The Korean Communications
in Statistics Vol. 4, No. 3, 1997
pp. 881 - 890

Shifted Nadaraya Watson Estimator

Sung S. Chung[1)]

Abstract

The local linear estimator usually has more attractive properties than Nadaraya-Watson estimator. But the local linear estimator gives bad performance where data are sparse. Müller and Song proposed Shifted Nadaraya Watson estimator which has treated data sparsity well. We show that Shifted Nadaraya Watson estimator has good performance not only in the sparse region but also in the dense region, through the simulation study. And we suggest the boundary treatment of Shifted Nadaraya Watson estimator.

1. Introduction

Nonparametric curve estimation is a useful and powerful tool for finding the structure in data. In particular, there is its strength in case that the structure in data is difficult to be ascertained by the parametric method. See, for example, the books of Silverman(1986), Eubank (1988), Müller (1988), Härdle (1990), Scott (1992) and Wand and Jones (1995).

Suppose that we have a set of bivariate data (X_i, Y_i), $i = 1, \cdots, n$. The regression relationship is commonly modelled as

$$Y_i = m(X_i) + \varepsilon_i, \qquad i=1,\cdots,n \tag{1}$$

where $\varepsilon_1, \cdots, \varepsilon_n$ are independent random variables with the expectation 0 and the variance σ^2. To estimate the regression function, $m(x) = E(Y|X=x)$, the kernel-based regression estimators are often used because of their simplicity and interpretability.

An intuitively attractive kernel-based regression estimator is Nadaraya-Watson estimator ($\widehat{m}_{NW}(x;h)$) proposed by Nadaraya (1964) and Watson (1964). An improved estimator is local linear estimator ($\widehat{m}_{LL}(x;h)$), proposed by Stone (1977), Cleveland (1979), Müller (1987) and Fan (1992). These estimators are based on moving locally weighted averaging. Fan (1992, 1993) showed that local linear estimator has the desirable asymptotic MSE properties as well

1) Assistant Professor, Department of Statistics, Chonbuk National University, Chonju, Chonbuk 560-756, Korea

as minimax optimality properties, in the interior of the design space. And Fan and Gijbels (1992) showed that it automatically adapts itself at boundaries to achieve good performance there too.

But in case when the data are sparse, local linear estimator often gives bad performance. This is due to its infinite variance in random design contexts. See Seifert and Gasser (1995), Hall and Marron (1996) and Hall, Marron, Neumann and Titterington (1996). Müller and Song (1993) suggested Shifted Nadaraya Watson estimator ($\hat{m}_{SNW}(x;h)$) which gives much better performance when the data are sparse.

Figure 1(a) shows an example of $\hat{m}_{NW}(x;h)$ and $\hat{m}_{LL}(x;h)$. $\hat{m}_{LL}(x;h)$ gives better behavior than $\hat{m}_{NW}(x;h)$ except the center region. But $\hat{m}_{LL}(x;h)$ shows terrible feature near the center. Figure 1(b) shows $\hat{m}_{SNW}(x;h)$ and $\hat{m}_{LL}(x;h)$ for the same data with Figure 1(a). $\hat{m}_{SNW}(x;h)$ is almost same to $\hat{m}_{LL}(x;h)$ where the data are dense. But in the center area where the data are sparse, it has much better performance than $\hat{m}_{LL}(x;h)$. The bandwidth used here, is the direct plug-in bandwidth, $\hat{h}_{DPI}$, of Ruppert, Sheather and Wand (1995).

In this paper, we show that $\hat{m}_{SNW}(x;h)$ has good performance when the design densities are sparse as well as when they are dense through the simulation study. And we suggest the boundary treatment of $\hat{m}_{SNW}(x;h)$ which wasn't suggested the exact method by Mammen and Marron (1996). In section 2, we describe the kernel-type regression estimators including $\hat{m}_{NW}(x;h)$, $\hat{m}_{LL}(x;h)$ and $\hat{m}_{SNW}(x;h)$. In section 3, we compare three estimators in the several setting through Monte Carlo studies.

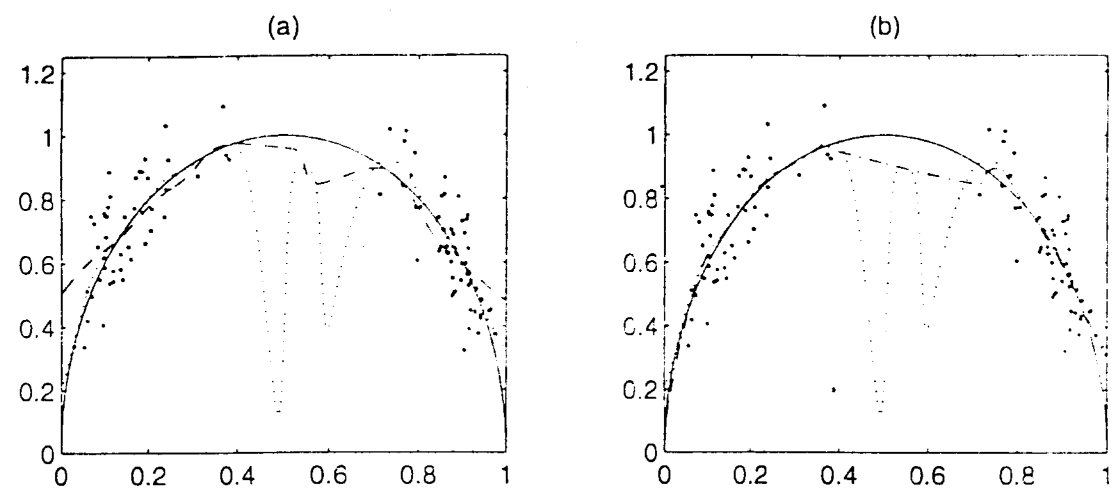

Figure 1. *Kernel regression estimates with the simulated data. The solid line is true regression function,* $m(x) = 4\sqrt{x}\sqrt{1-x}$ *with n=100,* $\sigma^2=0.01$ *and* $h=0.0432$ *for the normal kernel. (a) The dashed line is* $\hat{m}_{NW}(x;h)$*, the dotted line is* $\hat{m}_{LL}(x;h)$*; (b) the dashed line is* $\hat{m}_{SNW}(x;h)$*, the dotted line is* $\hat{m}_{LL}(x;h)$*.*

2. Shifted Nadaraya Watson Estimator

Given a set of bivariate data $(X_1, Y_1), \cdots, (X_n, Y_n)$, we estimate the regression function in (1), based on the data. For simplicity, we will assume X_i takes the value in [0, 1]. Nadaraya-Watson regression estimator is given by

$$\widehat{m}_{NW}(x;h) = \frac{n^{-1}\sum_{i=1}^{n} K_h(X_i - x)Y_i}{n^{-1}\sum_{i=1}^{n} K_h(X_i - x)}. \tag{2}$$

Note that the denominator in (2) is the kernel density estimator of $f(x)$, the marginal density of X. Here, $K_h(u) = K(u/h)/h$ and K is called the kernel function. The scale parameter h is called the bandwidth or smoothing parameter and is crucial to the performance of $\widehat{m}(x;h)$. The kernel function, K, is a continuous and symmetric probability function. It is well known that the choice of the kernel function, K, is of essentially negligible concern compared to the choice of the bandwidth (Silverman, 1986).

Local linear regression estimator is given by the value of $\widehat{b_0}$ when b_0 and b_1 are chosen to minimize the local least squares function

$$\sum_{i=1}^{n}\{Y_i - b_0 - b_1(X_i - x)\}^2 K_h(X_i - x).$$

And this can be expressed as follows

$$\widehat{m}_{LL}(x;h) = n^{-1}\sum_{i=1}^{n}\frac{\{\hat{s}_2(x;h) - \hat{s}_1(x;h)(X_i - x)\}\ K_h(X_i - x)Y_i}{\hat{s}_2(x;h)\ \hat{s}_0(x;h) - \hat{s}_1(x;h)^2}$$

where $\hat{s}_r(x;h) = n^{-1}\sum_{i=1}^{n}(X_i - x)^r K_h(X_i - x)$, $r = 0,1,2$.

Let $\hat{\xi}(x)$ denote the kernel based local center of mass, i.e.

$$\hat{\xi}(x) = \frac{\sum_{i=1}^{n} K_h(X_i - x)X_i}{\sum_{i=1}^{n} K_h(X_i - x)}. \tag{3}$$

Then at $\hat{\xi}(x)$, Shifted Nadaraya Watson estimate is defined as follows (Müller and Song, 1993)

$$\widehat{m}_{SNW}(\hat{\xi}(x);h) = \widehat{m}_{NW}(x;h) \tag{4}$$

Hence $\widehat{m}_{SNW}(x;h)$ is a correction, i.e. a horizontal shift, of $\widehat{m}_{NW}(x;h)$ taking into account the difference between the center, x, of the window and the local center of mass, $\hat{\xi}(x)$, of the design points, X_i's. For the normal kernel, $\hat{\xi}(x)$ is strictly increasing and so $\hat{\xi}^{-1}(x)$ exists. Then $\hat{m}_{SNW}(x;h)$ is uniquely defined for $x \in [\hat{\xi}(0), \hat{\xi}(1)]$ and (4) can be replaced by $\hat{m}_{SNW}(x;h) = \hat{m}_{NW}(\hat{\xi}^{-1}(x);h)$. Müller and Song (1993) showed that $\hat{m}_{SNW}(x;h)$ and $\hat{m}_{LL}(x;h)$ have the same asymptotic behavior for x in the interior of [0, 1]. But $\hat{m}_{NW}(x;h)$ has a more complicated bias expression than $\hat{m}_{LL}(x;h)$ and $O(h)$ boundary bias like Figure 1(a) (Fan, 1992, Wand and Jones, 1995).

And Mammen and Marron (1996) showed that $\hat{m}_{SNW}(x;h)$ has the good performance in the boundary regions. It is quantified by

$$\hat{m}_{SNW}(x;h) = m(x) + O_p(n^{-2/5})$$

for all $x \in [0, \hat{\xi}(0)]$. Note that this is same for $\hat{m}_{LL}(x;h)$. They suggested using the smooth linear extrapolation in the boundaries. But used for several data sets, the linear extrapolation gave very large variance. So we propose using the weighted average of the least squares estimates and $\hat{m}_{SNW}(x;h)$ near the left boundary region, i.e.

$$\hat{m}_{SNW}(x;h) = \begin{cases} l(x) & 0 < x < \hat{\xi}(0) \\ w(x)\,\hat{m}_{NW}(\hat{\xi}^{-1}(x);h) + (1-w(x))\,l(x) & \hat{\xi}(0) \le x < 2\hat{\xi}(0) \\ \hat{m}_{NW}(\hat{\xi}^{-1}(x);h) & x \ge 2\hat{\xi}(0) \end{cases} \qquad (5)$$

where $w(x) = \dfrac{2\hat{\xi}(0) - x}{2\hat{\xi}(0) - \hat{\xi}(0)}$ and $l(x)$ is the least squares fit of a line with the points in $[\hat{\xi}(0), 2\hat{\xi}(0)]$. Similarly, this method is applicable to the right boundary region.

3. Simulation Results and Discussions

Three estimators have same asymptotic properties in the interior region. In this section, we compare three estimators in several design settings through the simulation study. In our simulation study, we use MATLAB program in UNIX. We consider four regression functions as follows

(m1) $m(x) = \cos(2\pi x)$.

(m2) $$m(x) = \begin{cases} 1 & \text{if } x \in (0, \frac{1}{4}) \\ \cos(4\pi(x - 1/4)) & \text{if } x \in (\frac{1}{4}, \frac{3}{4}) \\ 1 & \text{if } x \in (\frac{3}{4}, 1) \end{cases}.$$

(m3) $m(x) = 1 - 4(x - \frac{1}{2})^2$.

(m4) $m(x) = 4\sqrt{x}\sqrt{1-x}$.

Also we consider the random designs

(d1) $X \sim U(0,1)$.

(d2) $X = \sqrt{U}$ where $U \sim U(0,1)$.

(d3) $X = \begin{cases} U_{(4)} & \text{with prob. } \frac{1}{2}, \\ U_{(17)} & \text{with prob. } \frac{1}{2}, \end{cases}$ where $U \sim U(0,1)$ and $U_{(i)}$ denotes the order statistic from $U_1, \cdots, U_{20}$. And we consider $\varepsilon_i \sim N(0, \sigma^2)$ independent of X with $\sigma = 0.2$. To estimate the bias and the variance of the regression estimator, 500 replications are performed. We use the normal function as the kernel function and the direct plug-in bandwidth selector, $\hat{h}_{DPI}$, of Ruppert, Sheather and Wand (1995) as the bandwidth. To reduce computational effort, we use the binning approach suggested by Fan and Marron (1994). This is useful because the data only need to be binned once. So binned computation has the advantage of requiring only $O(N)$ kernel evaluation, this allows very fast computation of $\widehat{m}(x;h)$ over the grid points. Here N denotes the number of grid points. We use N=401 recommended by Fan and Marron (1994). As the method for obtaining grid counts that has good properties, we use "linear binning" (see Hall and Wand, 1993).

Figure 2, 3 and 4 show the estimates of the expectations and mean squared errors of three estimates, and the true regression functions under design (d1), (d2) and (d3), respectively. Under these designs, three estimates are almost same in the dense region, but are pretty different in the boundaries or the sparse region. In Figure 2, we can see that $\widehat{m}_{NW}(x;h)$ has more bias but less variance than $\widehat{m}_{LL}(x;h)$ and $\widehat{m}_{SNW}(x;h)$. And $\widehat{m}_{SNW}(x;h)$ is almost same to $\widehat{m}_{LL}(x;h)$ at the boundary region. Note that the marginal density of X is uniform.

Figure 3 with the sparsity in the left boundary, shows that the bias of $\widehat{m}_{SNW}(x;h)$ is a little bigger than that of $\widehat{m}_{LL}(x;h)$ but the variance of $\widehat{m}_{SNW}(x;h)$ is much smaller than that of $\widehat{m}_{LL}(x;h)$ in the left area. Figure 4 with the bimodal marginal density of X, shows the same phenomenon with Figure 3. But we can see that $\widehat{m}_{LL}(x;h)$ suffers from very large variance in the boundaries and the center region. This was discussed in Fan(1992) and Mammen and Marron (1996).

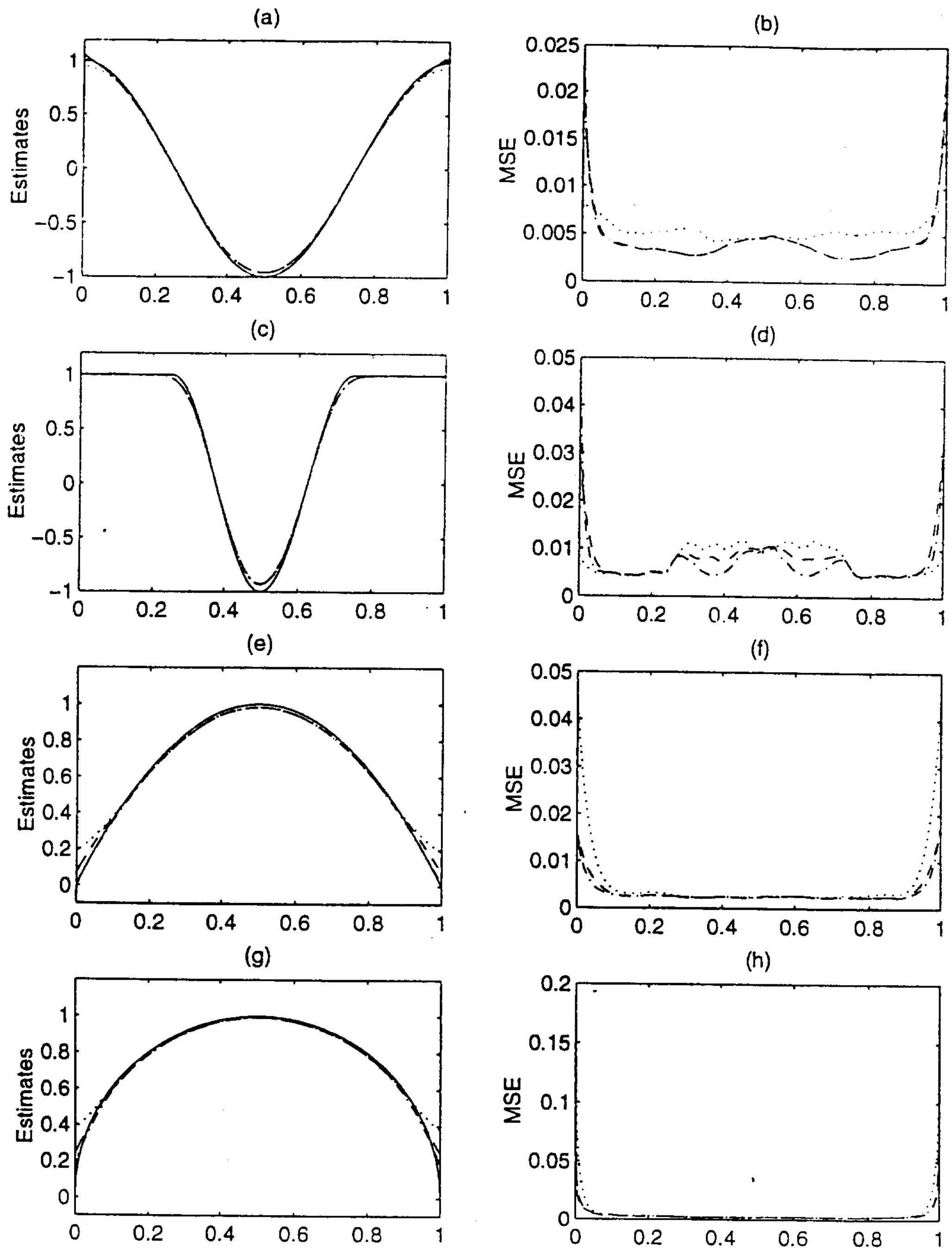

Figure 2. *Estimates of* $E(\widehat{m}(x;h))$ *and* $MSE(\widehat{m}(x;h))$ *under design (d1), and true regression functions. {(a),(b)}, {(c),(d)}, {(e),(f)} and {(g),(h)} are the estimates of {* $E(\widehat{m}(x;h))$, $MSE(\widehat{m}(x;h))$*} from the regression functions (m1)-(m4), respectively.* $m(x)$*(solid line),* $\widehat{m}_{SNW}(x;h)$*(dashed line),* $\widehat{m}_{NW}(x;h)$*(dotted line) and* $\widehat{m}_{LL}(x;h)$*(dot-dashed line).*

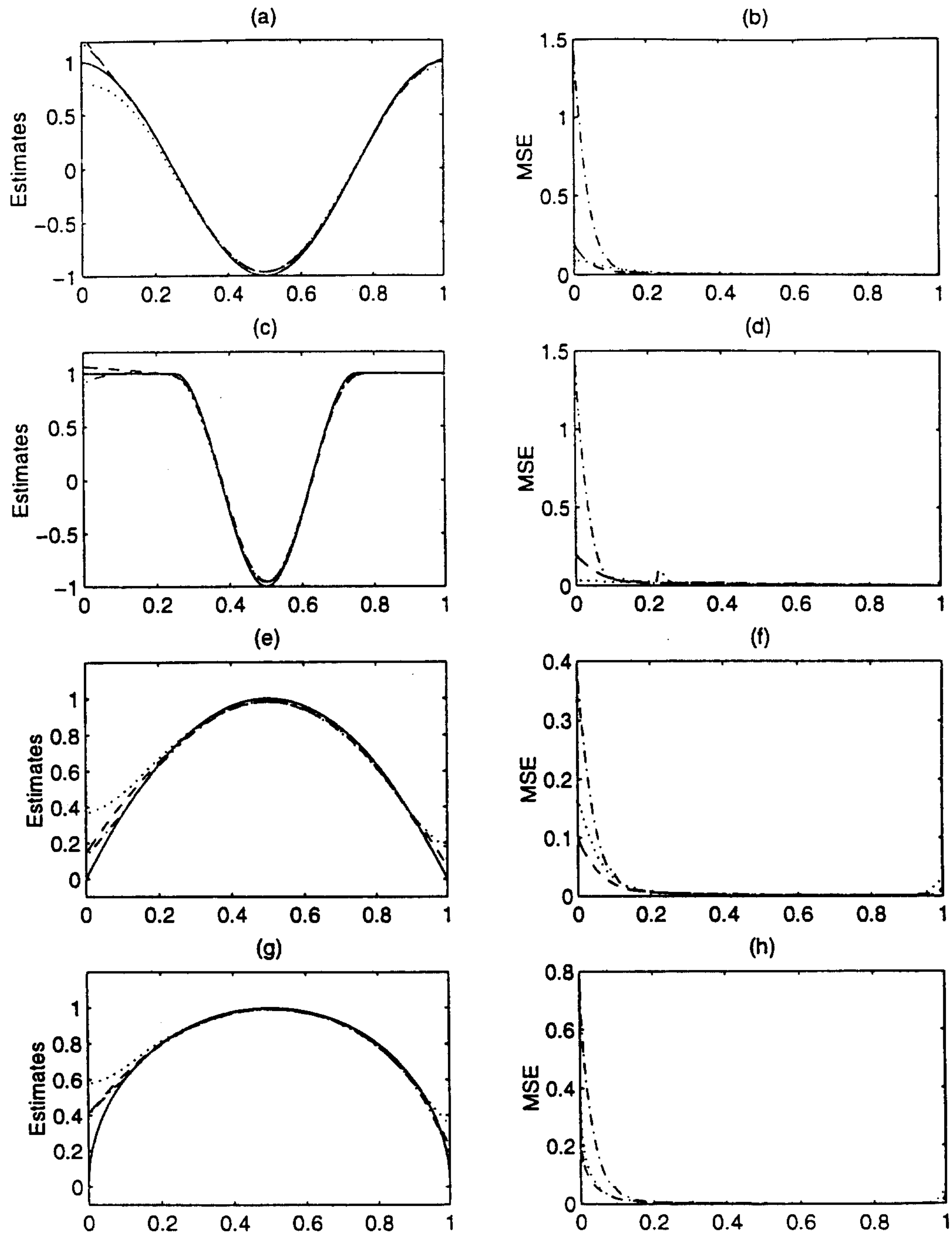

Figure 3. *Estimates of* $E(\widehat{m}(x;h))$ *and* $MSE(\widehat{m}(x;h))$ *under design (d2), and true regression functions.* *{(a),(b)}, {(c),(d)}, {(e),(f)} and {(g),(h)} are the estimates of* $\{E(\widehat{m}(x;h)), MSE(\widehat{m}(x;h))\}$ *from the regression functions (m1)-(m4), respectively.* $m(x)$*(solid line),* $\widehat{m}_{SNW}(x;h)$*(dashed line),* $\widehat{m}_{NW}(x;h)$*(dotted line) and* $\widehat{m}_{LL}(x;h)$*(dot-dashed line).*

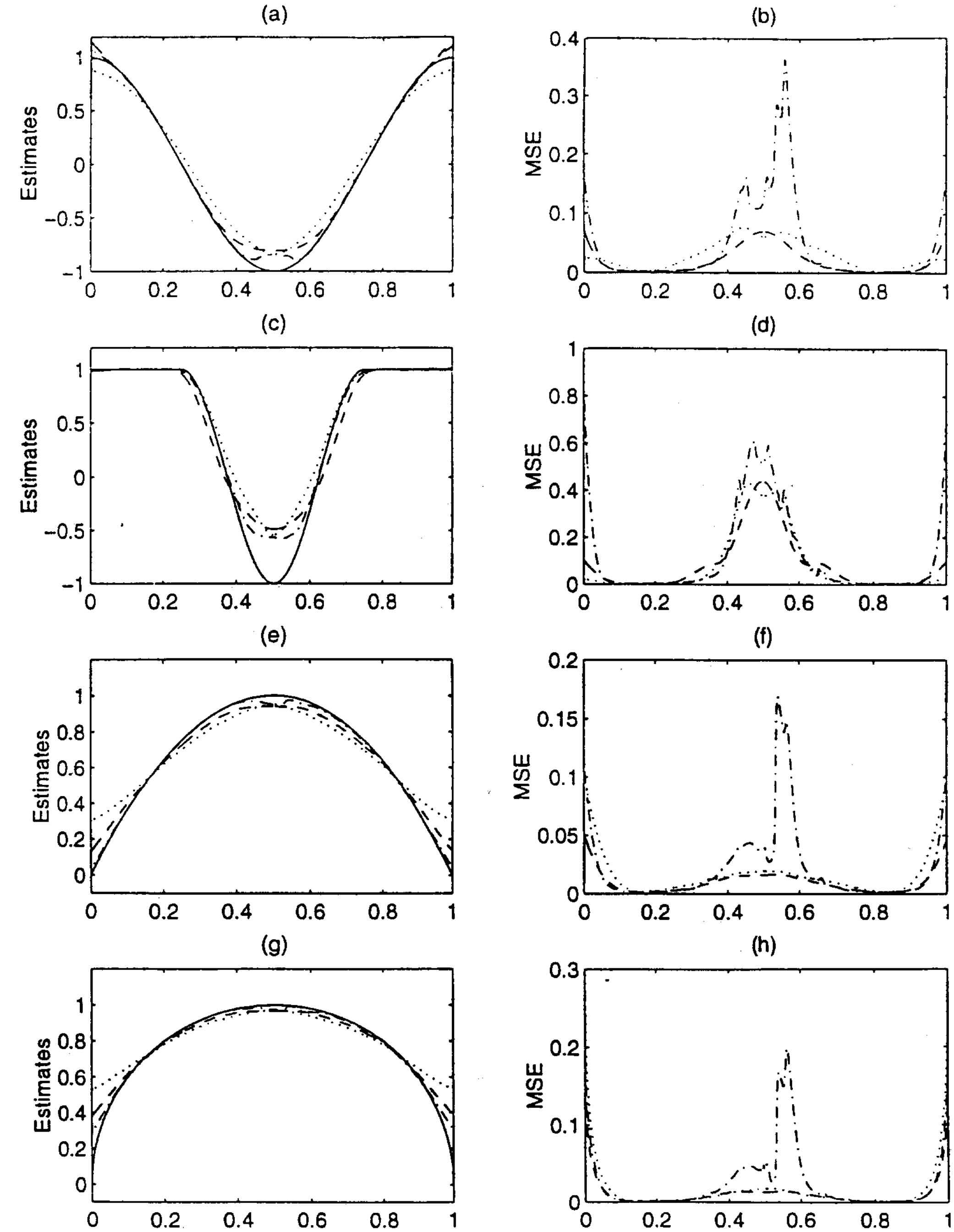

Figure 4. *Estimates of* $E(\widehat{m}(x;h))$ *and* $MSE(\widehat{m}(x;h))$ *under design (d3), and true regression functions. {(a),(b)}, {(c),(d)}, {(e),(f)} and {(g),(h)} are the estimates of {* $E(\widehat{m}(x;h))$, $MSE(\widehat{m}(x;h))$ *} from the regression functions (m1)-(m4), respectively.* $m(x)$*(solid line),* $\widehat{m}_{SNW}(x;h)$*(dashed line),* $\widehat{m}_{NW}(x;h)$*(dotted line) and* $\widehat{m}_{LL}(x;h)$*(dot-dashed line).*

4. Discussion

We have seen, through the simulation study, that $\widehat{m}_{SNW}(x;h)$ is better than the other two estimators when the data are sparse. Also $\widehat{m}_{SNW}(x;h)$ is not bad even when the marginal density of X are uniform compared to other estimators.

Under the designs (d2) and (d3), the bandwidth, $\hat{h}_{DPI}$, is often too small. So we use the interpolant of data in the sparse region. It is essential to use the kernel function with the unbounded support in order to avoid for denominator in $\widehat{m}(x;h)$ to be 0. When the marginal density of X has the unbounded support, the interval [0,1] should be replaced by the range of the design points.

Near the boundaries, there are still bias in all estimators. It could be improved by using the boundary kernel. $l(x)$ and $[\hat{\xi}(0), 2\hat{\xi}(0)]$ in (5) should be studied further to reduce the boundary bias.

References

[1] Cleveland, W. (1979), Robust locally weighted regression and smoothing scatterplots, *Journal of the American Statistical Association*, Vol. 74, 829-836.

[2] Eubank, R. (1988), *Spline smoothing and nonparametric regression*, Marcel Dekker, New York.

[3] Fan, J. (1992), Design-adaptive nonparametric regression, *Journal of the American Statistical Association*, Vol. 87, 998-1004.

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

[5] Fan, J. and Gijbels, I. (1992), Variable bandwidth and local linear regression smoothers, *The Annals of Statistics*, Vol. 20, 2008-2036.

[6] Fan, J. and Marron, J. S. (1994), Fast implementations of nonparametric curve estimators, *Journal of Computational & Graphical Statistics*, Vol 3, 35-56.

[7] Hall, P. and Marron, J. S. (1996), On the role of the ridge parameter in local linear smoothing, to appear *Probability Theory and Related Fields*.

[8] Hall, P., Marron, J. S., Neumann, M. H. and Titterington, D. M. (1997), Curve estimation when the design density is low, *The Annals of Statistics*, Vol. 25, 756-770.

[9] Hall, P. and Wand, M. P. (1994), On the accuracy of binned kernel density estimators, Submitted for publication.

[10] Härdle, W. (1990), *Applied nonparametric regression*, Cambridge University Press.

[11] Mammen, E. and Marron, J. S. (1996), Mass recentered kernel smoother, to appear *Biometrika*.

[12] Müller, H.-G. (1987), Weighted local regression and kernel methods for nonparametric curve fitting, *Journal of the American Statistical Association*, Vol. 82, 231-238.

[13] Müller, H.-G. (1988), *Nonparametric Regression Analysis of Longitudinal Data*, Springer-Verlag, Berlin.

[14] Müller, H.-G. and Song, K.-S. (1993), Identity reproducing nonparametric regression. *Journal of Multivariate Analysis*, Vol. 46, 237-253.

[15] Nadaraya, E. A. (1964), On estimating regression, *Theory of Probability and its Applications*, Vol. 10, 186-90.

[16] Ruppert, D. Sheather, S. J. and Wand, M. P. (1995), An effective bandwidth selector for local least squares regression, *Journal of the American Statistical Association*, Vol. 90, 1257-1270.

[17] Seifert, B. and Gasser, T. (1996), Finite sample variance of local polynomials: analysis and solution, *Journal of the American Statistical Association,* Vol. 91, 267-275.

[18] Silverman, B. W. (1986), *Density estimation for statistics and data analysis,* London: Chapman and Hall.

[19] Scott, D. W. (1992), *Multivariate density estimation: theory, practice and visualization,* Wiley, New York.

[20] Stone, C. J. (1977), Consistent nonparametric regression, *The Annals of Statistics,* Vol. 5, 595-620.

[21] Wand, M. P. and Jones, M. C. (1995), *Kernel Smoothing,* Chapman and Hall.

[22] Watson, G. S. (1964), Smooth regression analysis, *Sankhyā Series, A,* Vol. 26, 101-16.