Journal of Statistical Planning and Inference 93 (2001) 225–238

journal of statistical planning and inference

www.elsevier.com/locate/jspi

A study of local linear ridge regression estimators

Wen-Shuenn Deng[a], Chih-Kang Chu[a,*], Ming-Yen Cheng[b]

[a] *Department of Applied Mathematics, National Dong Hua University, Taiwan 974, Republic of China*

[b] *Department of Mathematics, National Taiwan University, Taiwan 106, Republic of China*

Received 8 July 1999; received in revised form 3 February 2000; accepted 16 June 2000

Abstract

In the case of the random design nonparametric regression, to correct for the unbounded finite-sample variance of the local linear estimator (LLE), Seifert and Gasser (J. Amer. Statist. Assoc. 91 (1996) 267–275) apply the idea of ridge regression to the LLE, and propose the local linear ridge regression estimator (LLRRE). However, the finite sample and the asymptotic properties of the LLRRE are not discussed there. In this paper, upper bounds of the finite-sample variance and bias of the LLRRE are obtained. It is shown that if the ridge regression parameters are not properly selected, then the resulting LLRRE has some drawbacks. For example, it may have a nonzero constant asymptotic bias, may suffer from boundary effects, or may be unable to share the nice asymptotic bias quality of the LLE. On the other hand, if the ridge regression parameters are properly selected, then the resulting LLRRE does not suffer from the above problems, and has the same asymptotic mean-square error as the LLE. For this purpose, the ridge regression parameters are allowed to depend on the sample size, and converge to 0 as the sample size increases. In practice, to select both the bandwidth and the ridge regression parameters, the idea of cross-validation is applied. Simulation studies demonstrate that the LLRRE using the cross-validated bandwidth and ridge regression parameters could have smaller sample mean integrated square error than the LLE using the cross-validated bandwidth, in reasonable sample sizes. © 2001 Elsevier Science B.V. All rights reserved.

MSC: 62J05; 62J07

Keywords: Asymptotic behavior; Boundary effect; Finite-sample behavior; Local linear ridge regression estimator; Local linear estimator; nonparametric regression; Ridge regression

1. Introduction

In the field of kernel regression estimation, it is well known that the local linear estimator (LLE) has many advantages. For example, it achieves full asymptotic minimax efficiency among all linear estimators (Fan, 1993), has a nice asymptotic bias

[*] Corresponding author.
E-mail address: chu@server.am.ndhu.edu.tw (C.-K. Chu).

0378-3758/01/$ - see front matter © 2001 Elsevier Science B.V. All rights reserved.
PII: S0378-3758(00)00161-0

quality and a superior asymptotic variance quantity (Wu and Chu, 1992), and adapts automatically to the boundary (Fan and Gijbels, 1992). For a detailed discussion of the LLE and other kernel regression function estimators, see, for example, the monographs by Eubank (1988), Müller (1988), Haȑdle (1990, 1991), Wand and Jones (1995), Fan and Gijbels (1996), and Simonoff (1996).

However, Seifert and Gasser (1996) show that there is a serious drawback to the LLE. The drawback is that the LLE has unbounded finite-sample conditional variance when a kernel function with compact support is used. Compactly supported kernels are often employed for computational convenience or for optimal performance (e.g. the Epanechnikov kernel minimizes mean square error among all nonnegative kernels; see Epanechnikov, 1969). In that case, the regression function estimate produced by the LLE sometimes has rough appearance. This adverse effect to the LLE is not shared by other popular kernel regression function estimators, for example, the Nadaraya–Watson estimator (Nadaraya, 1964; Watson, 1964) and the Gasser–Müller estimator (Gasser and Müller, 1979, 1984). The upper bounds of the finite-sample conditional variances of these two kernel regression function estimators are given in Section 2.

To correct for the above adverse effect to the LLE, Seifert and Gasser (1996) apply the idea of ridge regression to the LLE, and propose the local linear ridge regression estimator (LLRRE). But, theoretical properties of the LLRRE are not given there. The purpose of this article is to study the finite sample and the asymptotic behaviors of the LLRRE. For other approaches improving the adverse effect to the LLE, see, for example, Fan (1993) where a small positive quantity is added to the denominator of the LLE, and Hall and Marron (1997) which suggest shrinking the LLE towards another estimator with bounded mean-square error.

It is shown in Section 3 that if a kernel function with compact support is used, then the LLRRE has bounded finite-sample conditional (and unconditional) variance and bias. For the asymptotic properties of the LLRRE, it is also shown in Section 3 that if the ridge regression parameters are not properly selected, then the resulting LLRRE has some drawbacks. For example, it may have a nonzero constant asymptotic bias, may suffer from boundary effects, or may be unable to share the nice asymptotic bias quality of the LLE. On the other hand, if the ridge regression parameters are properly selected, then the resulting LLRRE does not suffer from the above problems, and has the same asymptotic mean-square error (AMSE) as the LLE. For this purpose, the ridge regression parameters are allowed to depend on the sample size, and converge to 0 as the sample size increases. In practice, to select both the bandwidth and the ridge regression parameters, we suggest using the idea of cross-validation. Simulation studies contained in Section 4 demonstrate that the LLRRE using the cross-validated bandwidth and ridge regression parameters could have smaller sample mean integrated square error (MISE) than the LLE using the cross-validated bandwidth, in reasonable sizes.

This article is organized as follows. A precise formulation of the LLRRE is described in Section 2. The finite sample and the asymptotic behaviors of the LLRRE are contained in Section 3. Simulation studies to gain additional insight to the theoretical results achieved in Section 3 are presented in Section 4. Finally, sketches of the proofs are given in the appendix.

2. Regression settings and estimators

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

$$Y_j = m(X_j) + \varepsilon_j, \tag{1}$$

for $j=1,\dots,n$. Here (X_j, Y_j) are independent and identically distributed bivariate random vectors, and ε_j are assumed to have mean 0 and variance σ^2, $0<\sigma^2<\infty$. The design points X_j are assumed to be independent of the regression errors ε_j, and are assumed to have the probability density function $f(x)$ supported on the bounded interval $[0,1]$. The purpose of the regression is to use the data points (X_j, Y_j) to estimate the regression function m.

The rest of this section is devoted to giving the formulation of the LLE and that of the LLRRE. For simplicity of presentation, assume that the regression function m has two continuous derivatives. Given the kernel function K as a probability density function supported on the interval $[-1,1]$ and the bandwidth $h=h_n$ tending to 0 as $n\to\infty$, the LLE $\tilde{m}(x)$ for $m(x)$ is constructed by minimizing the local weighted linear least squares

$$n^{-1}h^{-1}\sum_{j=1}^{n}(Y_j-\beta_0-\beta_1 Z_j)^2 K(Z_j) \tag{2}$$

for $x\in[0,1]$, where $Z_j=(x-X_j)/h$.

Through a straightforward calculation, the minimization problem (2) is equivalent to the problem of solving a system of linear equations

$$S\beta = T, \tag{3}$$

where

$$S=\begin{bmatrix} S_0 & S_1 \\ S_1 & S_2 \end{bmatrix},\quad \beta=\begin{bmatrix}\beta_0\\ \beta_1\end{bmatrix},\quad T=\begin{bmatrix}T_0\\ T_1\end{bmatrix}.$$

Here

$$S_k = n^{-1}h^{-1}\sum_{j=1}^{n} Z_j^k K(Z_j),$$

$$T_k = n^{-1}h^{-1}\sum_{j=1}^{n} Y_j Z_j^k K(Z_j)$$

for $k\geqslant 0$. Let $\tilde{\beta}_0$ and $\tilde{\beta}_1$ denote the solution of β_0 and β_1 in (3), respectively. Through a straightforward calculation, $\tilde{\beta}_0$ can be expressed as

$$\tilde{\beta}_0 = (T_0S_2 - T_1S_1)/(S_0S_2 - S_1^2).$$

By (2) and the first-order Taylor theorem, take $\tilde{m}(x)=\tilde{\beta}_0$. If the denominator of $\tilde{m}(x)$ is 0, then take $\tilde{m}(x)=0$.

By Cauchy–Schwartz inequality, the denominator $S_0S_2-S_1^2$ of $\tilde{m}(x)$ is nonnegative. In practice, it is possible that the denominator of $\tilde{m}(x)$ is 0. This case occurs when there is "no" or "one" design point falling in the compact window $[x-h,\ x+h]$ around x. The more sparse the distribution of the design points, the more often this case occurs. When it happens, the value of $\tilde{m}(x)$ is given as 0. However, such assignment might cause that $\tilde{m}(x)$ exhibits erratic behavior. The same drawback also happens to other kernel regression function estimators, for example, the Nadaraya–Watson estimator and the LLRRE. For the latter two estimators, such drawback occurs only when there is "no" design point falling in the compact window around x.

Even when the denominator of $\tilde{m}(x)$ is not 0, it might be nearly 0, and the resulting $\tilde{m}(x)$ suffers from the numerical instability problems. This case occurs when, for example, the design points falling in the compact window $[x-h,\ x+h]$ around x are all close to one another. In this special case, the closer these design points to one another, the larger the finite-sample conditional variance of the resulting $\tilde{m}(x)$. Hence, the finite-sample conditional variance of the LLE may become arbitrarily large. However, this adverse behavior of $\tilde{m}(x)$ is not present in the Nadaraya–Watson and the Gasser–Müller estimators, where for any kernel function the finite sample conditional variance is bounded by σ^2. For a detailed discussion of these facts, see Seifert and Gasser (1996).

To correct for the unbounded finite-sample conditional variance of the LLE, Seifert and Gasser (1996) use the idea of ridge regression to solve (3) and the resulting estimate of $m(x)$ is called the LLRRE. For a detailed discussion of ridge regression, see, for example, Montgomery and Peck (1982). Given the ridge regression parameters $\lambda_0\geqslant 0$ and $\lambda_1>0$, take the estimate of β by solving

$$(S+\Lambda)\beta=T, \tag{4}$$

where

$$\Lambda=\begin{bmatrix}\lambda_0 & 0\\ 0 & \lambda_1\end{bmatrix}.$$

Let $\hat{\beta}_0$ and $\hat{\beta}_1$ denote the solution of β_0 and β_1 in (4), respectively. Through a straightforward calculation, $\hat{\beta}_0$ can be expressed as

$$\hat{\beta}_0=\{T_0(\lambda_1+S_2)-T_1S_1\}/\{(\lambda_0+S_0)(\lambda_1+S_2)-S_1^2\}.$$

The LLRRE $\hat{m}(x)$ for $m(x)$ is taken as $\hat{m}(x)=\hat{\beta}_0$. If there is at least one design point falling in the compact window $[x-h,\ x+h]$ around x, then $S_0>0$ and, by Cauchy–Schwartz inequality, $S_0S_2-S_1^2\geqslant 0$. In this case, using the fact that $\lambda_0\geqslant 0$ and $\lambda_1>0$, the denominator $(\lambda_0+S_0)(\lambda_1+S_2)-S_1^2$ of $\hat{m}(x)$ is positive and $\hat{m}(x)$ is well defined. If the denominator of $\hat{m}(x)$ is 0, take $\hat{m}(x)=0$. If $\lambda_0=\lambda_1=0$, then $\hat{m}(x)=\tilde{m}(x)$.

To compute the LLRRE $\hat{m}(x)$, Seifert and Gasser (1996) suggest taking $\lambda_0=0$. Under this circumstance, if there is only one design point falling in the compact window around x, then $S_0S_2 - S_1^2 = 0$, $T_0S_2 - T_1S_1 = 0$, and the value of such $\hat{m}(x)$ is equal to that of T_0/S_0. Note that T_0/S_0 is the Nadaraya–Watson estimator for $m(x)$. Similarly, if the design points falling in the compact window around x are all close to one another, then both $S_0S_2 - S_1^2$ and $T_0S_2 - T_1S_1$ are roughly equal to 0, and the value of such $\hat{m}(x)$ approaches that of T_0/S_0. From these facts, it is expected that, like the Nadaraya–Watson estimator, the LLRRE has bounded finite-sample conditional variance.

The finite sample and the asymptotic behaviors of $\hat{m}(x)$ will be studied in Section 3.

3. Results

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

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

(A2) The design density f is Lipschitz continuous and positive on the interval $[0,1]$.

(A3) The kernel function K is a Lipschitz continuous and symmetric probability density function with support $[-1,1]$.

(A4) The value of h is selected on the interval $H_n = [\rho n^{-1+\delta}, \rho^{-1}n^{-\delta}]$, where the positive constants ρ and δ are arbitrarily small.

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

The following Theorem 3.1 gives upper bounds of the finite sample conditional variance and bias of $\hat{m}(x)$, and Theorem 3.2 shows the asymptotic variance and bias of $\hat{m}(x)$. Their proofs are given in the appendix. To state these theorems, we introduce the following notations. For $j \geqslant 1$, let f_j and m_j denote the jth derivatives of f and m, and m_j^0 and m^0 the maximum absolute value of $m_j(x)$ and $m(x)$ over $x \in [0,1]$, respectively, in each case. Let K^0 denote the maximum value of K over $[-1,1]$. Set $X^* = (X_1,\ldots,X_n)$, $\kappa_j = \int_{(x-1)/h}^{x/h} z^j K(z)\,\mathrm{d}z$, for $j \geqslant 0$, and

$$c_0 = \{\lambda_0 + f(x)\kappa_0\}\{\lambda_1 + f(x)\kappa_2\} - \{f(x)\kappa_1\}^2,$$

$$c_1 = f(x)f_1(x)(\kappa_1\kappa_2 - \kappa_0\kappa_3) - \lambda_1 f_1(x)\kappa_1 - \lambda_0 f_1(x)\kappa_3,$$

$$c_2 = (1/2)\{f(x)f_2(x)(\kappa_0\kappa_4 - 2\kappa_1\kappa_3 + \kappa_2^2) + 2f_1^2(x)(\kappa_1\kappa_3 - \kappa_2^2) + \lambda_0 f_2(x)\kappa_4 + \lambda_1 f_2(x)\kappa_2\},$$

$$c_3 = (-1)\lambda_0 m(x)\{\lambda_1 + f(x)\kappa_2\},$$

$$c_4 = \lambda_0 m(x)f_1(x)\kappa_3 - \lambda_1 m_1(x)f(x)\kappa_1,$$

$$c_5 = (1/2)[m_2(x)f(x)^2(\kappa_2^2 - \kappa_1\kappa_3) - \lambda_0 m(x)f_2(x)\kappa_4 + \lambda_1\kappa_2\{m_2(x)f(x) + 2m_1(x)f_1(x)\}],$$

$$c_6 = m(x)\{\lambda_1 f(x)\kappa_0 + f(x)^2\kappa_0\kappa_2 - f(x)^2\kappa_1^2\}.$$

Theorem 3.1. *Suppose that the assumptions given in Section* 2 *and* (A1)–(A4) *hold. Given the sample size* n, *if* $\lambda_0 \geqslant 0$ *and* $\lambda_1 > 0$, *then the finite-sample conditional variance and bias of* $\hat{m}(x)$ *are bounded above*, *respectively*, *by*

$$\mathrm{Var}\{\hat{m}(x)|X^*\} \leqslant \sigma^2(1 + 2\lambda_1^{-1}S_0)^2 \leqslant \sigma^2(1 + 2\lambda_1^{-1}h^{-1}K^0)^2, \tag{5}$$

$$\begin{aligned} |\mathrm{Bias}\{\hat{m}(x)|X^*\}| &\leqslant \begin{bmatrix} hm_1^0 + h^2m_2^0(1/2 + \lambda_1^{-1}S_0) & \textit{for } \lambda_0 = 0, \\ m^0 + hm_1^0 + h^2m_2^0(1/2 + \lambda_1^{-1}S_0) & \textit{for } \lambda_0 > 0, \end{bmatrix} \\ &\leqslant \begin{bmatrix} hm_1^0 + h^2m_2^0(1/2 + \lambda_1^{-1}h^{-1}K^0) & \textit{for } \lambda_0 = 0, \\ m^0 + hm_1^0 + h^2m_2^0(1/2 + \lambda_1^{-1}h^{-1}K^0) & \textit{for } \lambda_0 > 0 \end{bmatrix} \end{aligned} \tag{6}$$

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

Theorem 3.2. *Suppose that the assumptions given in Section* 2 *and* (A1)–(A5) *hold. If* $\lambda_0 \geqslant 0$ *and* $\lambda_1 > 0$, *then the asymptotic variance and bias of* $\hat{m}(x)$ *can be expressed*, *respectively*, *as*

$$\mathrm{Var}\{\hat{m}(x)\} = n^{-1}h^{-1}(v_1 + v_2) + \mathrm{o}(n^{-1}h^{-1}), \tag{7}$$

$$\mathrm{Bias}\{\hat{m}(x)\} = b_0 + b_1h + b_2h^2 + \mathrm{o}(h^2) + \mathrm{O}(n^{-1}h^{-1}) \tag{8}$$

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

$$v_1 = c_0^{-2}\sigma^2 f(x)\int_{(x-1)/h}^{x/h} K(z)^2\{\lambda_1 + f(x)\kappa_2 - zf(x)\kappa_1\}^2\,\mathrm{d}z,$$

$$\begin{aligned} v_2 = c_0^{-2}f(x)\int_{(x-1)/h}^{x/h} & K(z)^2[\{f(x)(\kappa_2 - 2z\kappa_1 + z^2\kappa_0) + \lambda_1\}\{m(x) - c_0^{-1}c_6\} \\ & - z^2\lambda_0c_0^{-1}c_6]\,\mathrm{d}z, \end{aligned}$$

$$b_0 = c_0^{-1}c_3,$$

$$b_1 = c_0^{-2}(c_0c_4 - c_1c_3),$$

$$b_2 = c_0^{-3}(c_0^2c_5 - c_0c_1c_4 + c_1^2c_3 - c_0c_2c_3).$$

If $\lambda_0 = \lambda_{0,n} = \mathrm{o}(h^2) \geqslant 0$ and $\lambda_1 = \lambda_{1,n} = \mathrm{o}(h) > 0$, then the asymptotic variance and bias of $\hat{m}(x)$ can be expressed, respectively, as

$$\mathrm{Var}\{\hat{m}(x)\} = n^{-1}h^{-1}v_3 + \mathrm{o}(n^{-1}h^{-1}), \tag{9}$$

$$\mathrm{Bias}\{\hat{m}(x)\} = h^2b_3 + \mathrm{o}(h^2) + \mathrm{O}(n^{-1}h^{-1}) \tag{10}$$

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

$$v_3 = f(x)^{-1}(\kappa_0\kappa_2 - \kappa_1^2)^{-2}\sigma^2\int_{(x-1)/h}^{x/h} K(z)^2(\kappa_2 - z\kappa_1)^2\,\mathrm{d}z,$$

$$b_3 = (1/2)(\kappa_0\kappa_2 - \kappa_1^2)^{-1}(\kappa_2\kappa_2 - \kappa_1\kappa_3)m_2(x).$$

We now close this section by the following remarks.

Remark 3.1 (*Upper bounds for the finite sample unconditional variance and bias of $\hat{m}(x)$*)**.** Since the upper bounds given on the right-hand side of (5) and (6) are all finite constants, $\hat{m}(x)$ has bounded finite sample unconditional variance and bias as well.

Remark 3.2 (*The asymptotic behavior of $\hat{m}(x)$ when $\lambda_0 = \mathrm{o}(h^2) \geqslant 0$ and $\lambda_1 = \mathrm{o}(h) > 0$*)**.** Note that the asymptotic variance and bias of such $\hat{m}(x)$ given, respectively, in (9) and (10) are the same as those of $\tilde{m}(x)$, for each $x \in [0,1]$. For the latters, see Theorem 1 of Fan (1993) and Theorem 4 of Fan and Gijbels (1992). On the other hand, for constructing $\hat{m}(x)$, Seifert and Gasser (1996) suggest taking λ_0 as 0 and λ_1 as a positive constant. Note that S&G use the quadratic forms $S_k = \sum_{j=1}^n (x - X_j)^k K\{(x - X_j)/h\}$, whereas we use $S_k = n^{-1}h^{-1} \sum_{j=1}^n \{(x - X_j)/h\}^k K\{(x - X_j)/h\}$. Thus, using our notation of S_k, the value of S&G's λ_1 becomes $\alpha n^{-1}h^{-3}$, for some $\alpha > 0$. By (9) and (10), if $n^{-1}h^{-3}$ is of smaller order than h, then $\hat{m}(x)$ suggested by S&G has the same AMSE as $\tilde{m}(x)$, for each $x \in [0,1]$. By Theorem 3.2, the value of the optimal h for constructing $\hat{m}(x)$, from the viewpoint of minimizing AMSE, is of order $n^{-1/5}$. In that case, the value of S&G's $\lambda_1 = \alpha n^{-1}h^{-3}$ is of order $n^{-2/5}$, and is of smaller order than that of the optimal h. For more discussion of the asymptotic performance of $\hat{m}(x)$ with $\lambda_0 = 0$, see Remarks 3.6 and 3.7.

Remark 3.3 (*The formulation of Hall and Marron's shrinkage estimator $\hat{m}_S(x)$*)**.** Since the asymptotic properties of $\hat{m}(x)$ discussed in the following remarks are related to those of $\hat{m}_S(x)$, the formulation of $\hat{m}_S(x)$ is now introduced. To overcome the problem that the denominator of $\tilde{m}(x)$ might be close to 0, one may shrink $\tilde{m}(x)$ by an amount ε in the direction of another estimator, $\tilde{m}_0(x)$ say, whose properties are less erratic than those of $\tilde{m}(x)$. That is, we choose β_0 and β_1 to minimize $n^{-1}h^{-1} \sum_{j=1}^n (Y_j - \beta_0 - \beta_1 Z_j)^2 K(Z_j) + \{\tilde{m}_0(x) - \beta_0\}^2 \varepsilon$, and take the shrinkage estimator $\hat{m}_S(x) = \beta_0$. It may be shown that

$$\hat{m}_S(x) = \hat{m}_{\mathrm{RLE}}(x) + \eta(S_0S_2 - S_1^2 + \eta)^{-1}\tilde{m}_0(x),$$

where $\hat{m}_{\mathrm{RLE}}(x) = (T_0S_2 - T_1S_1)/(S_0S_2 - S_1^2 + \eta)$ is the ridged linear estimator, and $\eta = \varepsilon S_2$. Note that H&M use the quadratic forms $S_k = \sum_{j=1}^n (x - X_j)^k K\{(x - X_j)/h\}$, whereas we use $S_k = n^{-1}h^{-1} \sum_{j=1}^n \{(x - X_j)/h\}^k K\{(x - X_j)/h\}$. Using our quadratic forms S_k and Theorems 2.1 and 2.2 and Remark 2.3 of H&M, if $\varepsilon = \mathrm{o}(h^2)$ and (A1)–(A5) hold, then both $\hat{m}_S(x)$ and $\hat{m}_{\mathrm{RLE}}(x)$ have the same AMSE as $\tilde{m}(x)$, for each $x \in [0,1]$. More results for the asymptotic performance of $\hat{m}_S(x)$ can be found in Remarks 3.4–3.7.

Remark 3.4 (*The asymptotic behavior of $\hat{m}(x)$ when λ_0 is a positive constant*)**.** If λ_0 is a positive constant and $m(x)$ is not equal to 0, then $b_0 \neq 0$. Under this circumstance, by (8), such $\hat{m}(x)$ has a nonzero constant asymptotic bias. Hence, it is not suggested taking λ_0 as a positive constant when computing $\hat{m}(x)$. Seifert and Gasser (1996) suggest taking $\lambda_0 = 0$ without providing any reason. Our result gives a nice explanation for that.

Remark 3.5 (*The finite sample and the asymptotic behaviors of $\hat{m}(x)$ when $\lambda_0 = \lambda_{0,n} > 0$ and $\lambda_1 = 0$*). To correct for the nonzero constant asymptotic bias of $\hat{m}(x)$ with λ_0 a positive constant in Remark 3.4, take $\lambda_0 = \lambda_{0,n} > 0$ and $\lambda_1 = 0$. Such $\hat{m}(x)$ is H&M's ridged linear estimator $\hat{m}_{\mathrm{RLE}}(x)$ produced by using $\varepsilon = \lambda_0$ and $\tilde{m}_0(x) \equiv 0$, for each $x \in [0,1]$. The same conclusion can be obtained for $\hat{m}(x)$ in Remark 3.4. By Remark 3.3, if $\lambda_0 = \mathrm{o}(h^2)$, then such $\hat{m}(x)$ has the same AMSE as $\tilde{m}(x)$, for each $x \in [0,1]$. This result agrees with our (9) and (10). However, there is a drawback to such $\hat{m}(x)$ in the finite sample case. If there is only one design point, X_1 say, falling in the compact window $[x-h, x+h]$ around x, and $x = X_1$, then $S_1 = S_2 = 0$, and for $\lambda_1 = 0$, the resulting $\hat{m}(x)$ has denominator zero, and is not well defined. The same drawback also happens to $\hat{m}_{\mathrm{RLE}}(x)$.

Remark 3.6 (*The asymptotic performance of $\hat{m}(x)$ when λ_0 is equal to 0 and λ_1 is a positive constant*). In this situation, if $x \in [h, 1-h]$, then $b_0 = b_1 = 0$ and b_2 becomes b_2^*, where

$$b_2^* = (1/2) m_2(x)\kappa_2 + f(x)^{-1}\{\lambda_1 + f(x)\kappa_2\}^{-1}\lambda_1 m_1(x) f_1(x)\kappa_2.$$

On the other hand, if $x \in [0,h) \cup (1-h, 1]$ and $m_1(x) \neq 0$, then $b_0 = 0$ and $b_1 \neq 0$. By these results, the asymptotic bias of such $\hat{m}(x)$ given in (8) depends on several factors f, f_1, m_1, and m_2, but that of $\tilde{m}(x)$ depends only on m_2. Also, the magnitudes of the asymptotic biases of such $\hat{m}(x)$ are of order h^2 and h for $x \in [h, 1-h]$ and for $x \in [0,h) \cup (1-h, 1]$, respectively. Hence, such $\hat{m}(x)$ has poor asymptotic bias quality, and suffers from the problem of boundary effects.

Remark 3.7 (*The asymptotic performance of $\hat{m}(x)$ when $\lambda_0 = 0$ and $\lambda_1 = \lambda_{1,n} > 0$*). To improve the asymptotic bias performance of $\hat{m}(x)$ with $\lambda_0 = 0$ and λ_1 a positive constant in Remark 3.6, take the value of λ_1 as $\lambda_1 = \lambda_{1,n} > 0$. Such $\hat{m}(x)$ is H&M's shrinkage estimator $\hat{m}_S(x)$ obtained by using $\varepsilon = \lambda_1 S_0/S_2$ and $\tilde{m}_0(x) = T_0/S_0$, the Nadaraya–Watson estimator. The same conclusion can be drawn for $\hat{m}(x)$ in Remark 3.6. By Remark 3.3, if $\lambda_1 = \mathrm{o}(h^2)$, then such $\hat{m}(x)$ has the same AMSE as $\tilde{m}(x)$, for each $x \in [0,1]$. However, this sufficient condition $\lambda_1 = \mathrm{o}(h^2)$ obtained from H&M does not agree with that $\lambda_1 = \mathrm{o}(h)$ given in this paper for (9) and (10). We now explain why the value of λ_1 in each of these two sufficient conditions has to be given by that way. To verify our claim, decompose

$$\hat{m}(x) - m(x) = (T_0 S_2 - T_1 S_1 + \lambda_1 T_0)/D - m(x) = B_1 + B_2,$$

where

$$D = S_0 S_2 - S_1^2 + \lambda_1 S_0,$$

$$B_1 = \{(T_0 S_2 - T_1 S_1) - m(x)(S_0 S_2 - S_1^2)\}/D,$$

$$B_2 = \lambda_1\{T_0 - m(x) S_0\}/D.$$

Our claim is a consequence of $E(B_1^2)=\text{AMSE}\{\tilde{m}(x)\}+\text{o}(h^4+n^{-1}h^{-1})$, $E(B_2^2)=\text{O}(\lambda_1^2h^2+\lambda_1^2n^{-1}h^{-1})=\text{o}(h^4+n^{-1}h^{-1})$, for each $x\in[0,1]$, and Cauchy–Schwartz inequality. On the other hand, H&M make use of a different decomposition

$$\hat{m}(x)-m(x)=B_3+B_4,$$

where

$$B_3=(T_0S_2-T_1S_1)/D-m(x),\quad B_4=\lambda_1T_0/D.$$

H&M's result is obtained by $E(B_3^2)=\text{AMSE}\{\tilde{m}(x)\}+\text{o}(h^4+n^{-1}h^{-1})$, $E(B_4^2)=\text{O}(\lambda_1^2)=\text{o}(h^4+n^{-1}h^{-1})$, for each $x\in[0,1]$, and Cauchy–Schwartz inequality. Comparing the magnitudes of $E(B_2^2)$ and $E(B_4^2)$, H&M require the stronger condition $\lambda_1=\text{o}(h^2)$ to ensure their result.

Remark 3.8 (*Practical choice of the values of the bandwidth h and the ridge regression parameters λ_0 and λ_1*). By (9), (10), and the results given in Remark 3.7, we suggest taking the value of λ_0 as 0. For constructing $\hat{m}(x)$, the optimal values h^* and λ_1^* of h and λ_1, respectively, are taken as the minimizer of the MISE of $\hat{m}(x)$. Given the values of h and λ_1, the MISE of $\hat{m}(x)$ is defined by $\text{MISE}_{\text{LLRRE}}(h,\lambda_1)=E\{\text{ISE}_{\text{LLRRE}}(h,\lambda_1)\}$. Here $\text{ISE}_{\text{LLRRE}}(h,\lambda_1)$ is defined by

$$\text{ISE}_{\text{LLRRE}}(h,\lambda_1)=\int_0^1\{\hat{m}(x)-m(x)\}^2f(x)\,\mathrm{d}x.$$

The weighting by f puts more emphasis on accuracy in regions with more data.

Since the optimal values h^* and λ_1^* for constructing $\hat{m}(x)$ are not available in practice, they are estimated respectively by the minimizer $\hat{h}$ and $\hat{\lambda}_1$ of the cross-validation score $\text{CV}_{\text{LLRRE}}(h,\lambda_1)$ defined by

$$\text{CV}_{\text{LLRRE}}(h,\lambda_1)=\sum_{i=1}^n\{\hat{m}_i(X_i)-Y_i\}^2.$$

Here $\hat{m}_i(X_i)$ is the "leave-one-out" version of $\hat{m}(X_i)$, that is, the observation (X_i,Y_i) is left out in constructing $\hat{m}(X_i)$. Haïdle and Marron (1985) show that, for the Nadaraya–Watson estimator, the cross-validated bandwidth is asymptotically optimal with respect to the conditional MISE. For other automatic smoothing parameter selection methods, see also Rice (1984), Haïdle et al. (1988), and Marron (1988).

The same argument for choosing the parameters for constructing $\hat{m}(x)$ can be applied to $\tilde{m}(x)$. Let $\text{ISE}_{\text{LLE}}(h)$, $\text{MISE}_{\text{LLE}}(h)$, and $\text{CV}_{\text{LLE}}(h)$ be similarly defined for $\tilde{m}(x)$, and h^0 and $\tilde{h}$ denote the minimizers of $\text{MISE}_{\text{LLE}}(h)$ and $\text{CV}_{\text{LLE}}(h)$, respectively. Simulation studies given in Section 4 demonstrate that $\hat{m}(x)$ using the cross-validated bandwidth $\hat{h}$ and ridge regression parameters $\lambda_0=0$ and $\lambda_1=\hat{\lambda}_1$ could have smaller sample MISE than $\tilde{m}(x)$ using the cross-validated bandwidth $\tilde{h}$, in reasonable sample sizes.

Note that it is very often to use the mean-average-squared error (MASE) to evaluate the performance of the kernel estimators. Our purpose for using a different criterion MISE is to make the advantage of the LLRRE over the LLE more visible. We now explain it. If the MASE criterion is used, then the value of $\tilde{m}(X_i)$ has to be calculated, for each $i = 1,\dots,n$. Note that when $\tilde{m}(X_i)$ is calculated, there will be at least one design point X_i falling in the compact window $[X_i - h, X_i + h]$ around X_i. In this case, $\tilde{m}(X_i)$ will not suffer from the drawback that there is no design point falling in the compact window around X_i. Hence, the erratic behavior of the LLE caused by this drawback cannot be felt by the MASE measure. The same remark applies to the LLRRE. Therefore, to compare the performance of the LLE and that of the LLRRE, the criterion MASE is not advisable.

4. Simulations

To investigate the practical implications of the results for the LLRRE $\hat{m}(x)$ and those for the LLE $\tilde{m}(x)$ presented in Section 3, an empirical study was carried out. The simulated regression settings are introduced in the following. Three sample sizes $n = 25, 50$, and 100 were considered. The regression function $m(x)$ was $m(x) = x^3(1-x)^3 I_{[0,1]}(x)$. The regression errors ε_i were Normal$(0, \sigma^2)$ variables, where $\sigma = 0.003$. Two design densities were employed. One is the Uniform$(0,1)$ density. The other is the Beta$(1/3, 1)$ density which is the density of the cubic Uniform$(0,1)$ variable. The data sparsity issue produced by the latter design density is more serious than that by the former one. The responses Y_i were generated from the regression model (1). For each sample size and each design density, 1000 independent sets of observations (X_i, Y_i) were generated. The kernel function used in $\hat{m}(x)$ and $\tilde{m}(x)$ was the Epanechnikov kernel $K(x) = (3/4)(1 - x^2)I_{[-1,1]}(x)$.

For each data set, the values of $\mathrm{ISE}_{\mathrm{LLRRE}}(h, \lambda_1)$ and $\mathrm{CV}_{\mathrm{LLRRE}}(h, \lambda_1)$ were calculated on an equally spaced logarithmic grid of 200 × 1000 values of h and λ_1. Here the 200 values of h were selected in $[0.05, 0.5]$ and the 1000 values of λ_1 were taken in $[0.0001, 10]$. 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 λ_1, the value of $\mathrm{ISE}_{\mathrm{LLRRE}}(h, \lambda_1)$ was approximated by $(1/u)\sum_{i=1}^{u}\{\hat{m}(t_i) - m(t_i)\}^2 f(t_i)$, where $t_i = (2i-1)/(2u)$ and $u = 1000$. Also, the values of $\mathrm{MISE}_{\mathrm{LLRRE}}(h, \lambda_1)$ and $\mathrm{SISE}_{\mathrm{LLRRE}}(h, \lambda_1)$, where $\mathrm{SISE}_{\mathrm{LLRRE}}(h, \lambda_1)$ denotes the standard deviation of $\mathrm{ISE}_{\mathrm{LLRRE}}(h, \lambda_1)$, were empirically approximated by the sample average and standard deviation, respectively, of $\mathrm{ISE}_{\mathrm{LLRRE}}(h, \lambda_1)$ over the 1000 pseudo-data sets. After evaluation on the grid, the global minimizers h^* and λ_1^* of $\mathrm{MISE}_{\mathrm{LLRRE}}(h, \lambda_1)$ and $\hat{h}$ and $\hat{\lambda}_1$ of $\mathrm{CV}_{\mathrm{LLRRE}}(h, \lambda_1)$ were taken on the grid.

When h^* and λ_1^* were obtained, the values of $\mathrm{MISE}_{\mathrm{LLRRE}}(h^*, \lambda_1^*)$ and $\mathrm{SISE}_{\mathrm{LLRRE}}(h^*, \lambda_1^*)$ were calculated. The former measures the best performance of $\hat{m}(x)$. On the other hand, the value of $\mathrm{MISE}_{\mathrm{LLRRE}}(\hat{h}, \hat{\lambda}_1)$ measures the performance of $\hat{m}(x)$ which can be attained in practice. The same computation procedures were applied to calculate $\mathrm{MISE}_{\mathrm{LLE}}(h^0)$, $\mathrm{SISE}_{\mathrm{LLE}}(h^0)$, and $\mathrm{MISE}_{\mathrm{LLE}}(\tilde{h})$. Here the values of $\mathrm{MISE}_{\mathrm{LLE}}(h)$ and $\mathrm{CV}_{\mathrm{LLE}}(h)$ were calculated on an equally spaced logarithmic grid of 200 values of h in the interval $[0.05, 0.5]$. The simulation results are summarized in Tables 1 and 2.

Table 1
Values of $\mathrm{MISE}_{\mathrm{LLRRE}}(h^*, \lambda_1^*)$ and $\mathrm{SISE}_{\mathrm{LLRRE}}(h^*, \lambda_1^*)$ (given in the parentheses) for $\hat{m}(x)$, and those of $\mathrm{MISE}_{\mathrm{LLE}}(h^0)$ and $\mathrm{SISE}_{\mathrm{LLE}}(h^0)$ for $\tilde{m}(x)$. These values have been multiplied by 10^6

n	$\hat{m}(x)$	$\tilde{m}(x)$
Uniform (0,1) density		
25	2.41 (1.76)	6.44 (2.93)
50	1.26 (0.73)	1.55 (1.05)
100	0.69 (0.37)	0.79 (0.43)
Beta(1/3,1) density		
25	2.81 (2.15)	20.5 (237.0)
50	1.27 (0.87)	2.89 (6.68)
100	0.65 (0.35)	1.19 (1.10)

Table 2
Values of the sample mean and standard deviation (given in the parentheses) of $\mathrm{MISE}_{\mathrm{LLRRE}}(\hat{h}, \hat{\lambda}_1)$ for $\hat{m}(x)$ and those of $\mathrm{MISE}_{\mathrm{LLE}}(\tilde{h})$ for $\tilde{m}(x)$, and the number of times N out of the 1000 pseudo-data sets that the values of $\mathrm{MISE}_{\mathrm{LLRRE}}(\hat{h}, \hat{\lambda}_1)$ are larger than those of $\mathrm{MISE}_{\mathrm{LLE}}(\tilde{h})$. The values of the sample means and standard deviations have been multiplied by 10^6

n	$\hat{m}(x)$	$\tilde{m}(x)$	N
Uniform (0,1) density			
25	7.26 (14.2)	3985.1 (37463.3)	2
50	2.33 (5.05)	292.1 (1512.6)	98
100	0.90 (0.59)	1.60 (2.70)	126
Beta(1/3,1) density			
25	7.12 (7.24)	583.6 (640.7)	0
50	3.20 (5.04)	324.7 (408.9)	4
100	1.10 (1.27)	25.9 (97.2)	4

Table 1 shows that, for each sample size and each design density, the best performance $\mathrm{MISE}_{\mathrm{LLRRE}}(h^*, \lambda_1^*)$ of $\hat{m}(x)$ is better than that $\mathrm{MISE}_{\mathrm{LLE}}(h^0)$ of $\tilde{m}(x)$. The magnitude of the difference between $\mathrm{MISE}_{\mathrm{LLRRE}}(h^*, \lambda_1^*)$ and $\mathrm{MISE}_{\mathrm{LLE}}(h^0)$ increases, when the design density moves from the Uniform$(0,1)$ design to the Beta$(1/3,1)$ design. This is a result of the fact that the latter design has a more serious data sparsity issue than the former one. The same situation also occurs when the sample size decreases. Table 2 contains the sample mean and standard deviation of $\mathrm{MISE}_{\mathrm{LLRRE}}(\hat{h}, \hat{\lambda}_1)$ and those of $\mathrm{MISE}_{\mathrm{LLE}}(\tilde{h})$. It also gives the number of times N out of the 1000 pseudo-data sets that the values of $\mathrm{MISE}_{\mathrm{LLRRE}}(\hat{h}, \hat{\lambda}_1)$ are larger than those of $\mathrm{MISE}_{\mathrm{LLE}}(\tilde{h})$. Considering the values of the sample mean and N, the practical performance of $\hat{m}(x)$ is still better than that of $\tilde{m}(x)$.

Acknowledgements

We thank the referees, the associate editor, and the editor for their many valuable comments which substantially improved the presentation. The research was supported by the National Science Council, Republic of China.

Appendix: Sketches of the proofs

The following notation and results will be used in this section. Let I_n denote a $n\times n$ identity matrix. Set $W_k=n^{-2}h^{-2}\sum_{j=1}^n Z_j^k K(Z_j)^2$, for $k\geqslant 0$, and $D=(\lambda_0+S_0)(\lambda_1+S_2)-S_1^2$. Since the kernel function K is supported on the interval $[-1,1]$,

$$|S_k|\leqslant n^{-1}h^{-1}\sum_{j=1}^n |Z_j^k|K(Z_j)\leqslant S_0 \quad \text{for each } k\geqslant 0. \tag{A.1}$$

Through a straightforward calculation, we have

$$W_0\leqslant S_0^2,\quad S_0\leqslant h^{-1}K^0,\quad D\geqslant\lambda_1 S_0,\quad D\geqslant\lambda_0(\lambda_1+S_2), \tag{A.2}$$

and $\hat{m}(x)$ can be expressed by

$$\hat{m}(x)=\sum_{j=1}^n q_j Y_j, \tag{A.3}$$

where

$$q_j=n^{-1}h^{-1}\{(\lambda_1+S_2)-Z_jS_1\}K(Z_j)D^{-1}.$$

Proof of Theorem 3.1. We first give the proof of (5). Using (A.1)–(A.3), the proof of (5) is complete by showing

$$\begin{aligned}\mathrm{Var}\{\hat{m}(x)\,|\,X^*\} &= \sigma^2 n^{-2}h^{-2}\sum_{j=1}^n(\lambda_1+S_2-Z_jS_1)^2K(Z_j)^2D^{-2}\\ &\leqslant \sigma^2W_0(\lambda_1+S_2+|S_1|)^2\lambda_1^{-2}S_0^{-2}\leqslant\sigma^2(\lambda_1+2S_0)^2\lambda_1^{-2}\\ &\leqslant \sigma^2(1+2\lambda_1^{-1}h^{-1}K^0)^2.\end{aligned}$$

We now give the proof of (6). For this, using (A.3) and applying the second order Taylor expansion to $m(X_j)$, through a straightforward calculation, the conditional bias of $\hat{m}(x)$ can be expressed by

$$\mathrm{Bias}\{\hat{m}(x)\,|\,X^*\}=\sum_{j=1}^n q_j m(X_j)-m(x)=A_1+A_2+A_3,$$

where

$$A_1=(-1)\lambda_0(\lambda_1+S_2)m(x)D^{-1},$$

$$A_2=(-1)h\lambda_1S_1m_1(x)D^{-1},$$

$$A_3=(1/2)h^2\{(\lambda_1+S_2)S_2^*-S_1S_3^*\}D^{-1}.$$

Here $S_k^*=n^{-1}h^{-1}\sum_{j=1}^k Z_j^kK(Z_j)m_2(\xi_j)$, for $k\geqslant 0$, where ξ_j lies inbetween x and X_j. Using (A.1), (A.2), and $|S_k^*|\leqslant m_2^0n^{-1}h^{-1}\sum_{j=1}^n|Z_j^k|K(Z_j)$, for $k\geqslant 0$, the proof of (6) is complete by showing

$$|A_1| \leqslant \begin{bmatrix} 0 & \text{for } \lambda_0 = 0, \\ m^0\lambda_0(\lambda_1 + S_2)\lambda_0^{-1}(\lambda_1 + S_2)^{-1} & \text{for } \lambda_1 > 0, \end{bmatrix} = \begin{bmatrix} 0 & \text{for } \lambda_0 = 0, \\ m^0 & \text{for } \lambda_1 > 0, \end{bmatrix}$$

$$|A_2| \leqslant hm_1^0\lambda_1 S_1 \lambda_1^{-1} S_0^{-1} \leqslant hm_1^0,$$

$$|A_3| \leqslant (1/2)h^2 m_2^0\{(\lambda_1 + S_2)S_0 + S_0 S_2\}\lambda_1^{-1}S_0^{-1} \leqslant h^2 m_2^0(1/2 + \lambda_1^{-1}h^{-1}K^0).$$

Hence, the proof of Theorem 3.1 is complete. □

Proof of Theorem 3.2. The proof of Theorem 3.2 is omitted since it is essentially the same as those of the asymptotic variance and bias of the Nadaraya–Watson estimator in (8.15) of Scott (1992) by using approximations to the standard errors of functions of random variables given in Section 10.5 of Stuart and Ord (1987).

References

Epanechnikov, V.A., 1969. Nonparametric estimation of a multidimensional probability density. Theory Probab. Appl. 14, 153–158.

Eubank, R.L., 1988. Spline Smoothing and Nonparametric Regression. Marcel Dekker, New York.

Fan, J., 1993. Local linear regression smoothers and their minimax efficiencies. Ann. Statist. 21, 196–216.

Fan, J., Gijbels, I., 1992. Variable bandwidth and local linear regression smoothers. Ann. Statist. 20, 2008–2036.

Fan, J., Gijbels, I., 1996. Local Polynomial Modeling and Its Application — Theory and Methodologies. Chapman & Hall, New York.

Gasser, T., Müller, H.G., 1979. Kernel estimation of regression functions. In: Gasser, T., Rosenblatt, M. (Eds.), Smoothing Techniques for Curve Estimation. Springer, Heidelberg, pp. 23–68.

Gasser, T., Müller, H.G., 1984. Estimating regression functions and their derivatives by the kernel method. Scand. J. Statist. 11, 171–185.

Hall, P., Marron, J.S., 1997. On the role of the shrinkage parameter in local linear smoothing. Probab. Theory Related Fields 108, 495–516.

Haïdle, W., 1990. Applied Nonparametric Regression. Cambridge University Press, New York.

Haïdle, W., 1991. Smoothing Techniques: with Implementation in S, Springer Series in Statistics. Springer, Berlin.

Haïdle, W., Hall, P., Marron, J.S., 1988. How far are automatically chosen regression smoothing parameters from their optimal? J. Amer. Statist. Assoc. 83, 86–101.

Haïdle, W., Marron, J.S., 1985. Optimal bandwidth selection in nonparametric regression function estimation. Ann. Statist. 13, 1465–1481.

Marron, J.S., 1988. Automatic smoothing parameter selection: a survey. Empirical Econom. 13, 187–208.

Marron, J.S., Wand, M.P., 1992. Exact mean integrated squared error. Ann. Statist. 20, 712–736.

Montgomery, D.C., Peck, E.A., 1982. Introduction to Linear Regression Analysis. Wiley, New York.

Müller, H.G., 1988. Nonparametric Regression Analysis of Longitudinal Data. Lecture Notes in Statistics, Vol. 46. Springer, Berlin.

Nadaraya, E.A., 1964. On estimating regression. Theory Probab. Appl. 9, 141–142.

Rice, J., 1984. Bandwidth choice for nonparametric regression. Ann. Statist. 12, 1215–1230.

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

Seifert, B., Gasser, T., 1996. Finite-sample analysis of local polynomials: Analysis and solutions. J. Amer. Statist. Assoc. 91, 267–275.

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

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

Wand, M.P., Jones, M.C., 1995. Kernel Smoothing. Chapman & Hall, New York.

Watson, G.S., 1964. Smooth regression analysis. Sankhya Ser. A 26, 359–372.

Wu, J.S., Chu, C.K., 1992. Double smoothing for kernel estimators in nonparametric regression. J. Nonparametric Statist. 1, 375–386.