Journal of the Korean
Data & Information Science Society
2000, Vol. 11, No. 2, pp. 335 ~ 345

A Comparison Study on the Error Criteria in Nonparametric Regression Estimators [1]

Sung. S. Chung [2]

Abstract

Most context use the classical norms on function spaces as the error criteria. Since these norms are all based on the vertical distances between the curves, these can be quite inappropriate from a visual notion of distance. Visual errors in Marron and Tsybakov(1995) correspond more closely to "what the eye sees". Simulation is performed to compare the performance of the regression smoothers in view of MISE and the visual error. It shows that the visual error can be used as a possible candidate of error criteria in the kernel regression estimation.

Key Words and Phrases: Nonparametric Regression Estimation, Local Polynomial Estimator, Visual Error Criteria

1. Introduction

The goal of regression curve fitting is to find a relationship between variables $\{X_i\}_{i=1}^n$ and $\{Y_i\}_{i=1}^n$, where we consider X_i to explain the value of Y_i.

Consider the problem of estimating the regression function $m(x) = E(Y|X = x)$ with the constant variance, $\sigma^2 = \text{var}(Y|X = x)$. Here, $m(x)$ is assumed to be a smooth but unknown function. This regression relationship is commonly modelled as

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

where ϵ_i's are independent random variables with the expectation 0 and the variance σ^2 denoting the variation of Y around $m(X)$.

[1]This research was supported by the Basic Science Research Institute Program, Project No. 1998-015-D00048

[2]Associate Professor, Division of Mathematics and Statistical Informatics, Chonbuk National University Chonju, Chonbuk, 561-756, Korea

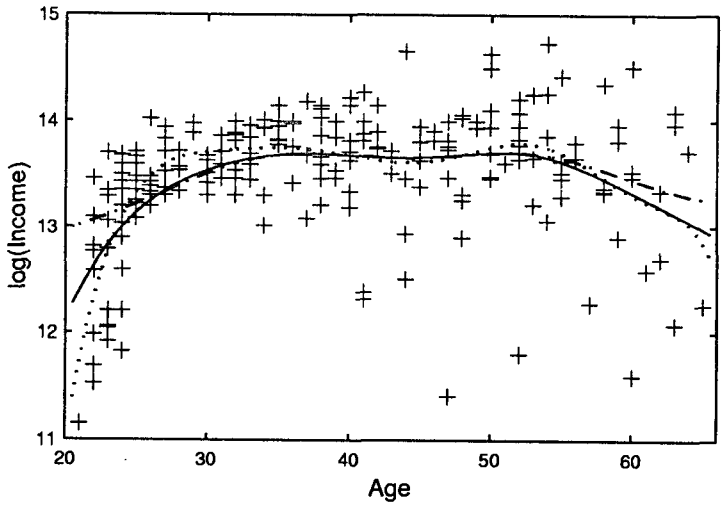
To estimate the regression function nonparametrically, kernel-based smoothers are often used because of their simplicity and implementation. There are a great deal of theoretical research on kernel-based smoothers. See Silverman(1986), Härdle (1990), Wand and Jones(1995) and Fan and Gijbels(1996). An intuitively attractive kernel-based regression estimator is Nadaraya-Watson estimator proposed by Nadaraya(1964) and Watson(1964). Priestley and Chao(1972) and Gasser and Müller(1984) introduced the alternative kernel regression estimators. An improved estimator is local polynomial estimator, proposed by Fan(1992). These estimators are based on moving locally weighted averaging.

As the approach of how to assess the performance of various estimators, most contexts is using the classical mathematical norms on function spaces (e.g., L_1, L_2, or L_∞). Since these norms are all based on the vertical distances between the curves, these can be quite inappropriate from a visual notion of distance. Marron and Tsybakov(1995) developed and investigated alternative measures of error which correspond more closely to "what the eye sees". Their visual error was using both vertical and horizontal information. This can be useful to capture the qualitative features, e.g., number of modes or inflection points. In this paper, their proposed error criterion is studied in cases of the regression problem. Although this criterion works well in a visual sense, it is not ideal for all statistical problem. In particular, if a regression is to be used solely for prediction purposes, then the L_2 fit has important optimality properties and is preferable to the visual error fit.

<Figure 1> shows three well-known estimates using the Gaussian kernel for the Canadian Earning Power Data from Ullah(1985). The raw data $\{(X_i, Y_i)\}_{i=1}^n$, $n = 205$, are overlaid with scatterplot. The bandwidth, h_{RSW}, in <Figure 1> is chosen by a plug-in method propsed by Ruppert, Sheather and Wand(1995). The solid line is the local linear estimate and seems to be a appropriate estimate. <Table 1> shows sum of squared error(SSE) and sum of squared visual error(SVE) of three estimates. Here SVE is defined as follows :

$$SVE = \sum_{i=1}^{n} \inf_{x \in R} ||(x_i, y_i) - (x, \widehat{m}(x; h))||^2$$

where $||\cdot||$ denotes the usual euclidean distance. Therefore, SVE is the sum of the minimum squared distances from each data point to the graph. SVE shows same pattern as SSE. And the local cubic estimate seems best among three estimates with respect to both error criteria but seems to be more wiggly.



<Figure 1> Kernel regression estimates for Canadian Earning Power Data, overlaid with scatterplot of the raw data. Solid line : $\widehat{m}_{LL}(x;h)$ (the local linear estimate), the dotted line : $\widehat{m}_{NW}(x;h)$ (Nadaraya-Watson estimate), and the dashed line : $\widehat{m}_{LC}(x;h)$ (the local cubic estimate). Ruppert, Sheather and Wand bandwidth, h_{RSW}, is used.

<Table 1> SSE and SVE of Kernel regression estimates

	$\widehat{m}_{NW}(x;h)$	$\widehat{m}_{LL}(x;h)$	$\widehat{m}_{LC}(x;h)$
SSE	56.601	53.810	51.626
SVE	56.343	52.609	49.808

In section 2, we describe the kernel-type regression estimators including Nadaraya-Watson estimator and local polynomial estimator. In section 3, the usual norms in L_2 and the visual error criterion are discussed. In section 4, we perform the simulation to compare the performance of four regression smoothers in four testing function. $\widehat{m}_{LC}(x;h)$ and $\widehat{m}_{VE}(x;h)$ has the good properties in aspect to both MISE and the visual error. In particular, $\widehat{m}_{VE}(x;h)$ is most stable, so it can be used as a good candidate in the kernel regression problem.

2. The Nonparametric Regression Smoothers

In this section we consider the kernel-type regression smoothers. Given a set of bivariate data $(X_1, Y_1), \cdots, (X_n, Y_n)$, we want to 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 estimator ($\widehat{m}_{NW}(x;h)$) independently proposed by Nadaraya(1964) and Watson (1964) 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. See Silverman(1986). Ruppert, Sheather and Wand(1995) proposed the "Direct Plug-in" bandwidth, h_{RSW}, which possessed attractive theoretical and practical properties.

As an improved estimator, there is the local polynomial regression estimator(Fan, 1992). This estimate the regression function, $m(x)$, at a particular point by locally fitting a pth degree polynomial to the data via weighted least squares. It is given by the value of $\widehat{b}_0$ when $\mathbf{b} = (b_0, \cdots, b_p)^t$ is chosen to minimize the weighted least squares function

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

Note that $\widehat{m}_{NW}(x;h)$ in (2) is the special case of the local polynomial regression estimator ($p = 0$). When $p = 1$, this is called the local linear regression estimator, and it can be expressed as follows

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

where $\widehat{s}_r(x;h) = n^{-1}\sum_{i=1}^{n}(X_i - x)^r K_h(X_i - x)$, $r = 0, 1, 2$. And when $p = 3$, it is called the local cubic regression estimator ($\widehat{m}_{LC}(x;h)$). Wand and Jones(1995) showed that odd degree polynomial fits had attractive bias and boundary properties and suggested using either the local linear or the local cubic estimator.

3. Error Criteria

The analysis of the performance of the estimator requires the specification of appropriate error criteria for measuring the error when estimating the regression function, $m(x)$, at a single point as well as over the whole real line. The classical approach is using the mathematical norms on the function spaces (e.g., L_1, L_2, or L_∞).

We often consider optimizing L_2 norm (e.g., the mean squared error(MSE), the integrated squared error(ISE) or the mean integrated squared error (MISE)) because of its mathematical simplicity.

MSE is to consider $\widehat{m}(x;h)$ as an estimator of $m(x)$ at a some point $x \in R$. Rather than simply estimating $m(x)$ at a fixed point, it is usually desirable, especially from a data analytic viewpoint, to estimate $m(x)$ over the whole real line. One such error criterion is the integrated squared error(ISE) given by

$$\mathrm{ISE}\{\widehat{m}(\,\cdot\,;h)\} = \int \{\widehat{m}(x;h) - m(x)\}^2 dx.$$

The ISE is appropriate if we are only concerned with data set at hand, but it does not take into account other possible data sets. Therefore, some researchers prefer to work with the expected ISE(MISE),

$$\mathrm{MISE}\{\widehat{m}(\,\cdot\,;h)\} = E\left[\mathrm{ISE}\{\widehat{m}(\,\cdot\,;h)\}\right] = E\int \{\widehat{m}(x;h) - m(x)\}^2 dx.$$

There has been an controversy over which should be called the "optimal bandwidth in the L_2 sense", between the respective minimizers of ISE and MISE. It is seen that despite important mathematical and conceptual differences there is often little difference in term of assessing bandwidth selectors in most practical situations. However MISE does give better resolution in comparing bandwidth selectors than ISE. So we use MISE in the simulation. See Grund, Hall and Marron(1992) for references and discussion.

The classical mathematical norms on the function spaces are based on the vertical distance between the two functions. But the eye does not works in this way. Marron and Tsybakov(1995) proposed the visual error criteria which can be appropriate from a graphical viewpoint. See that paper for the motivation and the detailed discussion.

A continuous function $m : [a,b] \to R$ can be represented by its "graph",

$$G_m = \{(x,y) : x \in [a,b], y = m(x)\} \subset R^2.$$

The shortest distance from the given point (x,y) to the graph G can be represented as follows:

$$d((x,y),G) = \inf_{(x',y')\in G} ||(x,y) - (x',y')||.$$

Then the asymmetric visual error criterion proposed by Marron and Tsybakov(1995) is

$$\mathrm{VE}(\widehat{m}_h \to m) = \Big[\int_a^b \{d((x,\widehat{m}(x;h)),G_m)\}^2 dx\Big]^{1/2}. \tag{3}$$

Lemma.

$$\begin{aligned} E(\mathrm{VE}(\widehat{m}_h \to m)^2) &= \int_a^b E\{d((x,\widehat{m}(x;h)), G_m)^2\}dx \\ &\approx \int_a^b \frac{E((\widehat{m}(x;h)-m(x))^2)}{1+(m'(x))^2}dx \\ &\approx \int_a^b \frac{b^2(x)+\sigma^2(x)}{1+(m'(x))^2}dx \end{aligned}$$

where $b(x) = \frac{h^2}{2}m''(x)\int_a^b u^2K(u)du$ and $\sigma^2(x) = \frac{\sigma^2}{nhf(x)}\int_a^b K(u)^2du$.

Proof. Similar to Marron and Tsybakov(1995).

Note that there is less error at x where $m(x)$ is steep (i.e., $|m'(x)|$ is large), but roughly the usual error at locations where $m(x)$ is flat (i.e., $m'(x) \approx 0$). And if we choose $h = O(n^{-1/5})$ as the bandwidth, $\mathrm{EVE} = E(\mathrm{VE}(\widehat{m}_h \to m)^2) = O(n^{-4/5})$, same asymptotic rate of MISE.

For computation of MISE and EVE, we need discretizing the continuous function. Denote a set of an equally spaced compact grid (with grid spacing Δx , say) of x locations by $\aleph$ and the discretized version of a graph G_m of a function $m(x)$ by $G_m^{\mathrm{discr}} = \{(x, m(x)) : x \in \aleph\}$.

The discretized version of $\mathrm{VE}(\widehat{m}_h \to m)$ is given by

$$\mathrm{VE}^{\mathrm{discr}}(\widehat{m}_h \to m) = \Big[\Delta x \sum_{x\in\aleph}\{d((x,\widehat{m}(x;h)), G_m^{\mathrm{discr}})\}^2\Big]^{1/2}.$$

Since the error criteria, VE, depends heavily depend on the relative scale of x and y, the figure will give results different from visual impression. For this reason it need linearly transforming x and y so that both $[a,b]$ and $[\inf_{x\in\aleph}m(x), \sup_{x\in\aleph}m(x)]$ are mapped to $[0,1]$.

4. Simulation

In this section, we carry out Monte Carlo simulation to compare the performance of the various regression smoothers in view of MISE and EVE.

We consider four testing regression functions as follows

$$\begin{aligned} &\text{(m1)} \quad m(x) = \sin(2\pi x). \\ &\text{(m2)} \quad m(x) = \sin(4\pi x). \\ &\text{(m3)} \quad m(x) = x(1-x). \\ &\text{(m4)} \quad m(x) = \frac{1}{0.1\sqrt{2\pi}}\exp\Big\{-\frac{(x-0.5)^2}{2(0.1)^2}\Big\}. \end{aligned}$$

We choose the function (m1) as a standard function since it is smoothed and easy to estimate. The function (m2) is more wiggly and (m3) is more smoothed in the interior region but has the more sharp end point. (m4) has a peak in the center and is flat near the boundary.

The sample size is taken as $n = 100$ and 500 replications are performed to estimate MISE and EVE. And we take the equally spaced fixed design because the sample size is relatively small to 401 grid points. We generate $\epsilon_i \sim N(0, \sigma^2)$ with $\sigma = 0.2$ from randn function of MATLAB. We use the Gaussian kernel as the kernel function, and the direct plug-in bandwidth selector, h_{RSW}, of Ruppert, Sheather and Wand(1995) as the bandwidth for $\hat{m}_{NW}(x;h), \hat{m}_{LL}(x;h)$ and $\hat{m}_{LC}(x;h)$. For $\hat{m}_{VE}(x;h)$, we apply the minimizer, h_{VE}, of $\mathrm{VE}(\hat{m}_h \to m)$ in (3). But as the bandwidth goes to zero, the estimate goes to interpolation of data points. So we use the local linear estimators as the function class of $\hat{m}_{VE}(x;h)$. Therefore, $\hat{m}_{VE}(x;h)$ is the estimate minimizing $\mathrm{VE}(\hat{m}_h \to m)$ among the local linear estimates.

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, and this allows very fast computation of $\hat{m}(x;h)$ 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" by Hall and Wand(1993).

<Table 2> shows the estimates of MISE and EVE of four regression smoothers. $\hat{m}_{NW}(x;h)$ and $\hat{m}_{LL}(x;h)$ have almost similar errors. The errors of $\hat{m}_{LC}(x;h)$ are very similar to those of $\hat{m}_{VE}(x;h)$ and are less than those of $\hat{m}_{NW}(x;h)$ and $\hat{m}_{LL}(x;h)$, as shown in Wand and Jones(1995). As we can expect, EVE is smaller than MISE. But the differences in (m3) are due to the linear transformation from $[\inf_{x\in\aleph} m(x), \sup_{x\in\aleph} m(x)]$ to $[0,1]$ in VE. As we can see in (m3) of <Figure 2> and <Figure 3>, $\hat{m}_{LC}(x;h)$ is unstable because it has largest errors(small bias but relatively large variance) among four estimators. This seems due to overfitting of $\hat{m}_{LC}(x;h)$ since (m3) is the quadratic function.

<Figure 2> shows the kernel regression estimates under four regression functions (m1)-(m4), respectively. $\hat{m}_{NW}(x;h)$ and $\hat{m}_{LL}(x;h)$ can not be discerned at almost middle points. $\hat{m}_{VE}(x;h)$ (dash-dotted line) has largest biases in the middle region in (m3) but it has smallest $MSE(\hat{m}(x;h))$ in <Figure 3>. This says that it has smallest variances and the biases are dominated by the variances. Except (m3), $\hat{m}_{LC}(x;h)$ and $\hat{m}_{VE}(x;h)$ have less biases than $\hat{m}_{NW}(x;h)$ and $\hat{m}_{LL}(x;h)$, especially in the region having the high curvature, i.e., the peak and the valley. This corresponds to the result of Wand and Jones(1995). <Figure 3> shows the estimates of $MSE(\hat{m}(x;h))$'s of four estimates under the regression functions (m1)-(m4), respectively. It shows that $MSE(\hat{m}(x;h))$'s of $\hat{m}_{NW}(x;h)$ and $\hat{m}_{LL}(x;h)$ have same features with the biases except (m3). $MSE(\hat{m}(x;h))$'s have similar pattern

as the biases of <Figure 2> except (m3). So we can say that the variances are dominated by the biases. The above discussion says that $\widehat{m}_{LC}(x;h)$ and $\widehat{m}_{VE}(x;h)$ are better than $\widehat{m}_{NW}(x;h)$ and $\widehat{m}_{LL}(x;h)$ for estimating the true function. And $\widehat{m}_{VE}(x;h)$ has good performance in all regression function including (m3) with respect to $MSE(\widehat{m}(x;h))$.

< Table 2 > Estimates of MISE and EVE_2

Error criteria	Estimator	Regression function			
		m1	m2	m3	m4
$\widehat{\text{MISE}}$	$\widehat{m}_{NW}(x;h)$	0.0069	0.0159	0.0014	0.0647
	$\widehat{m}_{LL}(x;h)$	0.0068	0.0160	0.0014	0.0647
	$\widehat{m}_{LC}(x;h)$	0.0028	0.0041	0.0023	0.0067
	$\widehat{m}_{VE}(x;h)$	0.0032	0.0056	0.0010	0.0077
$\widehat{\text{EVE}}$ $\times 10^2$	$\widehat{m}_{NW}(x;h)$	0.0576	0.0392	0.4679	0.0441
	$\widehat{m}_{LL}(x;h)$	0.0581	0.0369	0.4706	0.0435
	$\widehat{m}_{LC}(x;h)$	0.0182	0.0137	0.7225	0.0123
	$\widehat{m}_{VE}(x;h)$	0.0207	0.0146	0.3025	0.0144

5. Discussion

We have seen through the simulation that $\widehat{m}_{LC}(x;h)$ and $\widehat{m}_{VE}(x;h)$ have better performance than the other two estimates except (m3). In (m3), $\widehat{m}_{LC}(x;h)$ has much larger variance than other estimates. $\widehat{m}_{VE}(x;h)$ shows the good performance in all regression functions with respect to MISE and EVE. Therefore the visual error criteria can be used as a good candidate in assessing the kernel regression estimators since it correspond more closely to what the eye sees. It should be studied further to select the data-driven bandwidth minimizing EVE.

References

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

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

3. Fan, J. and Gijbels, I. (1996). *Local Polynomial Modelling and its Applications,* Chapman and Hall, London

4. Gasser, T. and Müller, H.-G. (1984). Estimating of regression functions and their derivatives by the kernel method, *Scandinavian Journal of Statistics,* 11, 171-185.

5. Grund, B., Hall, P. and Marron, J. S. (1995). Loss and risk in smoothing parameter selection, unpublished manuscript.

6. Hall, P. and Wand, M. P. (1993). On the accuracy of binned kernel density estimators. Working Paper 93-103, *Australian Graduate School of Management, University of New South Wales.*

7. Härdle, W. (1990). *Applied nonparametric regression,* Cambridge University Press.

8. Marron, J. S. and Tsybakov, A. B. (1995). Visual Error Criteria for Qualitative Smoothing, *Journal of the American Statistical Association,* 90, 499-507.

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

10. Priestley, M. B. and Chao, M. T. (1972). Non-parametric function fitting, *Journals of the Royal Statistical Society,* Series, B, 34, 385-392

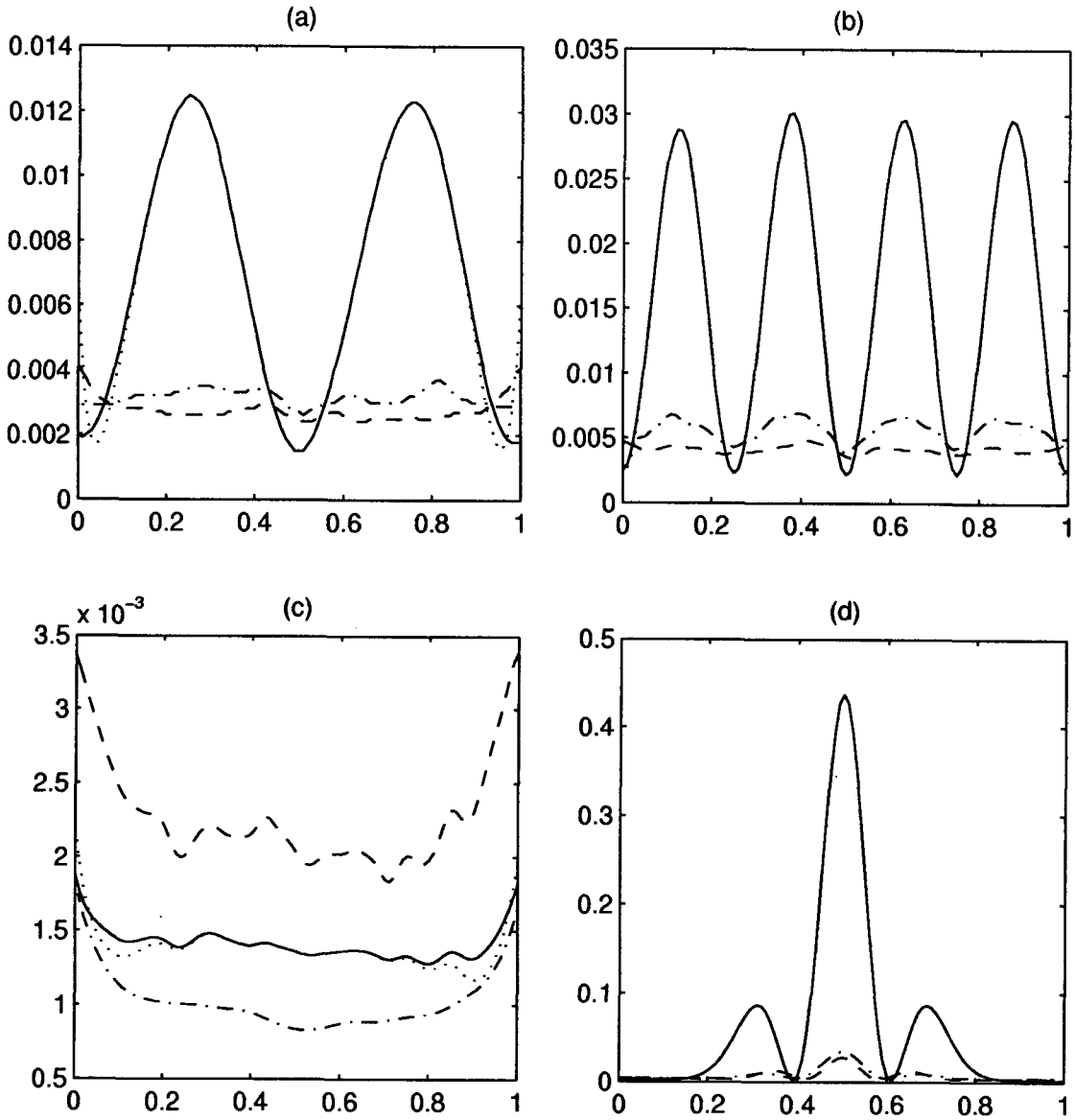
11. Ruppert, D., Sheather, S. J. and Wand, M. P. (1995). An effective bandwidth selector for local least squares regression, *Journal of the American Statistical Association,* 90, 1257-1270.

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

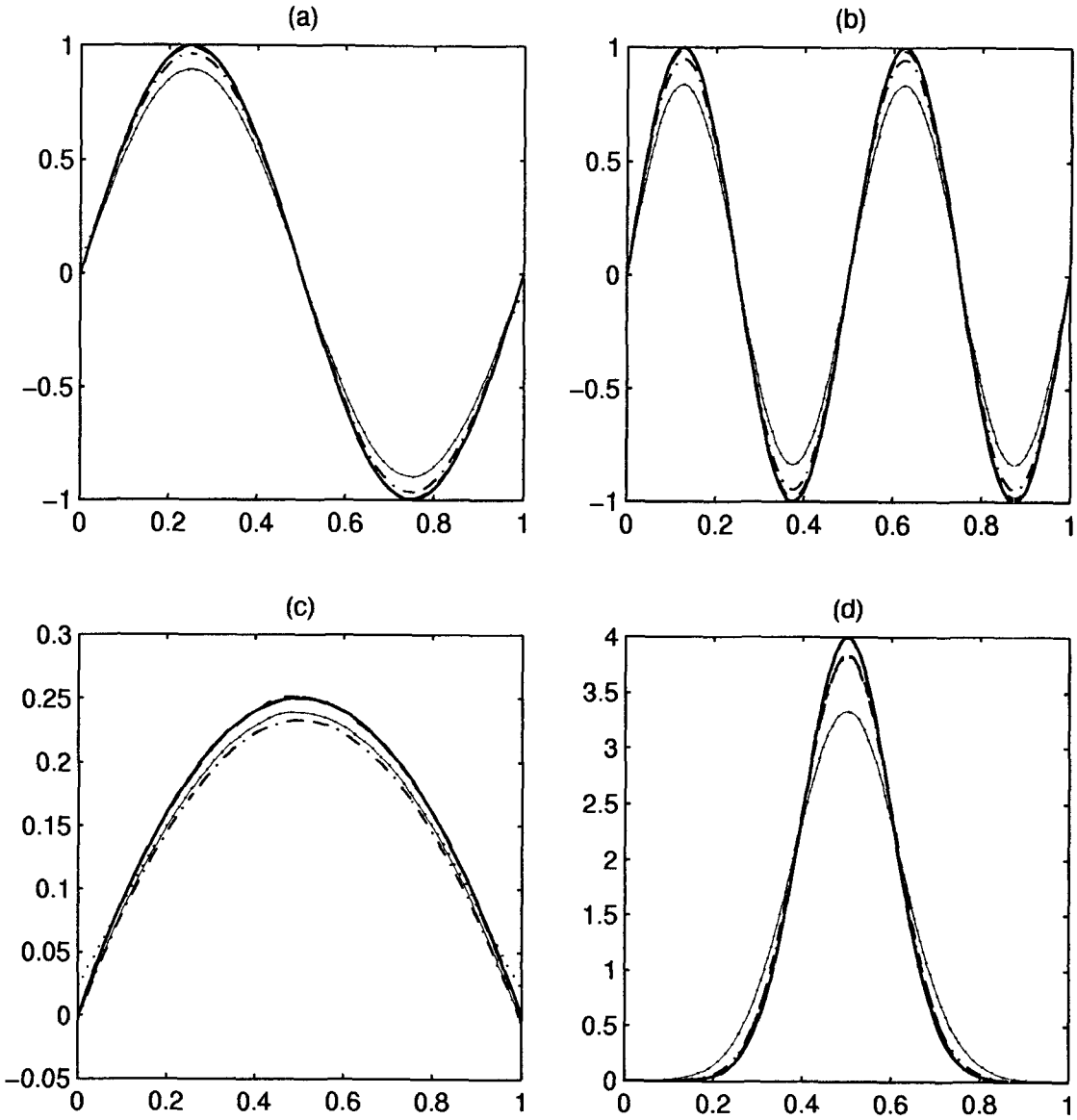
13. Ullah, A. (1985). Specification analysis of economietric models, *Journal of Quantitative Economics,* 2, 187-209.

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

15. Watson, G. S. (1964). Smooth regression analysis, *Sankhyā Series,* A, 26, 101-16.



<Figure 2> Kernel regression estimates under four regression functions (m1)-(m4), respectively. Heavy solid line(true function), light solid line ($\widehat{m}_{LL}(x;h)$) , dotted line ($\widehat{m}_{NW}(x;h)$), dashed line($\widehat{m}_{LC}(x;h)$) and dash-dotted line($\widehat{m}_{VE}(x;h)$). h_{RSW} is used.



<Figure 3> Estimates of $MSE(\widehat{m}(x;h))$ under four regression functions (m1)-(m4), respectively. Solid line ($\widehat{m}_{LL}(x;h)$), dotted line ($\widehat{m}_{NW}(x;h)$), dashed line ($\widehat{m}_{LC}(x;h)$) and dash-dotted line ($\widehat{m}_{VE}(x;h)$). h_{RSW} is used.