TAIWANESE JOURNAL OF MATHEMATICS
Vol. 9, No. 1, pp. 123-141, March 2005
This paper is available online at http://www.math.nthu.edu.tw/tjm/

INTEGRATED CROSS-VALIDATION FOR THE RANDOM DESIGN NONPARAMETRIC REGRESSION

Tzu-Kuei Chang, Wen-Shuenn Deng, Jung-Huei Lin, and C. K. Chu

Abstract. For the random design nonparametric regression, cross-validation is a popular bandwidth selector. It is constructed by using the criterion of "weighted" integrated square error. In practice, however, the weighting scheme by the design density in the criterion causes that its associated cross-validation function puts more emphasis in regions with more data, gives little attention to regions with few data, but has no consideration for regions without data. In such a case, the value of the cross-validated bandwidth depends on the distribution of the design points, but is independent of the location of the interval on which the regression function value is estimated. Hence, if there are sparse regions in the realization of the design, then the resulting cross-validated bandwidth is usually not large enough in magnitude such that its corresponding kernel regression function estimate has rough appearance in these sparse regions. To avoid this drawback to cross-validation, we suggest using the criterion of "unweighted" integrated square error to construct the bandwidth selector. Under the criterion, a bandwidth selector called integrated cross-validation is proposed, and the resulting bandwidth is shown to be asymptotically optimal. Empirical studies demonstrate that the kernel regression function estimate obtained by using our proposed bandwidth is better than that employing the ordinary cross-validated bandwidth, in both senses of having smoother appearance and yielding smaller sample unweighted integrated square error.

1. Introduction

In the field of kernel regression function estimation, it is well known that choosing a suitable value of bandwidth is the essence of the smoothing problem. See the

Received April 15, 2003; Revised August 20, 2003.
Communicated by Yuh-Jia Lee.
2000 *Mathematics Subject Classification*: Primary 62G05; secondary 62G20.
Key words and phrases: bandwidth selection, cross-validation, integrated cross-validation, nonparametric regression, sparse design, unweighted integrated square error, weighted integrated square error.

works by Eubank (1988), Müller (1988), Härdle (1990, 1991), Scott (1992), Wand and Jones (1995), Fan and Gijbels (1996), and Simonoff (1996) for a detailed introduction of the kernel regression function estimator. For independent observations, cross-validation introduced by Clark (1975) is an extremely popular data-driven bandwidth selector. It is constructed by using the criterion of weighted integrated square error (WISE) of the kernel regression function estimator. For asymptotic properties of the cross-validated bandwidth and asymptotic equivalence of some popular data-driven bandwidth selectors to cross-validation, see for example Rice (1984), Härdle and Marron (1985), and Härdle, Hall, and Marron (1988). For other bandwidth selectors, see also Marron (1988), a survey paper, and inferences cited therein.

However, in practice, the weighting scheme by the design density in the WISE criterion has an adverse effect. Its associated cross-validation function puts more emphasis in regions with more data, gives little attention to regions with few data, and no consideration for regions without data. In such a case, the magnitude of the cross-validated bandwidth depends on the distribution of the design points, but is independent of the location of the interval on which the regression function value is estimated. Hence, if there are sparse regions in the realization of the design, then the resulting cross-validated bandwidth is usually not large enough in magnitude such that its corresponding kernel regression function estimate has rough appearance in these sparse regions.

This drawback to ordinary cross-validation (OCV) is illustrated in Figure 1 using the shampoo data (Bayhan and Bayhan 1998). Figure 1a shows that the kernel regression function estimate produced by employing the ordinary cross-validated bandwidth has rough appearance; hence, it is difficult to explain the economic sense between the two variables considered by using this kernel estimate. The result is due to the fact that the magnitude of this ordinary cross-validated bandwidth $\hat{h}_{OCV} = 1.14$ is not large enough since it is less or slightly larger than the six largest spacings 2.5, 2.2, 1.6, 1.1, 1.1, and 1.1 among the design points and the boundary points of the interval on which value of the regression function is estimated. The formulations and the computation procedures for quantities in Figure 1 will be introduced in Sections 2 and 4, respectively.

To avoid the above drawback to OCV, we suggest using the criterion of "unweighted" integrated square error (UISE) to construct the bandwidth selector. This UISE criterion puts equal emphasis at each point on which the regression function value is estimated. Using the criterion, a bandwidth selector called integrated cross-validation (ICV) is proposed. The ICV function is constructed by the same idea of the OCV function in Härdle and Marron (1985). First, the unknown regression function value in the UISE criterion is replaced with its kernel estimate using the k-nearest neighbor (kNN) bandwidth (Härdle 1990). Then, to avoid the problem

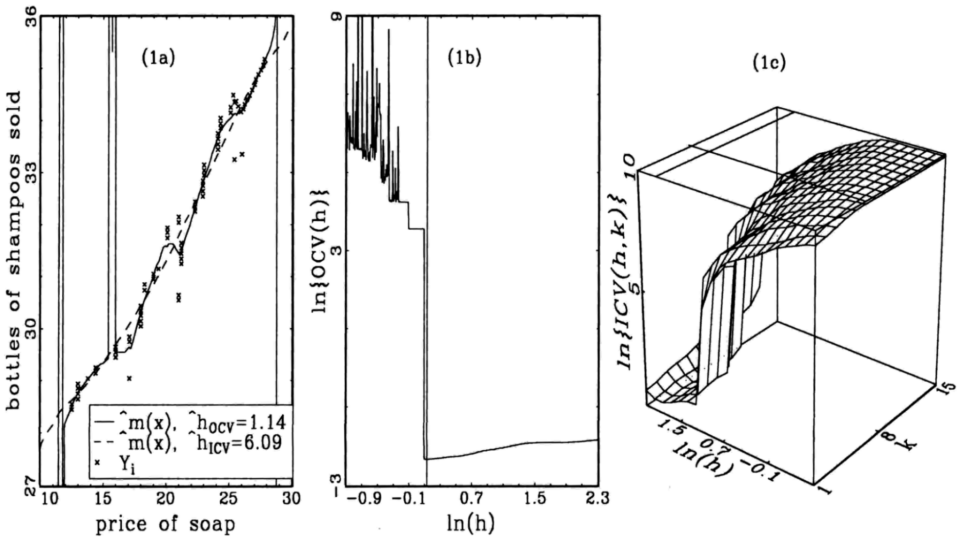

Fig. 1. Plot of the shampoo data of size $n = 75$ and their kernel regression function estimates (1a), the ordinary cross-validation function (1b), and the integrated cross-validation function (1c). The vertical line in (1b) and the solid lines on the top of (1c) stand for the locations of minimizers of their associated functions. For better visual performance, both the ordinary and the integrated cross-validations are expressed by taking natural logarithm.

of using the same data to both construct and assess the kernel regression function estimate, the kernel estimate using the global bandwidth originally in the UISE criterion is taken as its "leave-kNN-out" version. The minimizer of the ICV function over the global bandwidth is taken as the integrated cross-validated bandwidth. For the related modified cross-validation criterion in the case of dependent observations, see for example Chu and Marron (1991a) and references cited therein. The same idea of UISE criterion has been used in the field of kernel density estimation to produce the least-squares cross-validated bandwidth (Silverman 1986).

Figure 1a shows that the kernel regression function estimate using our integrated cross-validated bandwidth has smoother appearance and makes nice economic senses. For example, it shows a direct linear relationship between the two variables considered. Since soap and shampoo are close substitutes, when the price of soap goes up, people use more shampoo to replace soap.

The organization of this paper is as follows. Section 2 describes the regression settings and gives the precise formulation of our suggested bandwidth. Section 3 shows that the integrated cross-validated bandwidth is asymptotically optimal for the UISE criterion. Section 4 contains empirical results which demonstrate that ICV is better than OCV in the sense of yielding both smoother appearance and smaller sample UISE for the kernel regression function estimate. Finally, sketches of the

proofs are given in Section 5.

2. Regression Settings and Bandwidth Selectors

In this paper, the random design nonparametric regression model is given by

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

for $i = 1, \cdots, n$. Here (X_i, Y_i) are independently and identically distributed bivariate random vectors, and ϵ_i are unobservable regression errors with mean 0 and variance σ^2, $0 < \sigma < \infty$. For the sake of simplicity, the design points X_i are assumed to have the probability density function f supported on the bounded region [0,1], and are assumed to be independent of the regression errors ϵ_i. The purpose of the regression is to use the data points (X_i, Y_i) to estimate the regression function $m(x)$, for each $x \in [0, 1]$.

To estimate the regression function m, the local linear estimator (LLE; Fan 1992, 1993) using the global bandwidth is considered. It has many advantages. For example, it achieves full asymptotic minimax efficiency among all linear estimators (Fan 1993), has a nice asymptotic bias quality and a superior asymptotic variance quantity (Gasser and Engel 1990, Chu and Marron 1991b, Wu and Chu 1992), and adapts automatically to the boundary of the support of the design density (Fan and Gijbels 1992). However, it has a disadvantage of having unbounded finite sample conditional variance when a kernel function with compact support is used (Seifert and Gasser 1996). Compactly supported kernels are often employed for computational convenience (Härdle 1991) and for optimal performance (Epanechnikov 1969). This adverse effect causes that the regression function estimate produced by the LLE sometimes has rough appearance.

We now give the formulation of the LLE. For simplicity of presentation, assume that the regression function m has two derivatives. Given the kernel function K as a probability density function supported on [-1,1] and the bandwidth $h = h_n$ tending to 0 as $n \to \infty$, the LLE $\hat{m}(x)$ for $m(x)$ is defined by

$$\hat{m}(x) = \{T_0(x)S_2(x) - T_1(x)S_1(x)\}/\{S_0(x)S_2(x) - S_1^2(x)\},$$

for each $x \in [0, 1]$. Here

$$T_j(x) = n^{-1}h^{-1}\sum_{i=1}^{n} Y_i Z_i^j K(Z_i),\ S_j(x) = n^{-1}h^{-1}\sum_{i=1}^{n} Z_i^j K(Z_i),\ Z_i = (x - X_i)/h,$$

for $j \geq 0$. If the denominator of $\hat{m}(x)$ is 0, take $\hat{m}(x) = 0$.

For constructing $\hat{m}(x)$, the optimal value h_U of h is taken as the minimizer of the UISE of $\hat{m}(x)$ in this paper. For each given value of h, the UISE of $\hat{m}(x)$ is

defined by

$$d_U(h) = \int_0^1 \{\hat{m}(x) - m(x)\}^2 dx.$$

In practice, however, the value of h_U is not available because the quantity depends on the unknown function $m(x)$. Since the value of h_U can not be calculated, our ICV is designed with the purpose of providing an estimate of h_U.

We now give the motivation of ICV. Decompose $d_U(h)$ into

$$\int_0^1 \hat{m}^2(x)dx + (-2)\int_0^1 \hat{m}(x)m(x)dx + \int_0^1 m^2(x)dx. \tag{2.2}$$

Under some regularity conditions, the first term in (2.2) may be approximated by $\int_0^1 \hat{m}_k^2(x)dx$, and the second term may be estimated by (-2)$\int_0^1 \hat{m}_k(x)\ \hat{Y}_k(x)dx$. On the other hand, the third term in (2.2) is independent of h, and may be replaced by another term $\int_0^1 \hat{Y}_k^2(x)dx$ still independent of h. Here k is a given positive integer, $\hat{m}_k(x)$ is the "leave-kNN-out" version of $\hat{m}(x)$, that is, the observations $\{X_{(i)}, Y_{(i)}\}$ for $1 \le i \le k$ are left out in constructing $\hat{m}(x)$, and $\hat{Y}_k(x)$ is $\hat{m}(x)$ using only the data deleted by $\hat{m}_k(x)$, where for each $x \in [0,1]$, $X_{(i)}$ denote the rearranged X_i such that the values of $|x - X_i|$ are in ascending order, and $Y_{(i)}$ are the response variables corresponding to the design points $X_{(i)}$. Note that $\hat{Y}_k(x)$ is exactly $\hat{m}(x)$ using the kNN bandwidth $\varphi_k(x) = |x - X_{(k)}|$ for each x, where the window width varies with location, and it has the same number of design points in each window. Hwang (1995) shows that the regression function estimator $\hat{Y}_k(x)$ for $m(x)$ has similar advantages of the LLE. For example, it does not suffer from boundary effects. Combining these results, our ICV function is taken as

$$ICV(h,k) = \int_0^1 \{\hat{m}_k(x) - \hat{Y}_k(x)\}^2 dx.$$

Let $(\hat{h}_{ICV}, \hat{k}_{ICV})$ be the minimizer of $ICV(h,k)$ over (h,k), and $\hat{h}_{ICV}$ is called the integrated cross-validated bandwidth for h_U. By subtracting and adding the term $m(x)$, it will be shown in Section 5 that $ICV(h,k)$ approaches $d_U(h)+d_U\{\varphi_k(x)\}$, as the sample size n increases, where $d_U\{\varphi_k(x)\} = \int_0^1 \{\hat{Y}_k(x) - m(x)\}^2 dx$. By this result, it can be seen that $\hat{h}_{ICV}$ approaches h_U in some mode, and its asymptotic behavior is independent of the value of k. The asymptotic behavior of $\hat{h}_{ICV}$ will be studied in Section 3.

We now close this section by giving the OCV criterion for the purpose of comparison. It is constructed by taking the optimal bandwidth h_W as the minimizer of the WISE of $\hat{m}(x)$. For each given value of h, the WISE of $\hat{m}(x)$ is defined by

$$d_W(h) = \int_0^1 \{\hat{m}(x) - m(x)\}^2 f(x)dx.$$

In the uniform design case, $d_W(h) = d_U(h)$. The OCV function is taken as

$$OCV(h) = n^{-1} \sum_{i=1}^{n} \{\hat{m}_1(X_i) - Y_i\}^2.$$

Here $\hat{m}_1(X_i)$ is $\hat{m}_k(X_i)$ with $k = 1$. The ordinary cross-validated bandwidth $\hat{h}_{OCV}$ is taken as the minimizer of $OCV(h)$ over h. By subtracting and adding the term $m(X_i)$, $OCV(h)$ approaches $d_W(h) + \sigma^2$, as the sample size n increases, hence $\hat{h}_{OCV}$ converges to h_W in some mode.

3. Theoretical Results

In this section, we shall study the asymptotic behavior of $\hat{h}_{ICV}$. For this, in addition to the assumptions given in Section 2, we impose the following assumptions:

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

(A2) The design density f is Lipschitz continuous and bounded away from zero on the interval [0,1].

(A3) The regression errors ϵ_i are independently and identically distributed random variables with mean 0, variance σ^2, and all other moments finite.

(A4) The kernel function K is a Lipschitz continuous and symmetric probability density function with support [-1,1].

(A5) The values of h and k are selected on the intervals $H_n = [\rho\, n^{-1+2\rho},\, \rho^{-1} n^{-2\rho}]$ and $K_n = [\rho\, n^{\rho},\, \rho^{-1} n^{\rho}]$, respectively, where ρ is an arbitrarily small positive constant.

(A6) The total number of observations in this regression setting is n, with $n \to \infty$.

Following Shibata (1981), $\hat{h}_{ICV}$ is said to be asymptotically optimal with respect to the UISE criterion if

$$\lim_{n\to\infty} \{d_U(\hat{h}_{ICV}) \;/\; \inf_{h\in H_n} d_U(h)\} = 1$$

with probability one. The following Theorem 3.1 gives such optimality of $\hat{h}_{ICV}$, and its proof will be given in Section 5.

Theorem 3.1. *Given the regression model (2.1), if the assumptions (A1)-(A6) hold, then $\hat{h}_{ICV}$ is asymptotically optimal with respect to the UISE criterion.*

4. Empirical Results

To evaluate the performance of our integrated cross-validated bandwidth, empirical studies were carried out. Simulation studies and real data examples are given respectively in Subsections 4.1 and 4.2.

4.1 Simulations

In this subsection, a simulation study was performed to compare the performance of OCV and ICV. The simulation settings were as follows. Four regression functions $m_1(x) = x^3(1-x)^3$ with $\sigma = 0.003$ (Rice 1984), $m_2(x) = (3/10)\exp\{-4(4x-1)^2\} + (7/10)\exp\{-16(4x-3)^2\}$ with $\sigma = 1/10$ (Fan and Gijbels 1995), $m_3(x) = \sin(5\pi x)$ with $\sigma = 1/2$ (Ruppert, Sheather, and Wand 1995), and $m_4(x) = 2 - 5x + 5\ \exp\{-400(x-1/2)^2\}$ with $\sigma = (1/2)^{1/2}$ (Seifert and Gasser 1996) were chosen. Four design densities supported on [0,1], including f_1 : Uniform[0,1], f_2 : Normal$\{1/2, (1/3)^2\} \cap [0,1]$ (truncated-normal design), $f_3(x) = 4(1-b)|x-1/2| + b$ with $b = 1/5$ (central-hole design), and f_4 : Beta$(1/2, 1)$ (uniform-square design), were employed. These design densities have been considered by Seifert and Gasser (1996), Hall and Turlach (1997), and Deng, Chu, and Cheng (2001) for studying the performance of the LLE. For each regression function and each design density, three sample sizes $n = 50, 100$, and 200 were considered. For each setting, the regression errors ϵ_i were taken as the Normal$(0, \sigma^2)$ variables, and 500 independent sets of observations (X_i, Y_i) were generated by using the regression model (2.1). The kernel function K used by the LLE $\hat{m}(x)$ was the Epanechnikov kernel $K(u) = (3/4)(1-u^2)I_{[-1,1]}(u)$. It is the optimal kernel for constructing $\hat{m}(x)$ (Fan, Gasser, Gijbels, Brockmann, and Engel 1993).

Given each data set, the values of $ICV(h,k)$ were calculated on the equally spaced logarithmic grid of 200 values of h in [0.02,0.5] and $k = 1, \cdots, [n/5]$, and those of $d_U(h)$ and $OCV(h)$ were computed on the same grid of h employed by $ICV(h,k)$, where the notation $[x]$ denotes the integer part of x. 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 values of h and k, the values of $d_U(h)$ and $ICV(h,k)$ were approximated respectively by $(1/u)\sum_{i=1}^{u}\{\hat{m}(t_i) - m(t_i)\}^2$ and $(1/u)\sum_{i=1}^{u}\{\hat{m}_k(t_i) - \hat{Y}_k(t_i)\}^2$, where $t_i = (2i-1)/(2u)$ and $u = 500$. After evaluation on the grid, the global minimizers h_U of $d_U(h)$, $\hat{h}_{OCV}$ of $OCV(h)$, and $(\hat{h}_{ICV}, \hat{k}_{ICV})$ of $ICV(h,k)$ were taken on the grid. In our simulation study, the value of $\hat{k}_{ICV}$ derived from each data set is less than the right boundary point $[n/5]$ of the grid of k.

When the values of h_U over the 500 pseudo data sets were obtained, the sample average and standard deviation of their corresponding $d_U(h_U)$ were calculated. The former quantity measures the best performance of $\hat{m}$. On the other hand, the sample average of $d_U(\hat{h}_{OCV})$ and that of $d_U(\hat{h}_{ICV})$ over the 500 pseudo data sets measure

the performance of $\hat{m}(x)$ which can be obtained in practice by using the ordinary and the integrated cross-validated bandwidths, respectively. The simulation results are summarized in the following figures and tables.

Given the sample size $n = 200$ and the regression function $m_1(x)$, the practical performance of $\hat{m}(x)$ using the ordinary cross-validated bandwidth $\hat{h}_{OCV}$ and that

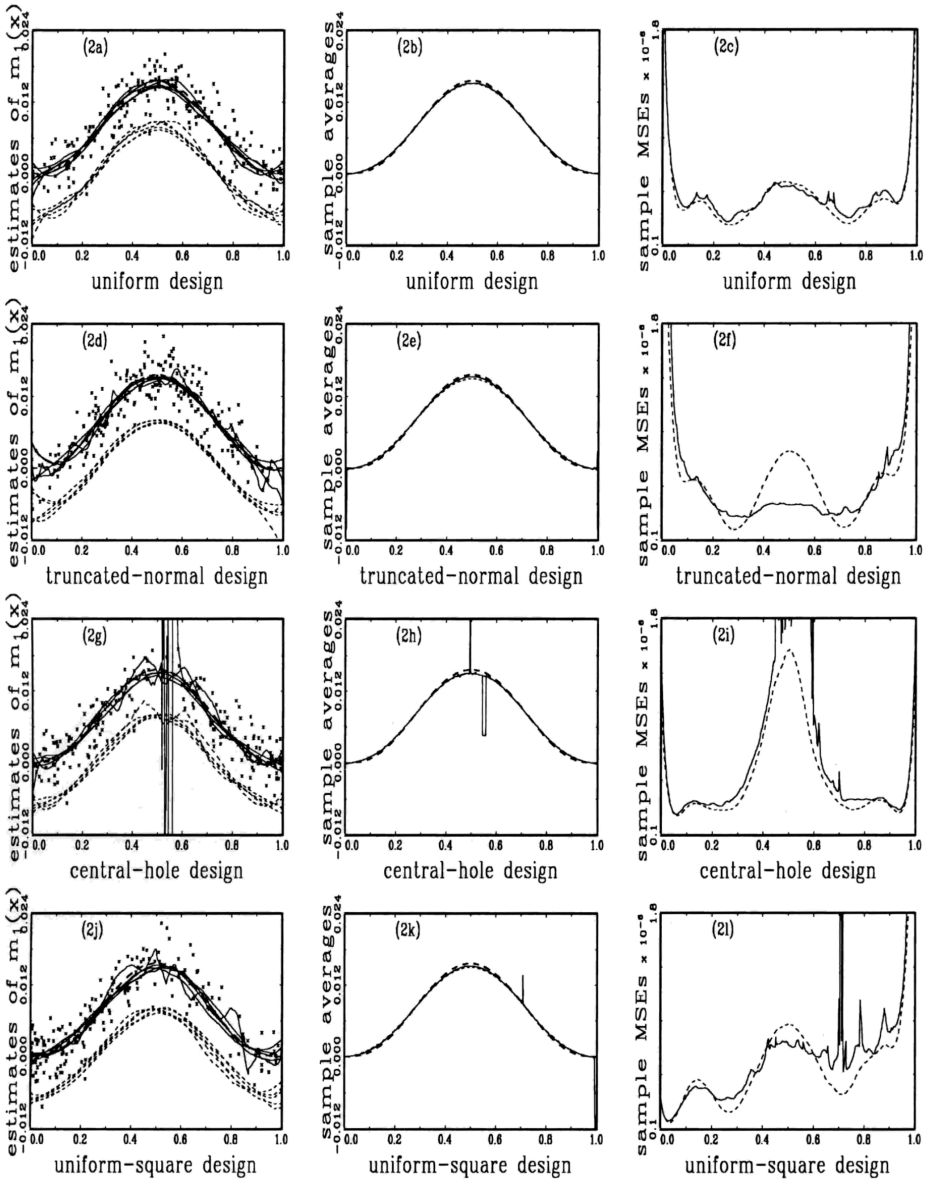

Fig. 2. Practical performance for $m_1(x)$.

employing the integrated cross-validated bandwidth $\hat{h}_{ICV}$ are presented in Figure 2. Figure 2a plots the regression function $m_1(x)$ (bold-faced dashed curve), one simulated data set (stars) with the uniform design, and 5 regression function estimates derived from 5 sets of the simulated data by $\hat{m}(x)$ using $\hat{h}_{OCV}$ (solid curves), and those employing $\hat{h}_{ICV}$ (dashed curves). Here, for better visual comparison, the regression function estimates produced by using $\hat{h}_{ICV}$ have been vertically shifted

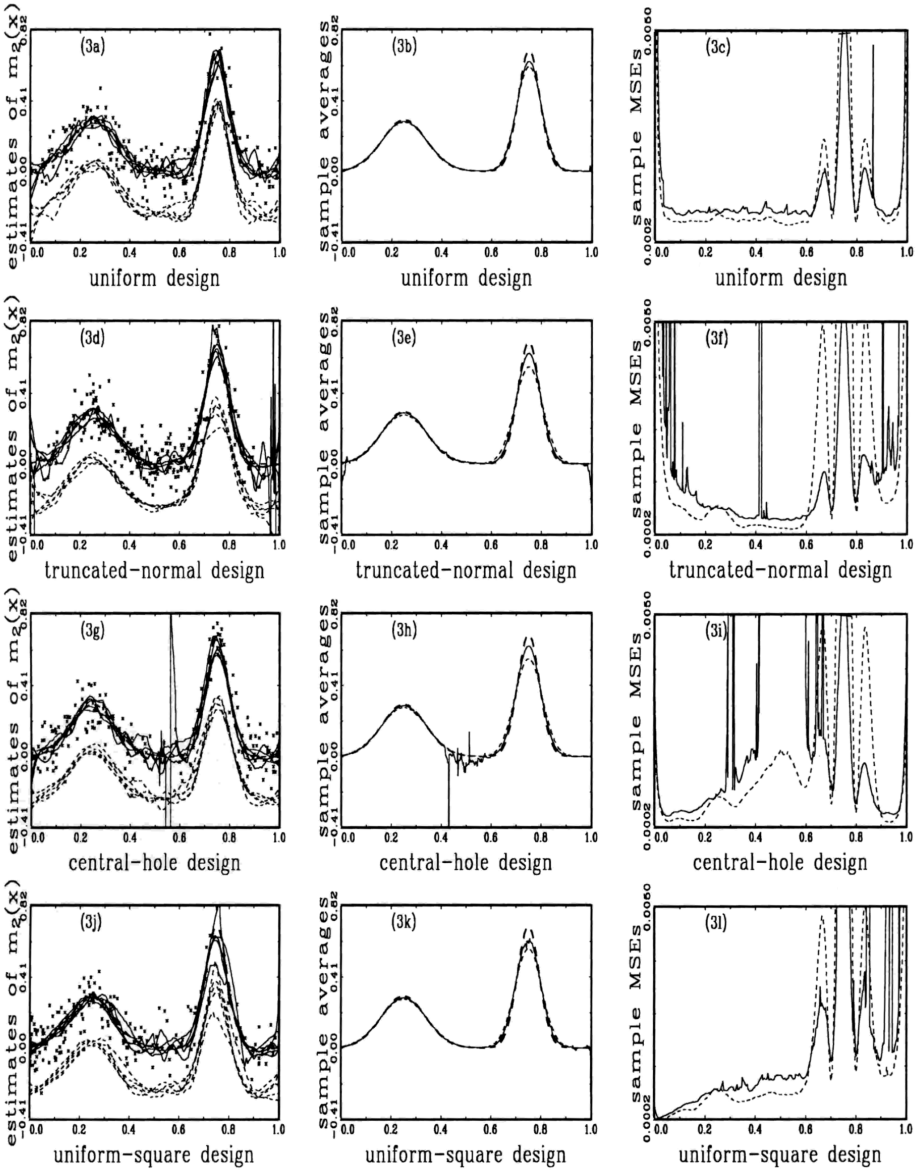

Fig. 3. Practical performance for $m_2(x)$.

below. Figure 2b contains the regression function $m_1(x)$ (bold-faced dashed curve), and the sample average of the 500 regression function estimates derived by $\hat{m}(x)$ using $\hat{h}_{OCV}$ (solid curve) and $\hat{h}_{ICV}$ (dashed curve). Figure 2c shows the sample mean square error (MSE) of the 500 regression function estimates derived by $\hat{m}(x)$ using $\hat{h}_{OCV}$ (solid curve) and $\hat{h}_{ICV}$ (dashed curve). The same descriptions given in (2a)-(2c) for the uniform design apply to (2d)-(2f) for the truncated-normal design,

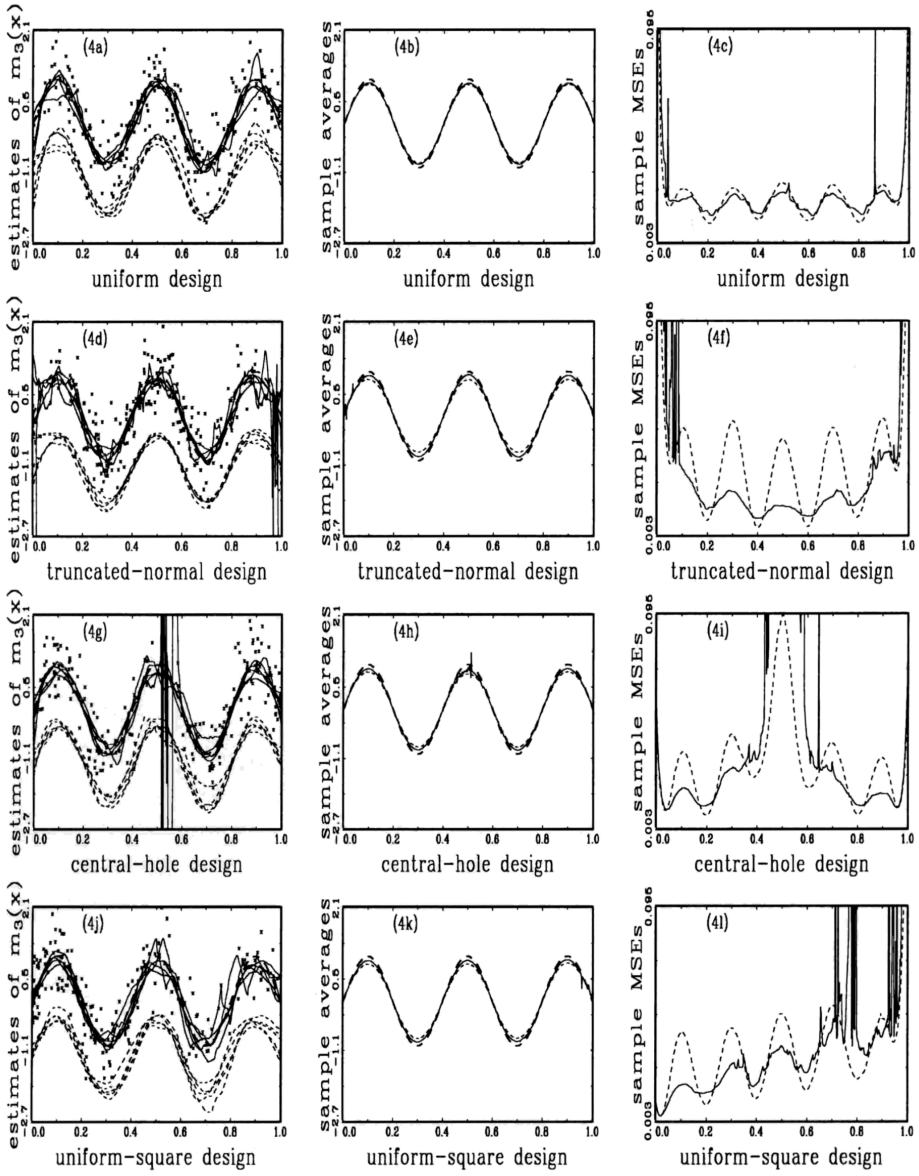

Fig. 4. Practical performance for $m_3(x)$.

to (2g) - (2i) for the central-hole design, and to (2j) - (2l) for the uniform-square design. Similar results for the regression functions $m_2(x)$, $m_3(x)$, and $m_4(x)$ are given respectively in Figures 3-5. These figures all show that the regression function estimate produced by using $\hat{h}_{ICV}$ has smoother appearance, and has smaller sample mean square error at the point outside the regions of peak and trough of the regression

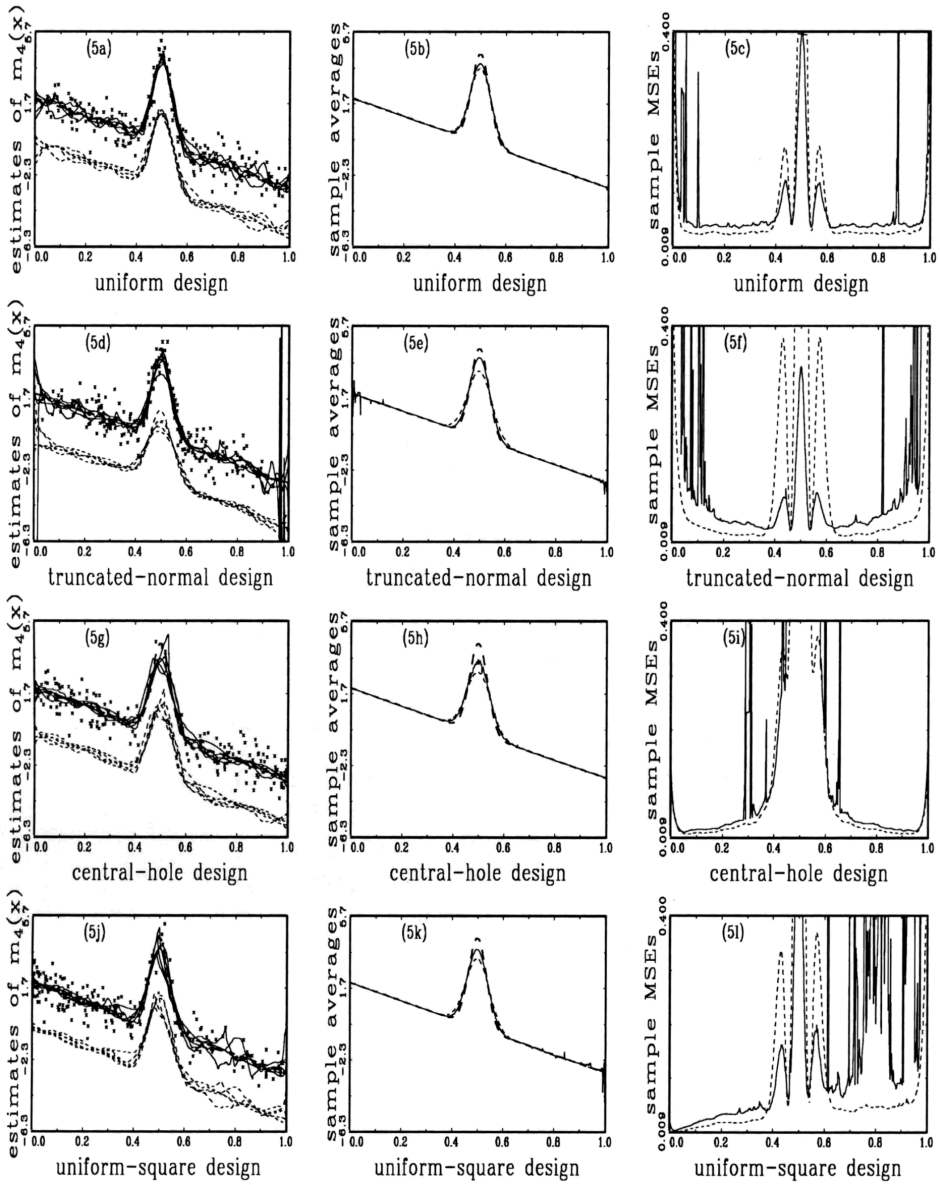

Fig. 5. Practical performance for $m_4(x)$.

function. Under our simulation settings, these results are caused by the fact that $\hat{h}_{ICV}$ is generally of larger magnitude than $\hat{h}_{OCV}$. This fact can be seen clearly from the curves for sample averages, since the curve obtained by using $\hat{h}_{ICV}$ has larger magnitude of sample bias at each point in the regions of peaks and troughs of the true regression function.

Table 1 contains, for each setting, the sample mean and standard deviation of $d_U(\hat{h}_{ICV})$, those of $d_U(\hat{h}_{OCV})$, and those of $d_U(h_U)$. Considering the values of the sample mean and standard deviation, for each setting, the practical performance

Table 1. Values of the sample mean and standard deviation (given in the parentheses) of $d_U(\hat{h}_{OCV})$, those of $d_U(\hat{h}_{ICV})$, and those of $d_U(h_U)$. Each value corresponding to the regression functions $m_1(x)$, $m_2(x)$, $m_3(x)$, and $m_4(x)$ has been multiplied respectively by 10^7, 10^3, 10^2, and 10^1. The positive integer in the upper index denotes the power of ten by which to multiply

	$d_U(\hat{h}_{OCV})$	$d_U(\hat{h}_{ICV})$	$d_U(h_U)$		$d_U(\hat{h}_{OCV})$	$d_U(\hat{h}_{ICV})$	$d_U(h_U)$
	$m_1(x)$ $n=50$				$m_3(x)$ $n=50$		
f_1	$1.02^5(1.62^6)$	18.7(11.4)	13.1(8.40)	f_1	$2.24^2(1.65^3)$	12.8(9.40)	8.64(6.17)
f_2	$8.73^5(3.84^6)$	29.9(33.5)	18.2(12.4)	f_2	$2.31^3(2.58^4)$	24.8(18.6)	14.5(12.1)
f_3	$6.42^5(3.67^6)$	24.2(20.0)	15.1(10.9)	f_3	$1.75^4(3.28^5)$	25.4(16.9)	11.0(6.72)
f_4	$6.36^5(4.25^6)$	27.9(21.6)	18.5(14.9)	f_4	$1.65^3(1.86^4)$	25.2(25.4)	14.5(12.0)
	$m_1(x)$ $n=100$				$m_3(x)$ $n=100$		
f_1	$1.11^4(2.25^5)$	9.48(5.20)	7.04(3.58)	f_1	9.66(61.3)	4.89(2.30)	3.66(1.59)
f_2	$1.10^5(8.14^5)$	13.4(7.40)	8.97(5.46)	f_2	$1.46^2(9.99^2)$	9.43(12.1)	5.62(4.39)
f_3	$2.30^5(1.63^6)$	11.1(7.61)	7.92(4.74)	f_3	$9.53^3(1.41^5)$	7.73(6.49)	5.03(3.01)
f_4	$6.95^4(7.65^5)$	13.5(8.38)	8.94(5.54)	f_4	$3.77^2(3.77^3)$	8.99(8.11)	5.21(2.68)
	$m_1(x)$ $n=200$				$m_3(x)$ $n=200$		
f_1	5.24(2.91)	4.92(2.39)	4.04(1.93)	f_1	2.45(2.91)	2.30(0.93)	1.90(0.75)
f_2	$3.10^3(5.01^4)$	7.00(3.90)	5.06(2.99)	f_2	$31.7(3.65^2)$	3.73(2.14)	2.52(1.26)
f_3	$9.01^3(1.64^5)$	5.93(3.74)	4.51(2.74)	f_3	$32.3(4.19^2)$	3.37(1.90)	2.40(1.16)
f_4	$7.27^3(1.12^5)$	6.57(3.61)	4.92(2.67)	f_4	$14.3(1.95^2)$	3.45(1.77)	2.48(1.06)
	$m_2(x)$ $n=50$				$m_4(x)$ $n=50$		
f_1	$3.89^4(8.55^5)$	9.50(7.63)	5.30(3.21)	f_1	$1.76^4(3.63^5)$	7.14(4.08)	3.32(1.81)
f_2	$8.16^2(4.33^3)$	18.7(57.0)	9.30(6.87)	f_2	$1.48^3(1.04^4)$	8.22(3.64)	4.93(2.42)
f_3	$9.02^3(1.69^5)$	13.9(9.86)	7.21(4.13)	f_3	$3.82^3(6.53^4)$	13.4(2.83)	5.61(3.75)
f_4	$8.89^2(8.34^3)$	21.8(18.2)	9.88(8.00)	f_4	$3.16^3(2.97^4)$	9.42(4.38)	5.11(2.78)
	$m_2(x)$ $n=100$				$m_4(x)$ $n=100$		
f_1	$1.33^2(2.68^3)$	3.35(1.89)	2.33(0.97)	f_1	$1.30^3(2.86^4)$	2.97(3.17)	1.33(0.52)
f_2	$2.52^2(2.19^3)$	6.04(4.30)	3.53(2.35)	f_2	$1.03^3(1.16^4)$	4.29(3.04)	2.18(1.38)
f_3	$1.13^3(1.24^4)$	5.05(3.72)	3.30(1.78)	f_3	$7.10^2(9.99^3)$	8.83(3.48)	2.61(2.44)
f_4	$3.41^2(2.64^3)$	8.42(2.27)	3.60(2.76)	f_4	$5.45^2(5.95^3)$	6.51(3.13)	2.10(1.08)
	$m_2(x)$ $n=200$				$m_4(x)$ $n=200$		
f_1	2.26(19.5)	1.44(0.67)	1.15(0.39)	f_1	1.23(6.58)	0.88(0.73)	0.65(0.21)
f_2	$35.8(2.40^2)$	2.41(1.35)	1.51(0.80)	f_2	$58.7(4.17^2)$	1.49(1.01)	0.89(0.42)
f_3	$6.39^2(1.18^4)$	2.20(0.91)	1.53(0.64)	f_3	9.07(62.8)	2.92(2.50)	1.05(0.71)
f_4	$55.4(1.05^3)$	2.30(1.44)	1.56(0.72)	f_4	$52.5(6.10^2)$	1.52(1.58)	0.89(0.41)

of the regression function estimate produced by $\hat{m}(x)$ using $\hat{h}_{ICV}$ is better than that employing $\hat{h}_{OCV}$. Note that this remark still holds in the special and important case of the uniform design. In this case, both $\hat{h}_{ICV}$ and $\hat{h}_{OCV}$ estimate the same value of h_U.

4.2 Applications

In this subsection, the performance of ICV is illustrated by using three data sets in Simonoff (1996), including the gasoline consumption data, the basketball player data, and the automobile data. The same computation procedures for these three data sets were also applied to Figure 1 in Section 1 for the shampoo data.

Given each data set, the LLE $\hat{m}(x)$ with the Epanechnikov kernel was used to estimate the underlying regression function. The global minimizer $(\hat{h}_{ICV}, \hat{k}_{ICV})$ of $ICV(h,k)$ was chosen on the equally spaced logarithm grid of 500 values of h in the interval $[w/50, w/2]$ and $k = 1, \cdots, [n/5]$. Here w stands for the width of the interval on which the regression function value is estimated. Given the values of h and k, that of $ICV(h,k)$ was approximated by the quantity $(1/u)\sum_{i=1}^{u}\{\hat{m}_k(t_i) - \hat{Y}_k(t_i)\}^2$, where $u = [100w]$ and t_i are equally spaced partition points of the interval on which the regression function value is estimated. The values of $\hat{k}_{ICV}$ derived from the three data sets considered in this section and the shampoo data discussed in Section 1 are 6, 2, 2, and 9, respectively. On the other hand, the ordinary cross-validated bandwidth $\hat{h}_{OCV}$ for $\hat{m}(x)$ was selected on the same grid of h used for choosing $\hat{h}_{ICV}$. The results are given in Figure 6.

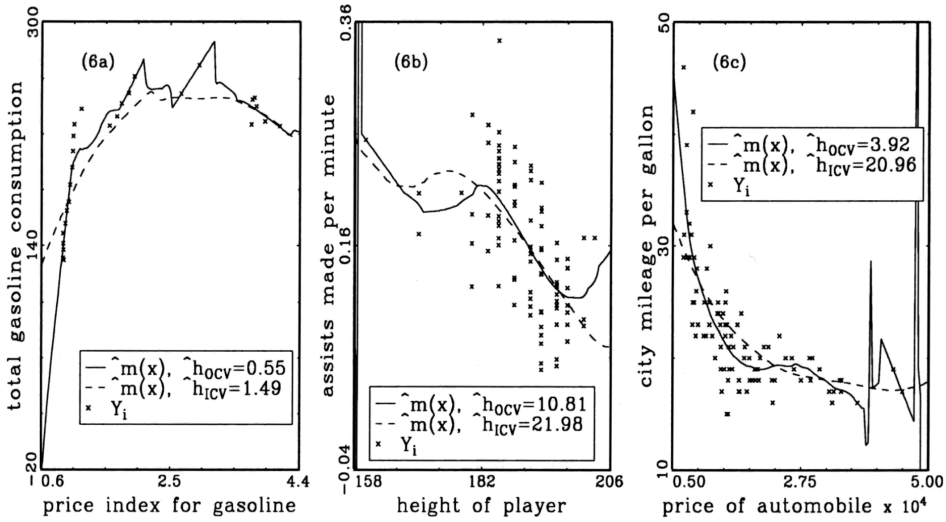

Fig. 6. Plot of the gasoline consumption data of size $n = 27$ (6a), the basketball player data of size $n = 96$ (6b), and the automobile data of size $n = 93$ (6c) and their local linear regression function estimates.

Figure 6 shows that, for each data set, the local linear regression function estimate produced by employing $\hat{h}_{OCV}$ has rough appearance, and exhibits erroneous behavior in some regions. For example, in Figure 6a, this regression function estimate shows unreasonably negative gasoline consumption in the left boundary region. Hence, the relationship between the predictor and the response variables can not be explained correctly by using such kernel estimate. On the other hand, for each data set, the drawback caused by $\hat{h}_{OCV}$ does not happen to $\hat{h}_{ICV}$ since the corresponding local linear regression function has smooth appearance, and has all values in the reasonable range.

5. Sketches of the Proofs

Proof of Theorem 3.1. The following notation will be used throughout this section. For each $x \in [0,1]$, define $v(x;h) = \sum_{i=1}^{n} \omega_i(x)\epsilon_i$, $b(x;h) = \sum_{i=1}^{n} \omega_i(x)\{m(X_i) - m(x)\}$, $\omega_i(x) = n^{-1}h^{-1}\, K(Z_i)\, \{S_2(x) - S_1(x)Z_i\}\, \{S_0(x)\, S_2(x) - S_1^2(x)\}^{-1}$, where $S_j(x)$ and Z_i have been defined in Section 2, and set $v_k(x;h)$ and $b_k(x;h)$ as the "leave-kNN-out" version of $v(x;h)$ and $b(x;h)$, respectively. Hence, $\hat{m}(x) - m(x) = v(x;h) + b(x;h)$ and $\hat{m}_k(x) - m(x) = v_k(x;h) + b_k(x;h)$. Set $v_s(x;h) = v_k(x;h) + v(x;h)$, $v_d(x;h) = v_k(x;h) - v(x;h)$, $b_s(x;h) = b_k(x;h) + b(x;h)$, and $b_d(x;h) = b_k(x;h) - b(x;h)$. Let $A_n = o_u(\alpha_n)$ and $B_n = O_u(\beta_n)$ mean that, as $n \to \infty$, A_n/α_n converges to 0 and $|B_n/\beta_n|$ is bounded above with probability one, and uniformly on [0,1], H_n, or K_n if A_n and B_n involves x, h, or k, respectively, in each case.

To prove Theorem 3.1, following essentially the same proof of Theorem 1 in Härdle and Marron (1985) for the optimality of $\hat{h}_{OCV}$ with respect to the WISE criterion, decompose $ICV(h,k)$ into

$$ICV(h,k) = d_U(h) + d_U\{\varphi_k(x)\} + A(h,k) + B(h,k),$$

where

$$A(h,k) = \int_0^1 \{b_d(x;h) + v_d(x;h)\}\{b_s(x;h) + v_s(x;h)\}dx,$$

$$B(h,k) = (-2)\int_0^1 \{\hat{m}_k(x) - m(x)\}\{\hat{Y}_k(x) - m(x)\}dx.$$

By the decomposition and the fact that the quantity $d_U\{\varphi_k(x)\}$ is independent of h, the proof of our Theorem 3.1 is complete by showing that

$$\sup_{h \in H_n} |\{d_U(h) - d_U^*(h)\}/\, d_U^*(h)| \to 0 \quad \text{with probability one,} \tag{5.1}$$

$$\sup_{h \in H_n, k \in K_n} |A(h,k)/\, d_U^*(h)| \to 0 \quad \text{with probability one,} \tag{5.2}$$

$$\sup_{h\in H_n, k\in K_n} |B(h,k)/\ d_U^*(h)| \to 0 \quad \text{with probability one,} \tag{5.3}$$

as $n \to \infty$. Here $d_U^*(h) = E\{d_U(h)\} = \int_0^1 E\{v^2(x;h)\}dx + \int_0^1 E\{b^2(x;h)\}dx$. Using (2.1), (A1)-(A6), and approximations to the standard errors of functions of random variables in Section 10.5 of Stuart and Ord (1987), through a straightforward calculation, we have

$$\int_0^1 E\{v^2(x;h)\}dx = a_1 n^{-1}h^{-1}\{1 + O(h + n^{-1}h^{-1})\},$$
$$\int_0^1 E\{b^2(x;h)\}dx = b_1 h^4\{1 + O(h + n^{-1}h^{-1})\},$$

where

$$a_1 = \sigma^2\{\int_{-1}^1 K^2(u)du\}\{\int_0^1 f^{-1}(x)dx\},$$
$$b_1 = (1/4)\{\int_{-1}^1 u^2K(u)du\}^2\{\int_0^1 m_2^2(x)dx\},$$

and m_2 denotes the second derivative of m; see Fan and Gijbels (1996).

We now give the proof of (5.1). Using (A1)-(A6), Whittle's inequality in Whittle (1960), and the large deviation theorem in Section 10.3.1 of Serfling (1980), through a straightforward calculation, we have the following asymptotic results: for any h, $h_1 \in H_n$, with $h \le h_1$,

$$v(x;h) = h^{-1}o_u(1), \quad b(x;h) = O_u(h^2),$$
$$\int_0^1 \{\rho_1(x;h) - \rho_1(x;h_1)\}dx = h^{-2}|(h-h_1)/h|o_u(1),$$
$$\int_0^1 \{\rho_2(x;h) - \rho_2(x;h_1)\}dx = h_1|(h-h_1)/h|o_u(1),$$
$$\int_0^1 \{\rho_3(x;h) - \rho_3(x;h_1)\}dx = h_1^3|(h-h_1)/h|o_u(1),$$

where $\rho_1(x;h) = v^2(x;h) - E\{v^2(x;h)\}$, $\rho_2(x;h) = v(x;h)b(x;h)$, and $\rho_3(x;h) = b^2(x;h) - E\{b^2(x;h)\}$. Using these asymptotic results, it is sufficient to restrict the supremum in (5.1) to a set H_n^* which is a subset of H_n so that $\#(H_n^*) \le n^{r+1}$ and so that for any $h \in H_n$ there is an $h_1 \in H_n^*$ with $h \le h_1$ and $|(h-h_1)/h| \le n^{-r}$. Then, for any constant $r \ge 3$, we have

$$\sup_{h\in H_n} |d_U(h) - d_U^*(h)| \le \sup_{h_1\in H_n^*} |d_U(h_1) - d_U^*(h_1)| +$$
$$\sup_{h\in H_n, h_1\in H_n^*, |(h-h_1)/h|\le n^{-r}} |\{d_U(h) - d_U^*(h)\} - \{d_U(h_1) - d_U^*(h_1)\}|$$
$$\le \sup_{h_1\in H_n^*} |d_U(h_1) - d_U^*(h_1)| + o_u(n^{-1}).$$

To prove (5.1), combining this result with the fact that $d_U^*(h) = O(h^4 + n^{-1}h^{-1})$, it is enough to show

$$\sup_{h\in H_n^*} | \int_0^1 \rho_1(x;h)dx \ / \ d_U^*(h)| \to 0 \quad \text{with probability one,} \tag{5.4}$$

$$\sup_{h\in H_n^*} | \int_0^1 \rho_2(x;h)dx \ / \ d_U^*(h)| \to 0 \quad \text{with probability one,} \tag{5.5}$$

$$\sup_{h\in H_n^*} | \int_0^1 \rho_3(x;h)dx \ / \ d_U^*(h)| \to 0 \quad \text{with probability one.} \tag{5.6}$$

To verify (5.4), given $\eta > 0$, for any $t = 1, 2, \cdots$, we have

$$\begin{aligned} &\text{Prob}(\sup_{h\in H_n^*} | \int_0^1 \rho_1(x;h)dx \ / \ d_U^*(h)| > \eta) \\ &\le \eta^{-2t} \#(H_n^*) \sup_{h\in H_n^*} E[\{\int_0^1 \rho_1(x;h)dx \ / \ d_U^*(h)\}^{2t}], \end{aligned}$$

where $\#(H_n^*) = O(n^{r+1})$. The proof of (5.4) is complete when it is seen that there is a constant $\tau > 0$, so that for $t = 1, 2, \cdots$, there are constants c_t so that

$$\sup_{h\in H_n^*} E[\{\int_0^1 \rho_1(x;h)dx \ / \ d_U^*(h)\}^{2t}] \le c_t n^{-\tau t}. \tag{5.7}$$

Using (5.7) and the Borel-Cantelli lemma, there is a sufficiently large t to make $r + 1 - \tau t < -1$, then, for any given $\eta > 0$,

$$\sum_{n=1}^{\infty} \text{Prob}(\sup_{h\in H_n^*} | \int_0^1 \rho_1(x;h)dx \ / \ d_U^*(h)| > \eta) \le c_t \sum_{n=1}^{\infty} n^{r+1-\tau t} < \infty.$$

Hence the proof of (5.4) is complete.

To prove (5.7), $\rho_1(x;h)$ is expressed as $\varphi_1(x;h) + \varphi_2(x;h)$, where

$$\varphi_1(x;h) = \sum_{i=1, i\neq j,}^{n} \sum_{j=1}^{n} \omega_i(x)\omega_j(x)\epsilon_i\epsilon_j, \quad \varphi_2(x;h) = \sum_{i=1}^{n} [\omega_i^2(x)\epsilon_i^2 - E\{\omega_i^2(x)\}\sigma^2].$$

Using (A1)-(A6), Whittle's inequality, and approximations to the standard errors of functions of random variables, through a straightforward calculation, we have

$$E[\{\int_0^1 \varphi_1(x;h)dx\}^{2t}] = O(n^{-2t}h^{-t}), \quad E[\{\int_0^1 \varphi_2(x;h)dx\}^{2t}] = O(n^{-3t}h^{-2t}).$$

Hence, the proof of (5.7) is complete.

The proofs for (5.5)-(5.6) and those for (5.2)-(5.3) are essentially the same as that of (5.4) and that of (5.1), respectively. Hence they are omitted. The proof of Theorem 3.1 is complete.

Acknowledgements

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. G. M. Bayhan, and M. Bayhan, Forecasting using autocorrelated errors and multicollinear predictor variables. *Computers and Industrial Engineering*, **34** (1998), 413-421.
2. C. K. Chu, and J. S. Marron, Comparison of two bandwidth selectors with dependent errors. *Annals of Statistics*, **19** (1991a), 1906-1918.
3. C. K. Chu, and J. S. Marron, Choosing a kernel regression estimator. *Statistical Science*, **6** (1991b), 404-436.
4. R. M. Clark, A calibration curve for radiocarbon data. *Antiquity*, **49** (1975), 251-266.
5. W. S. Deng, C. K. Chu, and M. Y. Cheng, A study of local linear ridge estimators. *Journal of Statistical Planning and Inference*, **93** (2001), 225-238.
6. V. A. Epanechnikov, Nonparametric estimation of a multivariate probability density. *Theory of Probability and Its Applications*, **14** (1969), 153-158.
7. R. L. Eubank, *Spline Smoothing and Nonparametric Regression*. Marcel Dekker Inc., New York, (1988).
8. J. Fan, Design-adaptive nonparametric regression. *Journal of the American Statistical Association*, **87** (1992), 998-1004.
9. Fan, J. Local linear regression smoothers and their minimax efficiencies. *Annals of Statistics*, **21** (1993), 196-216.
10. J. Fan, T. Gasser, I. Gijbels, M. Brookmann, and J. Engel, *Local polynomial fitting: A standard for nonparametric regression*. Discussion paper 9315. Institut de Statistique, Universite Catholique de Louvain, Belgium, (1993).
11. J. Fan, and I. Gijbels, Variable bandwidth and local linear regression smoothers. *Annals of Statistics*, **20** (1992), 2008-2036.
12. J. Fan, and I. Gijbels, Data-driven bandwidth selection in local polynomial fitting: variable bandwidth and spatial adaptation. *Journal of the Royal Statistical Society, Ser. B*, **57** (1995), 371-394.

13. J. Fan, and I. Gijbels, *Local Polynomial Modeling and Its Application – Theory and Methodologies*. New York: Chapman and Hall, (1996).

14. T. Gasser, and J. Engel, The choice of weights in kernel regression estimation. *Biometrika*, **77** (1990), 377-381.

15. W. Härdle, *Applied Nonparametric Regression*. Cambridge University Press, (1990).

16. W. Härdle, Smoothing Techniques: With Implementation in S. Springer Series in Statistics, Springer-Verlag, Berlin, (1991).

17. W. Härdle, P. Hall, and J. S. Marron, How far are automatically chosen regression smoothing parameters from their optimum? *Journal of the American Statistical Association*, **83** (1988), 86-101.

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

19. P. Hall, and B. A. Turlach, Interpolation methods for adapting to sparse design in nonparametric regression (with discussion). *Journal of the American Statistical Association*, **92** (1997), 466-476.

20. R. C. Hwang, Asymptotic properties of locally weighted regression. *Journal of Nonparametric Statistics*, **5** (1995), 303-310.

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

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

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

24. J. Rice, Bandwidth choice for nonparametric regression. *Annals of Statistics*, **12** (1984), 1215-1230.

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

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

27. B. Seifert, and T. Gasser, Finite-sample analysis of local polynomials: Analysis and solutions. *Journal of the American Statistical Association*, **91** (1996), 267-275.

28. R. Serfling, *Approximation Theorems of Mathematical Statistics*. Wiley, New York, (1980).

29. R. Shibata, An optimal selection of regression variables. *Biometrika*, **68** (1981), 45-54.

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

31. J. S. Simonoff, *Smoothing Methods in Statistics*. New York: Springer, (1996).

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

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

34. P. Whittle, Bounds for the moments of linear and quadratic forms in independent variables. *Theory of Probability and Its Applications*, **5** (1960), 302-305.

35. J. S. Wu, and C. K. Chu, Double smoothing for kernel estimators in nonparametric regression. *Journal of Nonparametric Statistics*, **1** (1992), 375-386.

Tzu-Kuei Chang
Department of Mathematics Education,
National Hualien Teachers College,
Hualien, Taiwan 970, R.O.C.
E-mail: tzu@sparc2.nhltc.edu.tw

Wen-Shuenn Deng
Department of Statistics,
Tamkang University,
Tamsui, Taiwan 251, R.O.C.

Jung-Huei Lin
General Education Center,
Taiwan Hospitality and Tourism College,
Hualien, Taiwan 974, R.O.C.

C. K. Chu
Department of Applied Mathematics,
National Donghua University,
Hualien, Taiwan 974, R.O.C.