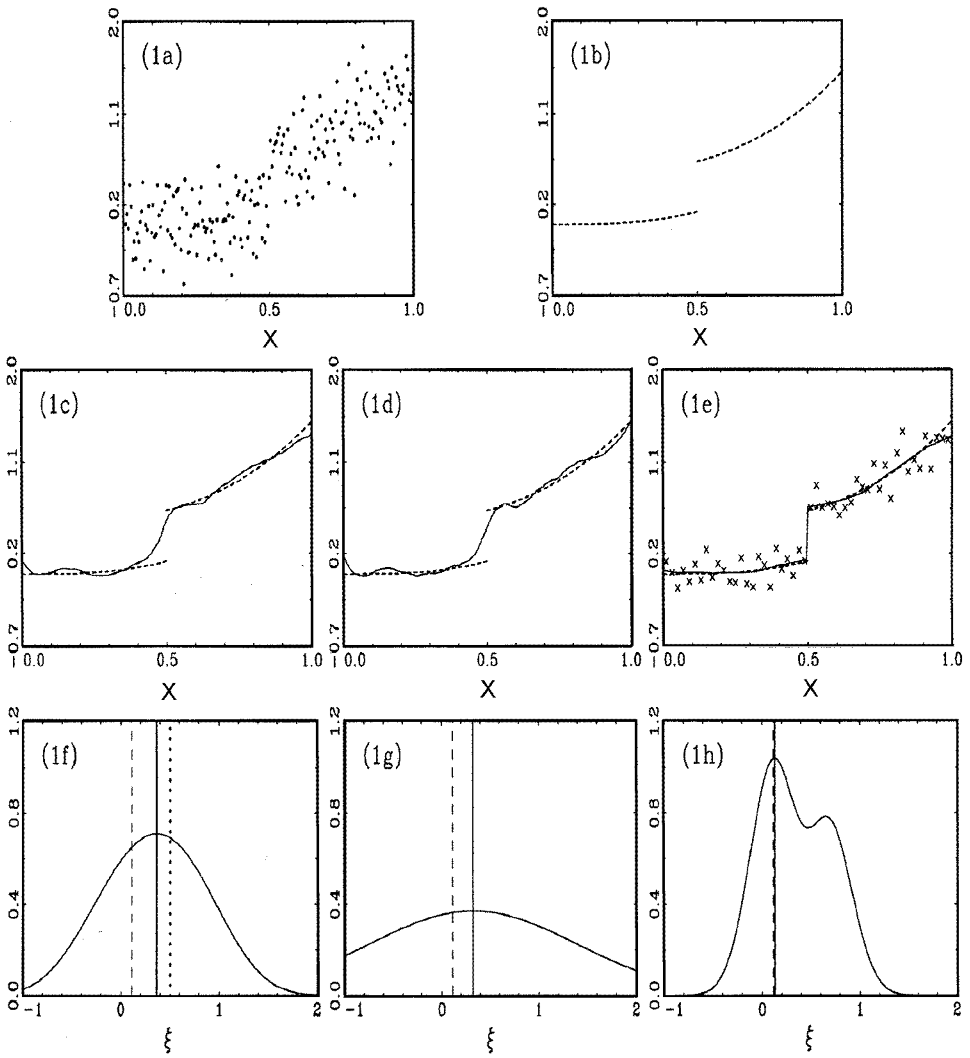

Fig. 1. An artificial example. (1a) One simulated data set of sample size $n = 200$. (1b) The true regression function. (1c) Best performance of $\hat{m}_{LCM}(x)$ using bandwidths $h = 0.0949$ and $g = 0.375$. (1d) Best performance of $\hat{m}_{RBT}(x)$ using $h = 0.0720$ and $g = 1$. (1e) Best performance of $\hat{m}_{MOD}(x)$ using $b = 0.0199, h = 0.157$, and $g = 0.203$. (1f) The M function $S_{LCM}(\xi; x = 0.48)$ for $\hat{m}_{LCM}(0.48)$ in (1c). (1g) The M function $S_{RBT}(\xi; x = 0.48)$ for $\hat{m}_{RBT}(0.48)$ in (1d). (1h) The M function $S_{MOD}(\xi; x = 0.48)$ for $\hat{m}_{MOD}(0.48)$ in (1e).

than that converging to $\mu(x) + d$. Similar remark can be applied to the case that $x > t$. By these arguments, the global maximizer of $S_{MOD}(\xi; x)$ over ξ is a reasonable estimator of $m(x)$, for each $x \in [0, 1]$.

We now close this section by giving an illustration of our modified LCM. For this, a simulation study was performed and the results are given in Figure 1. The simulation settings will be introduced in Section 4. It is difficult to distinguish visually from the simulated data in Figure 1a alone that the underlying regression function (dashed curves) in Figure 1b has a jump point at $x = 0.5$. Using the data in Figure 1a, Figures 1c-1e present respectively the best performance of the regression function estimate (solid curve) produced by each of $\hat{m}_{LCM}(x)$, $\hat{m}_{RBT}(x)$, and $\hat{m}_{MOD}(x)$. The values of the smoothing parameters used by the three estimators were taken as the minimizers of their sample integrated square errors. Under our simulation setting, the best performance of our modified LCM is better than that of the other two estimators, in the sense of yielding smaller sample integrated square error and showing more accurately the location of the jump point.

Figure 1f shows the disappearance of the bimodal structure of the mixture density estimate $S_{LCM}(\xi; x = 0.48)$ (solid curve), that is, the M function corresponding to $\hat{m}_{LCM}(0.48)$ in the jump region in Figure 1c. Such result explains the poor performance of $\hat{m}_{LCM}(0.48)$, and is due to the large relative magnitude of the noise in the original data to the jump size. The same remark applies to the mixture density estimate $S_{RBT}(\xi; x = 0.48)$ in Figure 1g for $\hat{m}_{RBT}(0.48)$ in Figure 1d. On the other hand, the binned data $(p_j, \bar{Y}_j)$ (stars) in Figure 1h produced by using the original data in Figure 1a show the small relative magnitude of their noise to the jump size. Their corresponding mixture density estimate $S_{MOD}(\xi; x = 0.48)$ (solid curve in Figure 1h) for $\hat{m}_{MOD}(0.48)$ in Figure 1e shows correctly a bimodal structure, and its global maximizer coincides to the value of $m(0.48)$. Comparing Figures 1f-1h, it can be seen that both binning and global maximizing have an effect on improving the inconsistency of $\hat{m}_{LCM}(x)$ in the jump region. The locations of the dashed and the solid vertical lines in each of Figures 1f-1h denote respectively the value of $m(0.48)$ and its corresponding estimate, and that of the dotted vertical line in Figure 1f stands for the value of Y_i in Figure 1a at the design point $x_i = 0.48$.

3. Theoretical Results

In this section, we shall study the asymptotic behavior of $\hat{m}_{MOD}(x)$. For this, in addition to the assumptions given in Section 2, we add the following ones:

(A1) The function μ defined on the interval [0,1] has two Lipschitz continuous derivatives.

(A2) The kernel function K is a symmetric and Lipschitz continuous probability density function supported on [-1,1], and L is the Gaussian function.

(A3) The regression errors ϵ_i are independent and identically distributed random variables with mean 0 and variance σ^2, $0 < \sigma^2 < \infty$.

(A4) The values of b, h, and g are selected on the interval $[\delta \cdot n^{-1+\delta}, \delta^{-1} \cdot n^{-\delta}]$, where δ is an arbitrarily small positive constant. Also, they satisfy the conditions $h^2 >> b$ and $g^2 >> h + n^{-1}b^{-1} + nbh^4 + n^{-3/2}b^{-3/2}h^{-2}$. The notation $\alpha_n >> \beta_n$ denotes $\beta_n/\alpha_n \to 0$, as $n \to \infty$.

(A5) The total number of observations in this regression setting is n, with $n \to \infty$.

The following Theorem 3.1 gives the asymptotic bias and variance of $\hat{m}_{MOD}(x)$. Its proof is given in Section 5. To state this theorem, we introduce the following notation. Decompose the interval [0,1] into the left boundary region R_1, the right boundary region R_2, the region R_3 just to the left of the jump point, the region R_4 just to the right of the jump point, and the smooth region R_5. Specifically,

$$R_1 = \{x : x = \lambda h, \lambda \in [0, 1]\}, \quad R_2 = \{x : x = 1 + \lambda h, \lambda \in [-1, 0]\},$$

$$R_3 = \{x : x = t + \lambda h, \lambda \in [-1, 0)\}, \quad R_4 = \{x : x = t + \lambda h, \lambda \in [0, 1]\},$$

$$R_5 = \{x : x \in (h, t - h) \cup (t + h, 1 - h)\}.$$

Set

$$\kappa_j = \int_{-1}^{1} z^j K(z)dz, \quad \kappa_{\ell,j} = \int_{-1}^{\lambda} z^j K(z)dz,$$

$$\tau_j = \int_{-1}^{1} z^j K(z)^2 dz, \quad \tau_{\ell,j} = \int_{-1}^{\lambda} z^j K(z)^2 dz,$$

$$D_{\ell,2} = \kappa_{\ell,0}\kappa_{\ell,2} - \kappa_{\ell,1}^2, \quad D_{\ell,3} = \kappa_{\ell,2}^2 - \kappa_{\ell,1}\kappa_{\ell,3},$$

$$N_{\ell,2} = \kappa_{\ell,2}^2\tau_{\ell,0} - 2\ \kappa_{\ell,1}\kappa_{\ell,2}\tau_{\ell,1} + \kappa_{\ell,1}^2\tau_{\ell,2},$$

for $j \geq 0$. Let $\kappa_{r,j}$ and $\tau_{r,j}$ be $\kappa_{\ell,j}$ and $\tau_{\ell,j}$ with the integration interval $[-1, \lambda]$ replaced by $[\lambda, 1]$, and $D_{r,2}$, $D_{r,3}$, and $N_{r,2}$ be defined similarly to $D_{\ell,2}$, $D_{\ell,3}$, and $N_{\ell,2}$, respectively, in each case.

Define quantities related to asymptotic biases and variances:

$$b_0 = \kappa_2, \quad b_{\ell,1} = \kappa_{\ell,0}^{-1}\kappa_{\ell,1}, \quad b_{\ell,2} = D_{\ell,2}^{-1}D_{\ell,3}, \quad b_{r,1} = \kappa_{r,0}^{-1}\kappa_{r,1}, \quad b_{r,2} = D_{r,2}^{-1}D_{r,3},$$

$$v_0 = \tau_0, \quad v_{\ell,1} = \kappa_{\ell,0}^{-2}\tau_{\ell,0}, \quad v_{\ell,2} = D_{\ell,2}^{-2}N_{\ell,2}, \quad v_{r,1} = \kappa_{r,0}^{-2}\tau_{r,0}, \quad v_{r,2} = D_{r,2}^{-2}N_{r,2}.$$

Theorem 3.1. *Suppose that the regression model (2.1) with $d \neq 0$ and the assumptions (A1)-(A5) hold. For x in each R_1, R_2, R_3, R_4, and R_5, the dominant terms of asymptotic bias and variance of $\hat{m}_{MOD}(x)$ can be expressed respectively by*

$$Bias\{\hat{m}_{MOD}(x)\} = (1/2)h^2\mu^{(2)}(x)b_{\ell,2}, \quad (1/2)h^2\mu^{(2)}(x)b_{r,2}, \quad (-1)h\mu^{(1)}(x)b_{r,1},$$

$$(3.1) \qquad (-1)h\mu^{(1)}(x)b_{\ell,1}, \quad (1/2)h^2\mu^{(2)}(x)b_0,$$

$$Var\{\hat{m}_{MOD}(x)\} = n^{-1}h^{-1}\sigma^2 v_{\ell,2}, \quad n^{-1}h^{-1}\sigma^2 v_{r,2}, \quad n^{-1}h^{-1}\sigma^2 v_{r,1},$$

$$(3.2) \qquad n^{-1}h^{-1}\sigma^2 v_{\ell,1}, \quad n^{-1}h^{-1}\sigma^2 v_0.$$

On the other hand, if $d = 0$, the results in (3.1) and (3.2) still hold by dropping out the cases R_3 and R_4.

We now close this section with the following remarks.

Remark 3.1. {The asymptotic properties of $\hat{m}_{MOD}(x)$} By (3.1) and (3.2), the bin size $2b$, the jump size d, where $d \neq 0$, and the Gaussian function L are not related to the dominant term of AMSE of $\hat{m}_{MOD}(x)$, for each $x \in [0, 1]$. However, the discontinuity of the regression function has an effect on increasing the order of asymptotic bias of $\hat{m}_{MOD}(x)$, but has no effect on the order of asymptotic variance of $\hat{m}_{MOD}(x)$, for x in the jump region $[t - h, t + h]$. By Theorem 3.1 of this paper and Theorem 4.1 of Ruppert and Wand (1994), for $x \notin [t - h, t + h]$, our $\hat{m}_{MOD}(x)$ has the same AMSE as the local linear smoother in Fan (1992, 1993). For such x away from jump point, in the sense of having smaller AMSE, using (5.7), (5.8), and Theorem 8 of Fan, Gasser, Gijbels, Brockmann, and Engel (1993), the optimal kernel function K satisfying the conditions given in (A2) for constructing $\hat{m}_{MOD}(x)$ is the Epanechnikov kernel $K(z) = (3/4)(1 - z^2)I_{[-1,1]}(z)$.

Remark 3.2. {Practical choice of the values of b, g, and h for constructing $\hat{m}_{MOD}(x)$} For constructing $\hat{m}_{MOD}(x)$, we suggest using the cross-validation criterion to choose the value of (b, h, g). The selected value of (b, h, g) is taken as the minimizer $(\hat{b}, \hat{h}, \hat{g})$ of the cross-validation score $CV_{MOD}(b, h, g)$ defined by

$$CV_{MOD}(b, h, g) = \sum_{i=1}^{n} \{\hat{m}_{MOD,i}(x_i) - Y_i\}^2;$$

see Section 5.1 of Härdle (1990). Here $\hat{m}_{MOD,i}(x_i)$ is the "leave-one-out" version of $\hat{m}_{MOD}(x_i)$, that is, the observation (x_i, Y_i) is left out in constructing $\hat{m}_{MOD}(x_i)$. For other automatic bandwidth selection methods, see also Härdle and Marron (1985), Rice (1984), and Marron (1988).

Remark 3.3. {A direct extension of $\hat{m}_{MOD}(x)$ to the random design data} By the suggestion of the reviewer of the paper, the result of this paper can be applied to the random design regression model. For this, the random design regression model is given by $Z_i = r(U_i) + e_i$, for $i = 1, \cdots, n$. Here the regression function $r(x)$ is the same as that $m(x)$ in (2.2), for each $x \in [0, 1]$, the random design data (U_i, Z_i) are independent and identically distributed bivariate random vectors, and the regression errors e_i are assumed to have mean 0 and variance σ^2, $0 < \sigma^2 < \infty$. The design points U_i are assumed to be independent of the regression errors e_i, and are assumed to have the probability density function supported on the bounded interval $[0, 1]$. The modified LCM $\hat{r}_{MOD}(x)$ for $r(x)$ is similarly defined as $\hat{m}_{MOD}(x)$ with the fixed design data (x_i, Y_i) replaced by the random design data (U_i, Z_i). The performance of $\hat{r}_{MOD}(x)$, for $x \in [0, 1]$, needs further study. In practice, the idea of cross-validation introduced in Remark 3.2 can be employed to select the values of the smoothing parameters for constructing $\hat{r}_{MOD}(x)$.

4. Simulations

In this section, a simulation study was performed to compare the performance of six kernel regression estimators dealing with discontinuity. These six estimators include the local linear smoother $\hat{m}_{LLS}(x)$ using locally varying bandwidths (Fan and Gijbels 1995), the diagnostic kernel regression estimators $\hat{m}_{DGN}(x)$ (Hall and Titterington 1992) and $\hat{m}_{QIU}(x)$ (Qiu 2003), the boundary modified LCM $\hat{m}_{LCM}(x)$ (Hwang 2002), the boundary modified robust kernel estimator $\hat{m}_{RBT}(x)$ (Härdle and Gasser 1984), and our proposed estimator $\hat{m}_{MOD}(x)$. They were computed without knowing in advance that the regression function under study is discontinuous. The formulations of $\hat{m}_{LCM}(x)$, $\hat{m}_{RBT}(x)$, and $\hat{m}_{MOD}(x)$ have been given in Section 2. The kernel function K used by each discussed estimator was the Epanechnikov kernel. Two regression functions $m_1(x) = x^3 + (1/2)\ I_{[0.5,1]}(x)$ with one point $t = 0.5$ of discontinuity, and $m_2(x) = 0.5\ I_{[0,0.3]}(x) + 0\ I_{(0.3,0.7)}(x) + 1\ I_{[0.7,1]}(x)$ with two points $t = 0.3$ and 0.7 of discontinuity were chosen. For each regression function, two sample sizes $n = 200$ and 500 were considered. For each regression function and each sample size, the regression errors ϵ_i were pseudo independent normal random variables $N(0, \sigma^2)$, where $\sigma = 0.25$, and 100 independent sets of observations were generated from the regression model (2.1).

Given each data set, the value of the integrated square error $ISE_{MOD}(b,\ h,\ g)$ and that of the cross-validation score $CV_{MOD}(b, h, g)$ for our proposed estimator $\hat{m}_{MOD}(x)$ were calculated on the equally spaced logarithmic grid of $51 \times 51 \times 51$ values of (b, h, g) in the region $[0.01, 0.1] \times [0.01, 0.3] \times [0.1, 0.5]$. See Marron and Wand (1992) for a discussion that an equally spaced grid of parameters is typically not a very efficient design for this type of grid search. For the given value (b, h, g), the value of $ISE_{MOD}(b, h, g)$ was empirically approximated by the quantity $(1/u) \sum_{i=0}^{u} \{\hat{m}_{MOD}(u_i) - m(u_i)\}^2$, where $u_i = i/u$ and $u = 200$. After evaluation on the grid, the global minimizers $(\tilde{b}, \tilde{h}, \tilde{g})$ of $ISE_{MOD}(b, h, g)$ and $(\hat{b}, \hat{h}, \hat{g})$ of $CV_{MOD}(b, h, g)$ were taken on the grid.

When the values of $(\tilde{b}, \tilde{h}, \tilde{g})$ over the 100 pseudo data sets were obtained, the sample average and standard deviation of their corresponding $ISE_{MOD}(\tilde{b}, \tilde{h}, \tilde{g})$

were calculated. The former quantity measures the best performance of $\hat{m}_{MOD}(x)$. The optimal regression function estimate produced by $\hat{m}_{MOD}(x)$ using the optimal

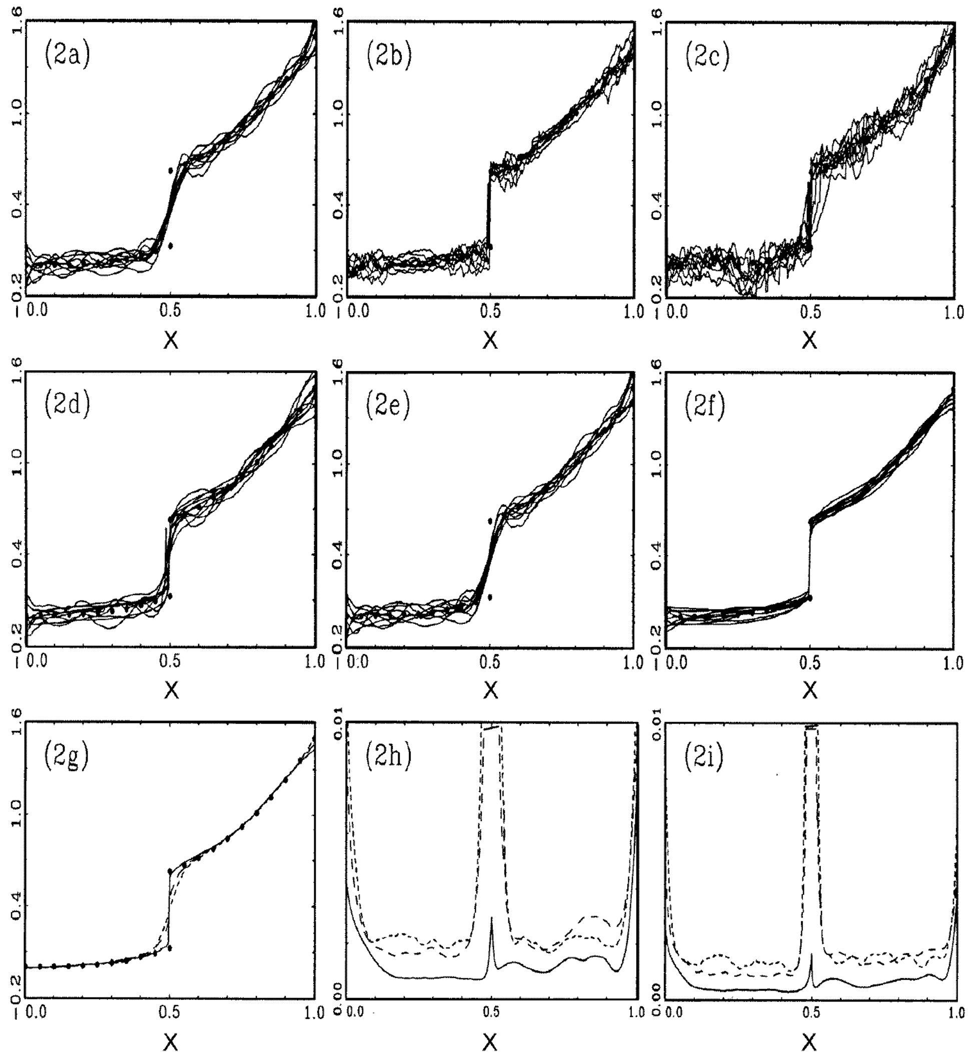

Fig. 2. With the sample size $n = 200$, (2a)-(2f) show the regression function $m_1(x)$ (dotted curve), and 10 optimal regression function estiamtes (solid curves) derived respectively by $\hat{m}_{LLS}(x)$, $\hat{m}_{DGN}(x)$, $\hat{m}_{QIU}(x)$, $\hat{m}_{LCM}(x)$, $\hat{m}_{RBT}(x)$, and $\hat{m}_{MOD}(x)$, and (2g)-(2h) plot $m_1(x)$ (dotted curve), and sample averages and sample mean square errors, respectively, of the 100 optimal regression function estimates derived by $\hat{m}_{LCM}(x)$ (dashed curve), $\hat{m}_{RBT}(x)$ (short-dashed curve), and $\hat{m}_{MOD}(x)$ (solid curve). (2i) gives sample mean square errors of the three estimators in (2h) with $n = 200$ replaced by $n = 500$.

value $(\tilde{b}, \tilde{h}, \tilde{g})$ of (b, h, g) was computed for each data set. On the other hand, the sample average and standard deviation of the corresponding $ISE_{MOD}(\hat{b}, \hat{h}, \hat{g})$ over the 100 pseudo data sets were also computed. The former measures the performance of $\hat{m}_{MOD}(x)$ which can be obtained in practice by using the cross-validated bandwidth. The obtainable regression function estimate produced by $\hat{m}_{MOD}(x)$ employing $(\hat{b}, \hat{h}, \hat{g})$ was computed for each data set.

The same procedures for computing both the best and the obtainable performance

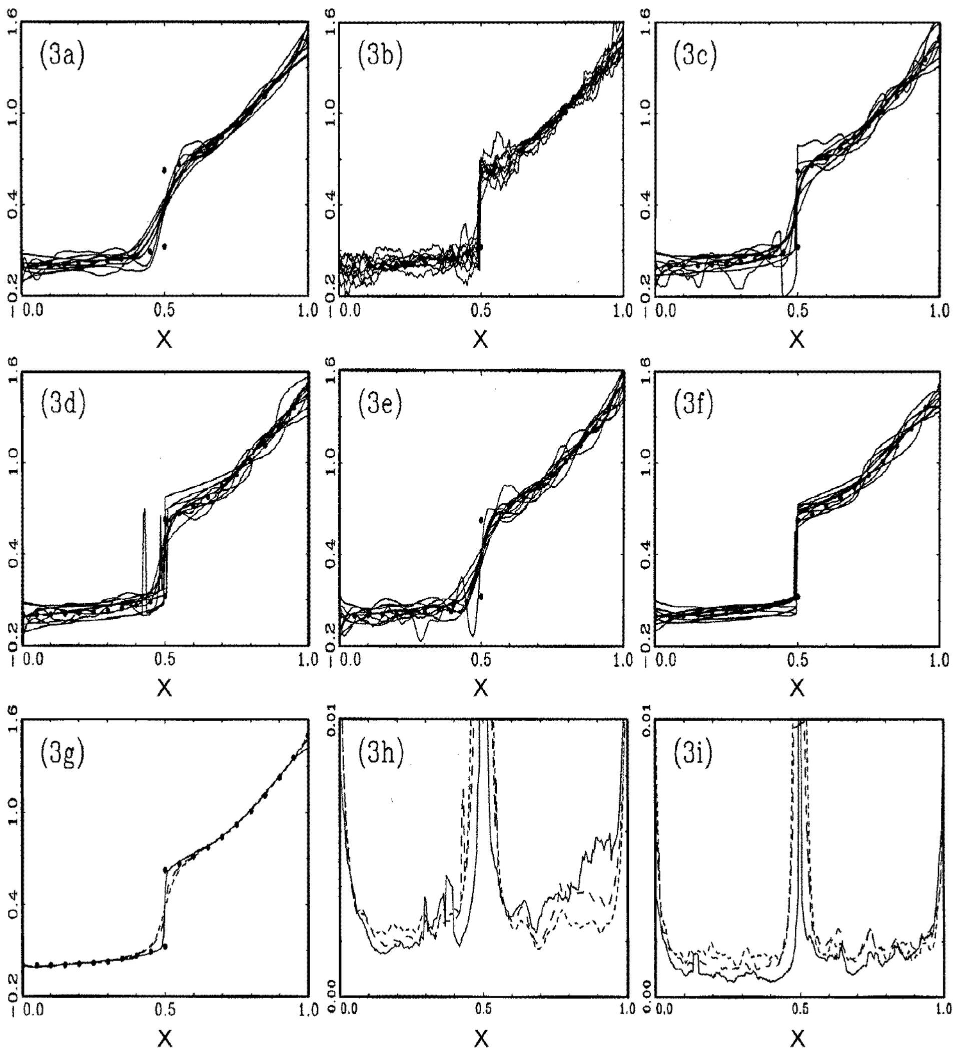

Fig. 3. The description of Figure 3 is the same as that of Figure 2 with optimal results replaced by obtainable results.

of $\hat{m}_{MOD}(x)$ were applied to the other five discussed estimators $\hat{m}_{LLS}(x)$, $\hat{m}_{DGN}(x)$, $\hat{m}_{QIU}(x)$, $\hat{m}_{LCM}(x)$, and $\hat{m}_{RBT}(x)$. For doing this, the same grid of values of h and g used by $\hat{m}_{MOD}(x)$ was employed. The best performance of $\hat{m}_{LLS}(x)$ was derived by using a globally constant bandwidth, but its obtainable performance by employing the locally varying bandwidth (Fan and Gijbels 1995). Let $ISE_{LLS}(\tilde{h})$ and $ISE_{LLS}(\hat{h})$ be similarly defined for $\hat{m}_{LLS}(x)$, $ISE_{DNG}(\tilde{h})$ and $ISE_{DNG}(\hat{h})$ for $\hat{m}_{DGN}(x)$, $ISE_{QIU}(\tilde{h})$ and $ISE_{QIU}(\hat{h})$ for $\hat{m}_{QIU}(x)$, $ISE_{LCM}(\tilde{h}, \tilde{g})$ and $ISE_{LCM}(\hat{h}, \hat{g})$ for $\hat{m}_{LCM}(x)$, and$ISE_{RBT}(\tilde{h})$ and$ISE_{RBT}(\hat{h})$ for $\hat{m}_{RBT}(x)$.

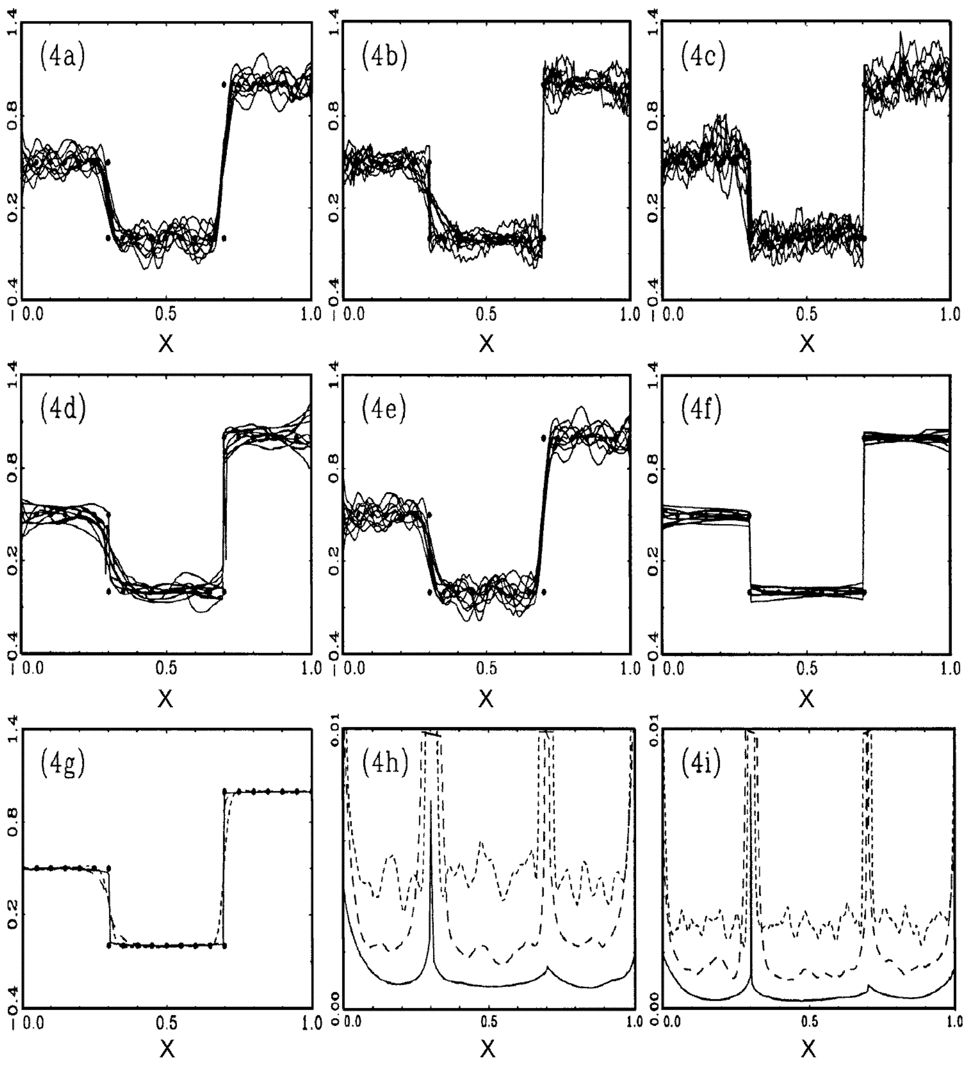

Fig. 4. The description of Figure 4 is the same as that of Figure 2 with the regression function $m_1(x)$ replaced by $m_2(x)$.

The simulation results are summarized in the following figures and tables.

Given the regression function $m_1(x)$, the best and the practical performance of the six discussed estimators are presented respectively in Figures 2 and 3. Given the sample size $n = 200$, Figures 2a-2f present 10 optimal regression function estimates derived from 10 sets of simulated data by the six discussed estimators. It is clear that the optimal regression function estimates produced by our $\hat{m}_{MOD}(x)$ show the location of jump point most accurately, and have the smoothest appearance in the smooth region. Also, Figures 2g-2h show respectively that, for x in the jump region,

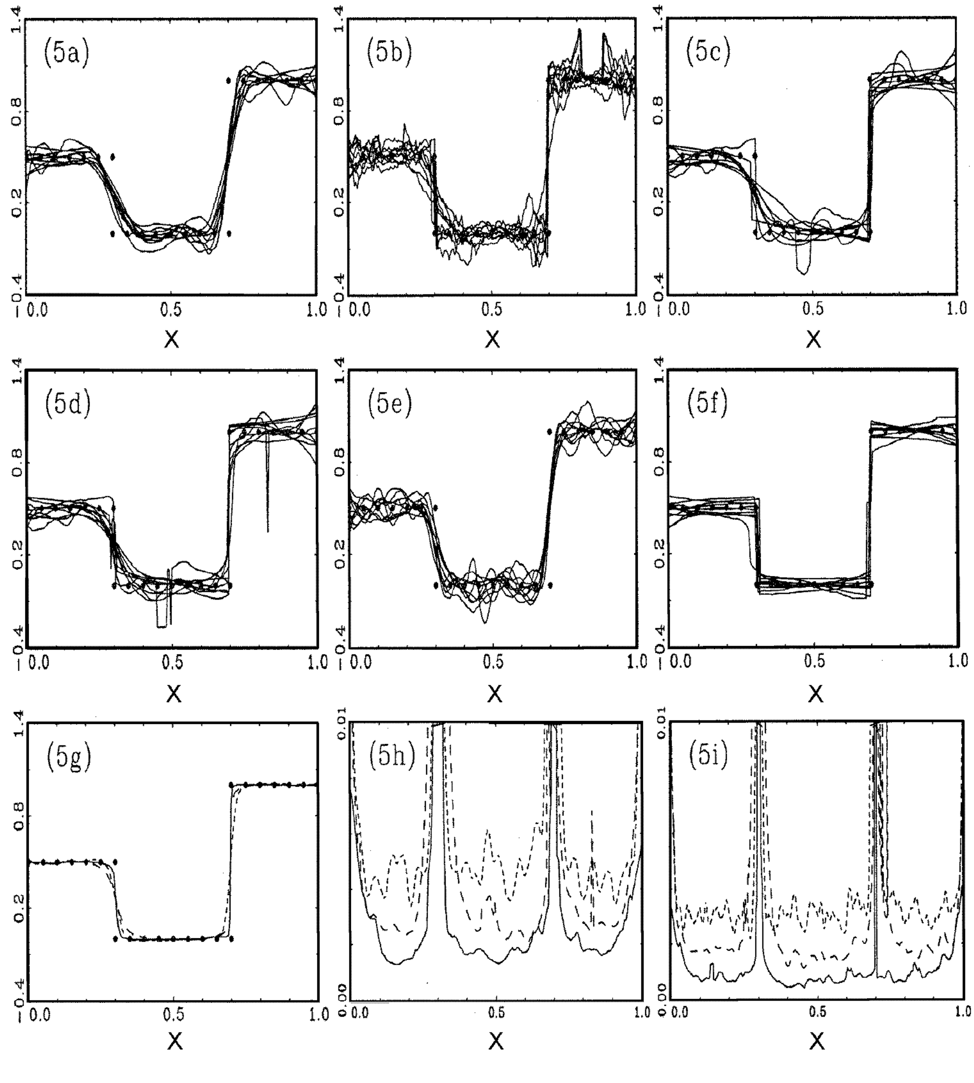

Fig. 5. The description of Figure 5 is the same as that of Figure 3 with the regression function $m_1(x)$ replaced by $m_2(x)$.

the optimal regression function estimate produced by $\hat{m}_{MOD}(x)$ does not suffer from inconsistency, and has smaller sample mean square error than $\hat{m}_{LCM}(x)$ and $\hat{m}_{RBT}(x)$. Given the sample size $n = 500$, Figure 2i present the same advantage of $\hat{m}_{MOD}(x)$ over $\hat{m}_{LCM}(x)$ and $\hat{m}_{RBT}(x)$ as they do in Figure 2h. Under the simulation setting, the results in Figure 2 show that our modified LCM improves the inconsistency of the ordinary LCM in the jump region. These remarks made from Figure 2 for the optimal regression function estimate can be applied to the obtainable regression function estimate in Figure 3. Similar conclusions drawn from Figures 2-3 for the regression function $m_1(x)$ can be made for the regression function $m_2(x)$ in Figures 4-5. Finally, the regression function estimates produced by the six discussed estimators with the sample size $n = 500$ are not reported here since their performance is the same as that shown in Figures 2-5 for the sample size $n = 200$.

Table 1 shows that, for each regression function and each sample size, our suggested estimator $\hat{m}_{MOD}(x)$ has better both the optimal and the practical performance than the other five discussed estimators, in the sense of yielding smaller sample mean integrated square error.

Table 1. Values of the sample mean and standard deviation (contained in the parentheses) of $ISE_{LLS}(\tilde{h})$ and $ISE_{LLS}(\hat{h})$ for $\hat{m}_{LLS}(x)$, $ISE_{DNG}(\tilde{h})$ and $ISE_{DNG}(\hat{h})$ for $\hat{m}_{DNG}(x)$, $ISE_{QIU}(\tilde{h})$ and $ISE_{QIU}(\hat{h})$ for $\hat{m}_{QIU}(x)$, $ISE_{LCM}(\tilde{h},\tilde{g})$ and $ISE_{LCM}(\hat{h},\hat{g})$ for $\hat{m}_{LCM}(x)$, $ISE_{RBT}(\tilde{h})$ and $ISE_{RBT}(\hat{h})$ for $\hat{m}_{RBT}(x)$, and $ISE_{MOD}(\tilde{b},\tilde{h},\tilde{g})$ and $ISE_{MOD}(\hat{b},\hat{h},\hat{g})$ for $\hat{m}_{MOD}(x)$. Each of these values has been multiplied by 10^3.

	best performance		practical performance	
	$n = 200$	$n = 500$	$n = 200$	$n = 500$
	regression function $m_1(x)$			
$\hat{m}_{LLS}(x)$	4.82(1.13)	3.06(0.58)	4.94(1.29)	3.20(0.84)
$\hat{m}_{DNG}(x)$	3.34(1.53)	1.28(0.58)	4.79(2.08)	1.85(0.84)
$\hat{m}_{QIU}(x)$	7.15(2.74)	3.79(1.51)	8.25(3.44)	4.15(1.67)
$\hat{m}_{LCM}(x)$	3.83(1.60)	2.22(0.96)	5.04(2.06)	2.79(1.22)
$\hat{m}_{RBT}(x)$	4.72(1.13)	2.99(0.58)	5.32(1.33)	3.31(0.83)
$\hat{m}_{MOD}(x)$	1.35(0.84)	0.65(0.33)	4.51(2.94)	1.72(1.21)
	regression function $m_2(x)$			
$\hat{m}_{LLS}(x)$	10.6(1.58)	6.63(0.84)	13.1(2.50)	8.66(2.56)
$\hat{m}_{DNG}(x)$	5.89(2.37)	2.42(1.08)	8.67(4.38)	3.26(1.45)
$\hat{m}_{QIU}(x)$	9.78(3.37)	5.02(1.46)	10.8(4.24)	5.52(1.70)
$\hat{m}_{LCM}(x)$	4.87(1.73)	2.91(0.91)	6.33(2.28)	3.86(1.38)
$\hat{m}_{RBT}(x)$	9.81(1.60)	6.11(0.85)	10.3(1.70)	6.47(1.01)
$\hat{m}_{MOD}(x)$	1.19(0.86)	0.56(0.34)	4.75(3.06)	2.14(1.58)

5. Sketches of the Proofs

Proof of Theorem 3.1. The following notation and asymptotic results will be used in this section. For each $x \in [0,1]$, set

$$A(x) = \sum_{j=1}^{q} \alpha_j(x) L^{(1)}[\{\bar{Y}_j - m(x)\}/g],$$

$$B(x) = \sum_{j=1}^{q} \alpha_j(x) L^{(2)}[\{\bar{Y}_j - m(x)\}/g],$$

$$U(x) = \hat{m}_{MOD}(x) - m(x).$$

Using the regression model (2.1) and (A1)-(A5), through a straightforward calculation, for each $x \in R_1$, R_2, R_3, R_4, R_5, the dominant terms of the asymptotic expectation and variance of $A(x)$ and those of $B(x)$ can be expressed respectively by

$$\mathrm{E}\{A(x)\} = (1/2)h\, g^{-1} L^{(2)}(0)\{h\mu^{(2)}(x)b_{\ell,2}, \quad h\mu^{(2)}(x)b_{r,2},$$

$$(-2)\mu^{(1)}(x)\kappa_{r,1}, \quad (-2)\mu^{(1)}(x)\kappa_{\ell,1}, \quad h\mu^{(2)}(x)\kappa_2\}, \tag{5.1}$$

$$\mathrm{Var}\{A(x)\} = \sigma^2 n^{-1} h^{-1} g^{-2} L^{(2)}(0)^2 \{v_{\ell,2}, \quad v_{r,2}, \quad \tau_{r,0}, \quad \tau_{\ell,0}, \quad \tau_0\}, \tag{5.2}$$

$$\mathrm{E}\{B(x)\} = L^{(2)}(0)\ \{1, \quad 1, \quad \kappa_{r,0}, \quad \kappa_{\ell,0}, \quad 1\}, \tag{5.3}$$

$$\mathrm{Var}\{B(x)\} = \sigma^4 n^{-2} h^{-1} b^{-1} g^{-2} L^{(4)}(0)^2 \{v_{\ell,2}, \quad v_{r,2}, \quad \tau_{r,0}, \quad \tau_{\ell,0}, \quad \tau_0\}. \tag{5.4}$$

Here the notation κ_j, $\kappa_{\ell,j}$, $\kappa_{r,j}$, τ_j, $\tau_{\ell,j}$, $\tau_{r,j}$, $b_{\ell,2}$, $v_{\ell,2}$, $b_{r,2}$, and $v_{r,2}$ has been defined in Section 3.

We now prove (3.1) and (3.2). Taking the first derivative of $S_{MOD}(\xi;\, x)$ in (2.3) with respect to ξ, and applying the first order Taylor expansion to the result, we have

$$0 = (\partial/\partial\xi) S_{MOD}(\xi; x)|_{\xi=\hat{m}_{MOD}(x)} = A(x) - U(x) g^{-1} B^*(x), \tag{5.5}$$

where $B^*(x)$ is $B(x)$ with $\bar{Y}_j - m(x)$ in $B(x)$ replaced by $\bar{Y}_j - m(x) + Q_j(x)$ and $Q_j(x)$ satisfy $|Q_j(x)| \le |U(x)|$ for all j. Using (5.5), the Lipschitz continuity of $L^{(2)}$ and the fact that $\sum_{j=1}^{q} |\alpha_j(x)| = O(1)$, through a straightforward calculation, we have

$$0 = A(x) - U(x) g^{-1} B(x) + O\{g^{-2} U(x)^2\}. \tag{5.6}$$

Using (5.1)-(5.4) and comparing the magnitudes of $A(x)$ and $B(x)$ in both sides of (5.6), a straightforward calculation leads to

$$U(x) = o_p(1). \tag{5.7}$$

Hence $Q_j(x) = o_p(1)$, for all j.

By (5.5), we have

$$U(x) = g\ A(x)/\ B^*(x).$$

Using the result, (5.1)-(5.4), (5.7), (A1)-(A5), and approximation to the standard errors of functions of random variables given in Section 10.5 of Stuart and Ord (1987), through a straightforward calculation, (3.1) and (3.2) follow. Hence the proof of Theorem 3.1 is complete.

Acknowledgment

The research was supported by National Science Council, Republic of China. The authors thank Editor and referees for their valuable suggestions which greatly improve the presentation of this paper.

References

1. P. J. Besl, J. B. Birch and L. T. Watson, Robust window operators. *Machine Vision and Applications*, **2** (1989), 179-191.
2. C. K. Chu, I. Glad, F. Godtliebsen and J. S. Marron, Edge-preserving smoothers for image processing (with discussion). *J. Amer. Statistical Assoc.*, **93** (1998), 526-556.
3. R. L. Eubank, *Spline Smoothing and Nonparametric Regression*, Marcel Dekker Inc., New York, 1988.
4. J. Fan, Design-adaptive nonparametric regression, *J. Amer. Statistical Assoc.*, **87** (1992), 998-1004.
5. J. Fan, Local linear regression smoothers and their minimax efficiencies. *Annals of Statistics*, **21** (1993), 196-216.
6. J. Fan, T. Gasser, I. Gijbels, M. Brookmann and J. Engels, *Local polynomial fitting: A standard for nonparametric regression. Discussion paper 9315*, Institut de Statistique, Universite Catholique de Louvain, Belgium, 1993.
7. J. Fan and I. Gijbels, Data-driven bandwidth selection in local polynomial fitting: Variable bandwidth and spatial adaption. *J. Royal Statistical Soc. Ser. B*, **57** (1995), 371-394.
8. J. Fan and I. Gijbels, Local Polynomial Modeling and Its Application - Theory and Methodologies, Chapman and Hall, New York, 1996.

9. J. Fan, T. Hu and Y. Truong, Robust non-parametric function estimation. *Scandinavian J. Statistics*, **21** (1994), 433-446.

10. J. Fan and J. Jiang, Variable bandwidth and one-step local M-estimator, *Sci. Chin. Ser. A*, **29** (1999), 1-15.

11. P. Hall and D. M. Titterington, Edge-preserving and peak-preserving smoothing, *Technometrics*, **34** (1992), 429-440.

12. F. Hample, E. Ronchetti, P. Rouseeuw and W. Stahel, *Robust Statistics: The Approach Based on Influence Functions*, Wiley, New York, 1986.

13. W. Härdle, *Applied Nonparametric Regression*, Cambridge University Press, 1990.

14. W. Härdle, *Smoothing Techniques: With Implementation in S.*, Springer Series in Statistics, Springer-Verlag, Berlin, 1991.

15. W. Härdle and T. Gasser, Robust nonparametric function fitting. *J. Royal Statistical Soc. Ser. B*, **46** (1984), 42-51.

16. W. Härdle and J. S. Marron, Optimal bandwidth selection in nonparametric regression function estimation, *Ann. Statistics*, **13** (1985), 1465-1481.

17. P. J. Huber, *Robust Statistics*, Wiley, New York, 1981.

18. R. C. Hwang, A new version of the local constant M-smoother. *Comm. Statistics-Theory Methods*, **31** (2002), 833-848.

19. J. S. Marron, Automatic smoothing parameter selection: A survey. *Empirical Economics*, **13** (1988), 187-208.

20. J. S. Marron and M. P. Wand, Exact mean integrated squared error. *Ann. Statistics*, **20** (1992), 712-736.

21. J. A. McDonald and A. B. Owen, Smoothing with split linear fits. *Technometrics*, **28** (1986), 195-208.

22. H. G. Müller, *Nonparametric Regression Analysis of Longitudinal Data*, Lecture Notes in Statistics, No. 46, Springer-Verlag, Berlin 1988.

23. H. G. Müller, Change-points in nonparametric regression analysis. *Ann. Statistics*, **20** (1992), 737-761.

24. P. Qiu, Discontinuous regression surfaces fitting, *Ann. Statistics*, **26** (1998), 2218-2245.

25. P. Qiu, A jump-preserving curve fitting procedure based on local piecewise-linear kernel estimation, *J. Nonparametric Statistics*, **15** (2003), 437-453.

26. J. Rice, Bandwidth choice for nonparametric regression, *Ann. Statistics*, **12** (1984), 1215-1230.

27. H. Rue, C. K. Chu, F. Godtliebsen and J. S. Marron, M-smoother with local linear fit. *J. Nonparametric Statistics*, **14** (2002), 155-168.

28. D. Ruppert and M. P. Wand, Multivariate locally weighted least squares regression, *Ann. Statistics*, **22** (1994), 1346-1370.

29. D. W. Scott, *Multivariate Density Estimation: Theory, Practice, and Visualization*, Wiley, New York 1992.

30. J. H. Shiau, *Smoothing spline estimation of functions with discontinuities*, Unpublished PhD. Dissertation, Department of Statistics, University of Wisconsin. Madison, Wisconsin, 1985.

31. B. W. Silverman, *Density Estimation for Statistics and Data Analysis*, Chapman and Hall, New York, 1986.

32. J. S. Simonoff, *Smoothing Methods in Statistics*, Springer, New York, 1996.

33. A. Stuart and J. K. Ord, *Kendall's Advanced Theory of Statistics, Volume 1*, Oxford University Press, New York, 1987.

34. A. B. Tsybakov, Robust reconstruction of functions by the local approximation, *Problems of Information Trans.*, **22** (1986), 133-146.

35. M. P. Wand and M. C. Jones, *Kernel Smoothing*, Chapman and Hall, London, 1995.

36. J. S. Wu and C. K. Chu, Kernel type estimators of jump points and values of a regression function. *Ann. Statistics*, **23** (1993a), 1545-1566.

37. J. S. Wu and C. K. Chu, Nonparametric function estimation and bandwidth selection for discontinuous regression functions. *Statistica Sinica*, **3** (1993b), 557-576.

Jung-Huei Lin
General Education Center,
Taiwan Hospitality and Tourism College,
Hualien 974, Taiwan

Tzu-Kuei Chang
Department of Mathematics,
National Hualien University of Education,
Hualien 970, Taiwan

Chih-Kang Chu
Department of Applied Mathematics,
National Dong Hua University,
Hualien 974, Taiwan
E-mail: chu@mail.ndhu.edu.tw