Statistics and Computing (1997) **7**, 11–17

On the shrinkage of local linear curve estimators

MING-YEN CHENG[1,2], PETER HALL[1] and D. M. TITTERINGTON[1,3]

[1]*Centre for Mathematics and its Applications, Australian National University, Canberra, ACT 0200, Australia*
[2]*Institute of Mathematical Statistics, National Chung Cheng University, Minghsiung, Chiayi, Taiwan, Republic of China*
[3]*Department of Statistics, University of Glasgow, Glasgow G12 8QW, UK*

Received December 1995 and accepted October 1996

Local linear curve estimators are typically constructed using a compactly supported kernel, which minimizes edge effects and (in the case of the Epanechnikov kernel) optimizes asymptotic performance in a mean square sense. The use of compactly supported kernels can produce numerical problems, however. A common remedy is 'ridging', which may be viewed as shrinkage of the local linear estimator towards the origin. In this paper we propose a general form of shrinkage, and suggest that, in practice, shrinkage be towards a proper curve estimator. For the latter we propose a local linear estimator based on an infinitely supported kernel. This approach is resistant against selection of too large a shrinkage parameter, which can impair performance when shrinkage is towards the origin. It also removes problems of numerical instability resulting from using a compactly supported kernel, and enjoys very good mean squared error properties.

Keywords: Bandwidth, bias, compactly supported kernel, kernel estimator, mean squared error, ridge parameter, smoothing, variance

1. Introduction

Local polynomial methods have attractive numerical and theoretical properties, which have received considerable attention in recent years. They are particularly adaptive (Hastie and Loader, 1993), and enjoy minimax optimality (Fan, 1993). Jones (1993) investigated applying local polynomial regression techniques to density estimation. Therefore it comes as no surprise to find that local polynomial methods are widely represented in software for curve estimation, such as LOESS; see for example Cleveland (1979, 1993), Cleveland and Devlin (1988) and Cleveland and Grosse (1991). Nevertheless, they do suffer drawbacks, of which perhaps the most serious are numerical problems in cases of sparse design density. In its 'naive' form, a local linear estimator based on a compactly supported kernel may equal a non-zero number divided by zero, and so may not be well defined. In less extreme cases, small but non-zero values of the denominator in the definition of the estimator can produce erratic fluctuations, which impair performance.

There are a variety of ways of overcoming this difficulty, all of which have side effects. One is based on using an empirical bandwidth determined by the design density, producing a larger bandwidth in places where design points are sparse. Ideally, however, local bandwidth choice should be determined by the curvature of the target function, as well as by the design density and the local variance, and is particularly prone to stochastic error when the design is sparse. Interpolation methods were discussed in Hall and Turlach (1995). Another remedy involves shrinking the local linear estimator towards the origin, by incorporating a ridge parameter.

The ridge acts to some extent as a smoothing parameter, however, with the result that the increase in numerical stability produced by the ridge parameter can be accompanied by an increase in bias. As a result, overall performance may be impaired. In this paper we suggest a general formulation of the shrinkage approach, allowing a local linear estimator to be shrunken in the direction of a general curve estimator which may be chosen to be free from the instability properties that we seek to remove. An advantage of this

approach, relative to shrinkage towards the origin, is that it produces an estimator that is robust against excessively large choice of the shrinkage parameter—in the case of our rule, taking that quantity to be infinite still produces a proper curve estimator, rather than simply zero. This largely overcomes the bias problem mentioned earlier.

Next we introduce our method. Suppose data $\chi = \{(X_i, Y_i),\ 1 \le i \le n\}$ are generated by the model $Y_i = g(X_i) + e_i$, where the X_is are independent and identically distributed random variables with density f, and, conditional on $\chi = \{X_1, \ldots, X_n\}$, the e_is have zero mean and are uncorrelated, with $\mathrm{var}(e_i|\chi) = \sigma(X_i)^2$. Here, σ^2 is a smooth, non-negative function. A local linear estimator of $g(x)$, using a kernel K and bandwidth h, may be defined as $a = a(x)$ where a and b are chosen to minimize

$$\sum_{i=1}^{n} \{Y_i - a - b(x - X_i)\}^2 K\{(x - X_i)/h\}$$

A general approach to shrinking an amount $\epsilon = \epsilon(x)$ away from the solution of this problem, and towards another estimator $\tilde{g}$, is as follows. Let K be a compactly supported kernel and h a bandwidth. Choose a and b to minimize

$$\sum_{i=1}^{n} \{Y_i - a - b(x - X_i)\}^2 K\{(x - X_i)/h\} + \epsilon\{\tilde{g}(x) - a\}^2$$

or equivalently to minimize

$$\sum_{i=1}^{n+1} \{Y_i - a - b(x - X_i)\}^2 K_i$$

where $K_i = K\{(x - X_i)/h\}$ if $1 \le i \le n$, $K_{n+1} = \epsilon$, $Y_{n+1} = \tilde{g}(x)$ and $X_{n+1} = x$. Let $\hat{g} = a$ be the solution of this new problem. Thus, $\hat{g}$ is obtained by applying local linear smoothing to the original data χ, augmented by the addition of a single new value $(x, \tilde{g}(x))$ which is given a special weight ϵ.

Explicitly

$$\hat{g} = \left(\sum_{i=1}^{n+1} w_i Y_i\right) \Big/ \left(\sum_{i=1}^{n+1} w_i\right)$$

where $w_i = \{s_2 - (x - X_i)s_1\}K_i$ and

$$s_k = \sum_{i=1}^{n+1} (x - X_i)^k K_i = \sum_{i=1}^{n} (x - X_i)^k K\{(x - X_i)/h\} \quad (1)$$

Thus, for $1 \le i \le n$ the weights w_i are exactly as in the case of the naive local linear estimator, $\bar{g}$ say, defined by taking $\epsilon = 0$; and

$$\hat{g} = \left(\sum_{i=1}^{n} w_i Y_i + \epsilon s_2 \tilde{g}\right) \Big/ \left(\sum_{i=1}^{n} w_i + \epsilon s_2\right) \quad (2)$$

Formula (2) makes it clear that $\hat{g}$ is obtained by (a) shrinking the numerator in the definition of $\bar{g}$ towards $\tilde{g}$, in the classical sense of shrinkage; (b) shrinking the denominator towards the constant 1, using the same shrinkage parameter as the numerator; and (c) re-forming the ratio. We say that $\hat{g}$ results from shrinking $\bar{g}$ towards $\tilde{g}$. Note that $\hat{g} = \bar{g}$ when $\epsilon = 0$, while $\epsilon = \infty$ gives $\hat{g} = \tilde{g}$.

Taking $\tilde{g} = 0$ in (2) we obtain the classical prescription for ridging a local linear estimator, where a small positive weight (here, ϵs_2) is added to the denominator to prevent it from straying too close to zero. For example, Fan (1993) suggested letting $\epsilon = n^{-1}$ or $(ns_2)^{-1}$, and Hall and Marron (1995) discussed the general choice of ridge parameter. See also Hall *et al.* (1995). Taking $\tilde{g} = \bar{Y} = n^{-1}\Sigma_{i \le n} Y_i$ we have the result of shrinking a local linear estimator towards the mean of the response variables. More reasonably, in our view, $\tilde{g}$ could be a proper curve estimator, such as one based on the Nadaraya–Watson method (e.g. Härdle, 1990, Section 3.1; or Wand and Jones, 1995, p. 119), or on interpolation (Clark, 1977, 1980), or on convolution (Gasser and Müller, 1979; Müller, 1988).

We suggest taking $\tilde{g}$ to be another local linear estimator, employing the same bandwidth as $\bar{g}$ but using a kernel with infinite support. The infinite support helps to alleviate problems caused by data sparseness. A stabilization method proposed by Seifert and Gasser (1995, 1996) may be viewed as shrinkage towards a Nadaraya–Watson kernel estimator with the same kernel as the original local linear estimator.

Section 2 will introduce our method from a slightly different viewpoint, and outline its theoretical properties. Numerical performance will be addressed in Section 3. An outline proof of the main result of Section 2 will be given in the Appendix.

2. Methodology

Let K and L be bounded, non-negative functions, satisfying

$$\int u^2 M(u)\mathrm{d}u < \infty \quad \text{and} \quad \int uM(u)\mathrm{d}u = 0$$

where M denotes either K or L. We ask that K be compactly supported; and that L be monotone for sufficiently large positive and large negative values of its argument, and such that const.$(1 + |x|)^{-\alpha_2} \le L(x) \le$ const.$(1 + |x|)^{-\alpha_1}$ for some $3 < \alpha_1 \le \alpha_2 < \infty$. We refer to these conditions as (C_1). An appropriate L would be a Student's t density with three or more degrees of freedom. The final estimator is obtained by 'mixing' the local linear weights for K and L in the proportion $1 : \eta$, respectively, where $\eta = \eta(n) > 0$ is proportional to a shrinkage parameter.

Define s_k as in (1), and $w_i^{(K)}(x) = \{s_2 - (x - X_i)s_1\}$ $K\{(x - X_i)/h\}$. Let $w_i^{(L)}$ denote the version of $w_i^{(K)}$ which results if K is replaced by L. Our estimator is given by

$$\hat{g} = \left\{\sum_{i=1}^{n} (w_i^{(K)} + \eta w_i^{(L)}) Y_i\right\} \Big/ \left\{\sum_{i=1}^{n} (w_i^{(K)} + \eta w_i^{(L)})\right\} \quad (3)$$

To appreciate that $\hat{g}$ has the form suggested in Section 1, let $\tilde{g}$ there be the local linear estimator based on L rather than K, i.e.

$$\tilde{g} = \left(\sum_{i=1}^{n} w_i^{(L)} Y_i\right) \Big/ \left(\sum_{i=1}^{n} w_i^{(L)}\right)$$

and put $\epsilon = \eta s_2^{-1} \Sigma_{i \le n} w_i^{(L)}$. Then the definition of $\hat{g}$ at (2) produces the estimator at (3).

The advantages of using a local linear estimator based on a compactly supported kernel, rather than an infinitely supported one, include reduced edge effects and lower mean squared error (if the Epanechnikov kernel is employed). On the other hand, an infinitely supported kernel produces an estimator which enjoys a lower level of numerical problems. The shrunken form, $\hat{g}$, allows access to the best of both these worlds. To achieve this it is usually necessary for L to be a kernel with tails heavier than those of the Normal. Numerical evidence is provided in Section 3. For smaller sample sizes, and in contexts where edge effects are less important, the optimal version of $\hat{g}$ may be closer to $\tilde{g}$ than to the local linear estimator based on K, but for large sample sizes and in cases where edge effects are significant, the optimal $\hat{g}$ will be close to the naive local linear estimator based on K. Section 3 will explore these properties.

Next we state a result which demonstrates that, under mild conditions, the unconditional mean square performance of $\hat{g}$ is asymptotic to the renowned conditional performance of $\bar{g}$. Assume that $h + (nh)^{-1} = O(n^{-\delta})$ for some $\delta > 0$; that $\eta \to 0$ and $\eta^{-1} = O(n^{\gamma})$ for some $\gamma > 0$; that f is supported on $\mathcal{I} = [0, 1]$, and continuous and non-vanishing there; that g has two continuous derivatives on $\mathcal{I}$; and that σ is continuous on $\mathcal{I}$. We call these conditions (C_2). Let $\mathcal{J}$ denote any open subset of $\mathcal{I}$, and define

$$\kappa_1 = \left(\int K^2\right) \Big/ \left(\int K\right)^2, \quad \kappa_2 = \frac{1}{4} \left\{\int u^2 K(u) \mathrm{d}u \Big/ \int K\right\}^2$$

Theorem 2.1. Assuming conditions (C_1) and (C_2)

$$E(\hat{g} - g)^2 = (nh)^{-1} \kappa_1 \sigma^2 f^{-1} + h^4 \kappa_2 (g'')^2 + o\{(nh)^{-1} + h^4\} \qquad (4)$$

uniformly in $x \in \mathcal{J}$.

The range of validity of (4) may be extended right to the very ends of $\mathcal{I}$ if minor qualifications are made. First, the variance contribution to the right-hand side of (4), $v = (nh)^{-1} \kappa_1 \sigma^2 f^{-1}$, only admits that formula away from the ends of $\mathcal{I}$. Within distance $O(h)$ of the ends it is inflated by a constant factor, for example to $2v$ at zero and one. The order of magnitude of variance remains, however, $(nh)^{-1}$. Secondly, owing to the fact that L is not compactly supported, the second term in (4), representing the squared bias contribution, is not applicable at the ends of $\mathcal{I}$ unless η is sufficiently small. The squared bias contribution remains at $h^4 \kappa_2 (g'')^2 + o(h^4)$ if $\eta = o(h^2)$, and equals $O(h^4)$ if $\eta = O(h^2)$. Noting these changes, a number of variants of Theorem 2.1 are available for describing performance of the estimator uniformly on $\mathcal{I}$, rather than just on the subinterval $\mathcal{J}$. One of them is as follows: if conditions (C_1) and (C_2) hold, and if $\eta = O(h^2)$, then

$$E(\hat{g} - g)^2 = O\{(nh)^{-1} + h^4\}$$

uniformly in $x \in \mathcal{I}$.

The estimator $\hat{g}$ employs two local linear estimators based on different kernels K and L. It is important to rescale the kernels properly, for example using the canonical kernels of Marron and Nolan (1989), so that they yield the same amount of smoothing. This is incorporated in all the numerical studies of Section 3.

3. Numerical properties

In this section we summarize a simulation study which examines finite sample properties of $\hat{g}$. Throughout we took K to be the biweight kernel, and L to be Student's t density with 5 degrees of freedom. Figure 1 illustrates typical regression estimates obtained by naive local linear fitting based on Normal and biweight kernels (i.e. with $\eta = 0$), and our shrinkage method with $\eta = 50^{0.5}$ or 50^{20}. (Taking $\eta = 50^{20}$ produces basically the native local linear estimator based on L.) The regression mean was linear with a Gaussian peak

$$g(x) = 2 - 5x + 5 \exp\{-400(x - 0.5)^2\} \qquad (5)$$

The design density was Uniform on $\mathcal{I} = [0, 1]$, errors were Normally distributed with variance $\sigma^2 = 0.5$, and sample size was $n = 50$. The plus signs in this and other figures indicate data points.

Both of the naive local linear estimates depicted in Fig. 1 are particularly rough, and stray well away from the true curve in places where design points are sparse. On the other hand, the modified local linear estimates suffer less from numerical fluctuation. Compared to the local linear estimate with shrinkage parameter $\eta = 50^{20}$, using $\eta = 50^{0.5}$ gives a more erratic curve in places where it should be linear, but produces a better estimate of the peak.

Using the same regression mean, design density and error distribution as before, but employing a wider range of values of n and η, we conducted extensive simulations to investigate properties of the mean integrated squared error (MISE) of $\hat{g}$ as an estimator of g, defined at (5). We took $n = 25$, 50, 100, 250, 500 or 1000, and $\eta = n^{-4}$, n^{-2}, $n^{-0.5}$, $n^{-0.2}$, $n^{0.2}$, $n^{0.5}$ or n^{10}. To avoid edge effects we estimated g on the interval $[-1, 2]$ rather than $[0, 1]$, and generated n uniformly distributed design points on each subinterval $[-1, 0]$, $[0, 1]$ and $[1, 2]$. (The influence of edge effects when the design points are restricted to $[0, 1]$ will be addressed

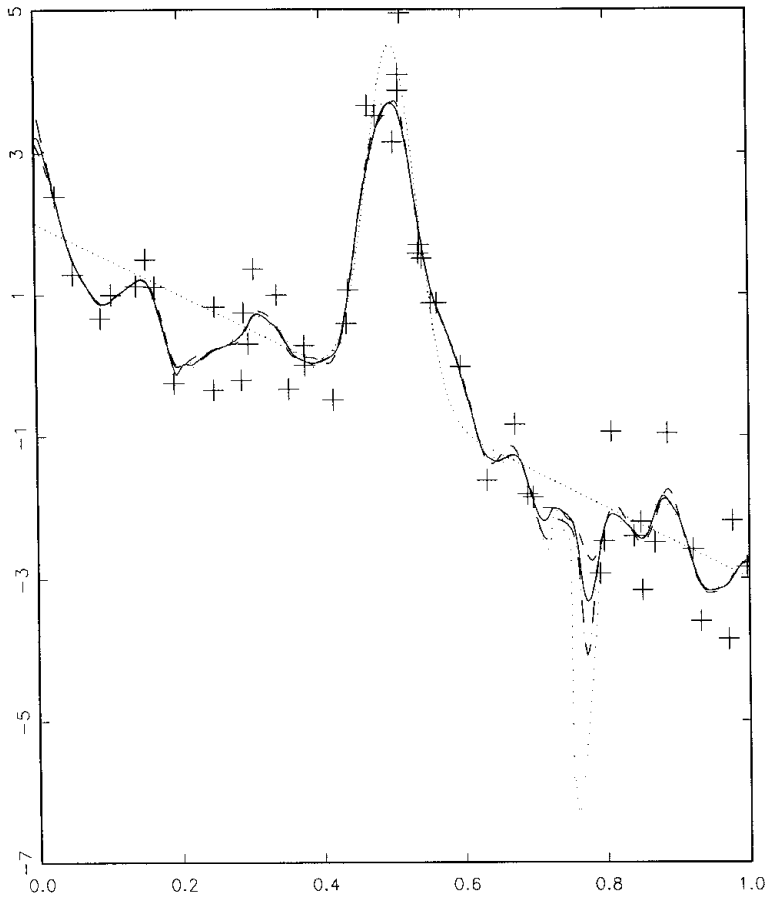

Fig. 1. *Typical estimates of the regression mean at* (5) (*closely-spaced dotted line*) *constructed by naive local linear regression based on the biweight* (*dotted line*) *or Normal* (*dotted-and-dashed line*) *kernels and the suggested modifications using shrinkage parameter* $\eta = 50^{0.5}$ (*solid line*) *or* 50^{20} (*short-dashed line*). *For all estimates, the asymptotically optimal bandwidths were used. Data are indicated by the plus signs*

shortly.) Here and in other MISE comparisons below, each MISE was estimated as the average of 4000 realizations of integrated squared error, the values of which were calculated by the trapezoidal rule from corresponding squared errors evaluated on a grid of 400 equally spaced points. There were 51 logarithmically equispaced bandwidths, ranging from half of the asymptotically optimal bandwidth when $n = 1000$ and $\sigma^2 = 0.05$, to four times the asymptotically optimal bandwidth when $n = 25$ and $\sigma^2 = 0.5$. Figure 2 shows the simulated MISE curves for $n = 25$, 50 and 100. When $n = 250$, 500 or 1000, the seven values of η yielded simulated MISE curves very close to each other and are not shown in the figure. For particularly small values of η, not illustrated in the figure, estimated MISE took very large values, as predicted theoretically. (The MISE is infinite when $\eta = 0$.)

For smaller sample sizes ($n = 25$, 50 and 100), design sparseness occurs relatively often. Thus, we see from Fig. 2 that larger values of η, equal to positive powers of sample size, produce minimum MISEs. The optimal η's are 25^{10}, 50^{10} and $100^{0.2}$ in the cases of those respective sample sizes, and $n^{0.5}$ is never worse than second best. For larger sample sizes (250, 500 and 1000), design sparseness is less of a problem, and optimal shrinkage parameters are smaller and equal to negative powers of sample size. Indeed, for those sample sizes the optimal ηs are 250^{-4}, 500^{-4} and 1000^{-4}, and neither n^{-2} nor n^{-4} is ever worse than second best. When $n \geq 250$ the values of minimum MISE when $\eta = n^c$ for $-4 \leq c \leq -2$ are virtually indistinguishable from one another, and vary little in the range $-4 \leq c \leq 0.5$. Generally, taking $\eta = \sqrt{n}$ produces good performance, at the optimal bandwidth, for all $n \geq 25$.

We compared the performance of our estimator $\hat{g}$ and the local linear ridge estimators suggested by Seifert and Gasser (1995, 1996). The ridge parameters for Seifert and Gasser's methods (i) and (ii), in their notation, were fixed at $c_1 = 0.001$ and $c_2 = 1$, respectively. Their methods require bandwidth adjustment when there are very few data within the smoothing region; we selected the bandwidth to ensure that at least two points were always included. Using the same parameter settings as before, and employing $\eta = \sqrt{n}$ throughout, our approach produces lower MISE than Seifert and Gasser's method for small to moderate n, and virtually identical MISE for larger n. The simulated MISE curves for $n = 25$, 50 and 100 are illustrated in Fig. 3.

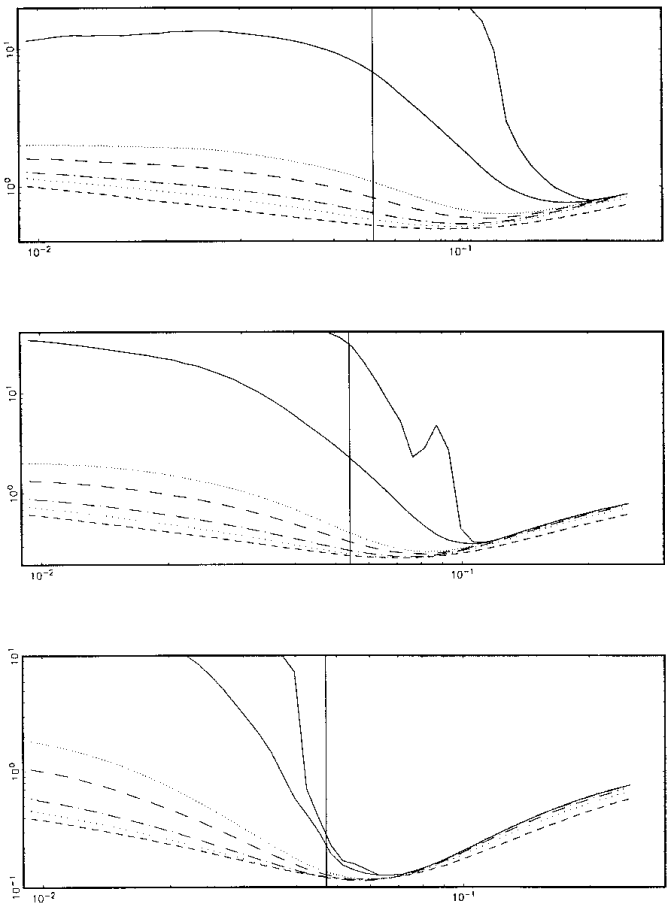

Fig. 2. *Simulated MISE as a function of bandwidth (on logarithmic scales) for the estimator $\hat{g}$, for sample sizes $n = 25$ (upper panel), 50 (middle panel) or 100 (lower panel) and seven different values of the shrinkage parameter (n^{-4}: upper solid line; n^{-2}: lower solid line; $n^{-0.5}$: closely-spaced dotted line; $n^{-0.2}$: dashed line; $n^{0.2}$: dotted-and-dashed line; $n^{0.5}$: dotted line; n^{0}: short-dashed line). The regression mean was that defined at (5), design points were Uniformly distributed, and errors were Normally distributed with variance 0.5. The design was extended to $[-1, 2]$ to avoid edge effects. Here and in other figures, vertical lines indicate asymptotically optimal bandwidths for the biweight kernel*

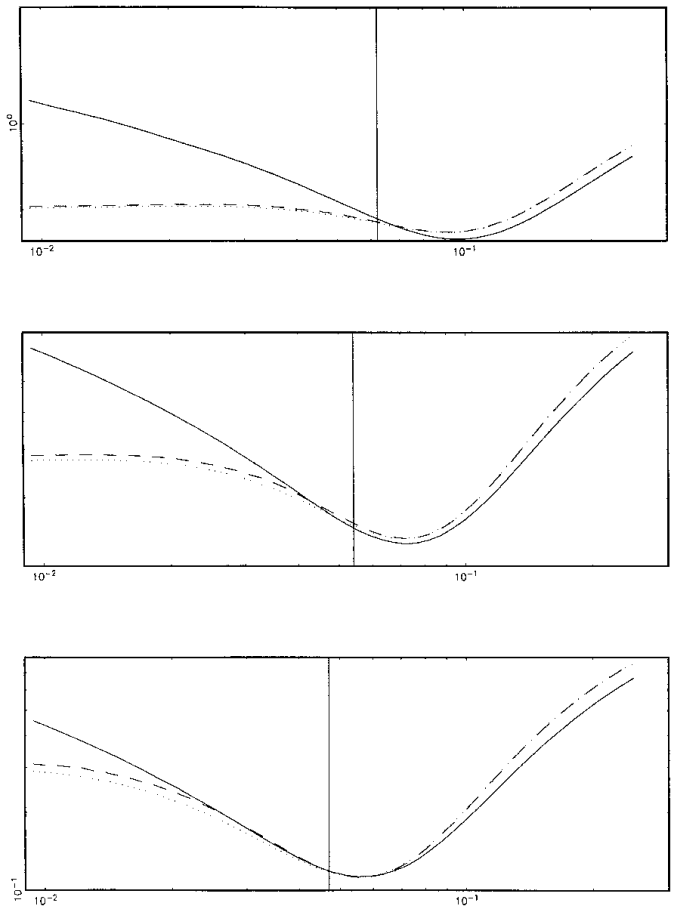

Fig. 3. *Simulated MISE as a function of bandwidth (on logarithmic scales) for the estimator $\hat{g}$ and sample sizes $n = 25$ (upper panel), 50 (middle panel) or 100 (lower panel). The solid line represents the MISE curve for $\hat{g}$ using $\eta = \sqrt{n}$. Seifert and Gasser's methods (i), using ridge parameter 0.001, and (ii), with ridge parameter 1, are represented by dotted and dashed lines, respectively. Other parameter settings, including extension of design to $[-1, 2]$, are as in Fig. 2*

Next we examine edge effects. To make estimation particularly difficult for all methods, we simulated in the case where the regression mean is given by

$$g_2(x) = 2 - 5x + 5\exp(-400x^2) + 5\exp\{-400(x - 0.5)^2\} + 5\exp\{-400(x - 1)^2\} \quad (6)$$

representing a linear function with three Gaussian peaks. Sample size was $n = 50$, design was Uniform and confined to $[0, 1]$, and errors were Normally distributed with variance $\sigma^2 = 0.5$. Figure 4 depicts representative estimates with different levels of shrinkage. Using a large value of η (e.g. 50^{20}) gives rise to relatively serious edge effects, particularly in the region $[0, 0.05] \cup [0.85, 1]$, but produces no numerical aberration arising from sparse design. Without any shrinkage, i.e. with $\eta = 0$, the local linear estimate has only minor boundary problems but exhibits a relatively high level of numerical aberration. Taking $\eta = 50^{0.5}$ gives a good compromise between these extremes.

To study the influence of edge effects on MISE we generated data using the parameter settings of Fig. 2, except that no design point was chosen outside $[0, 1]$. In particular, the regression mean at (5) was employed. The MISE curves are similar to those in Fig. 2 and are not shown here. In the cases of sample sizes 25, 50, 100, 250, 500 and 1000, the optimal values of η are respectively 25^{10}, 50^{10}, $100^{0.5}$, $250^{0.2}$, 500^{-4} and 1000^{-4}. Again, as in the case of Fig. 2, $\eta = \sqrt{n}$ never produces worse than second-best performance when $n = 25$, 50 or 100; neither n^{-2} nor n^{-4} produces worse than second-best performance when $n = 500$ or 1000; and $\eta = \sqrt{n}$ gives good performance, at the optimal bandwidth, for all $n \geq 25$.

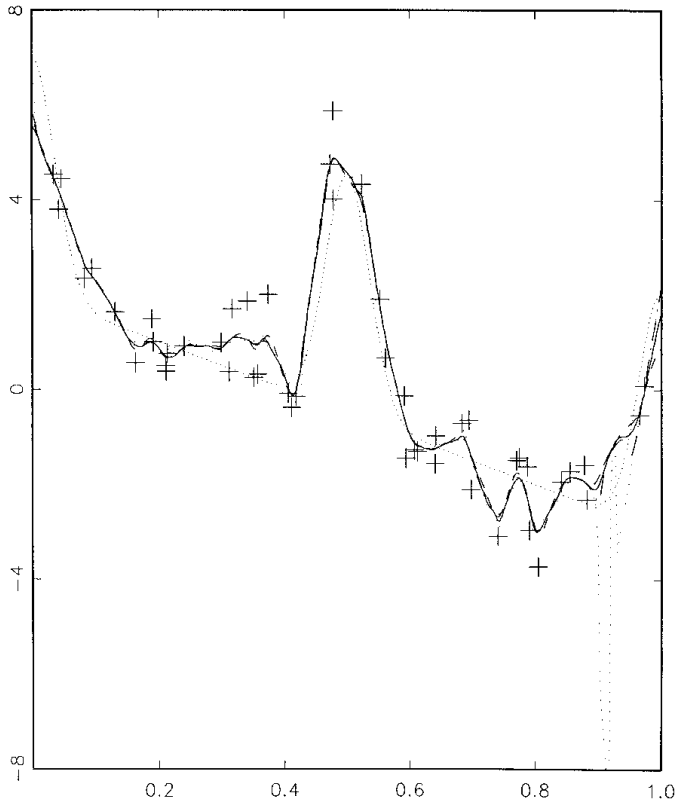

Fig. 4. *Typical estimates of the 'three peaks' regression mean defined at (6) (closely-spaced dotted line), constructed by naive local linear regression based on the biweight (dotted line) or Normal (dotted-and-dashed line) kernels and the suggested modifications with parameters $\eta = 50^{.20}$ (short-dashed line) or $\eta = 50^{.5}$ (solid line). The asymptotically optimal bandwidths were used*

Acknowledgment

We are grateful to Berwin Turlach for helpful discussions during the preparation of this paper.

References

Clark, R. M. (1977) Nonparametric estimation of a smooth regression function. *Journal of the Royal Statistical Society*, Series B **39**, 107–13.

Clark, R. M. (1980) Calibration, cross-validation and carbon-14, II. *Journal of the Royal Statistical Society*, Series A **143**, 177–94.

Cleveland, W. S. (1979) Robust locally weighted regression and smoothing scatterplots. *Journal of the American Statistical Association*, **74**, 829–36.

Cleveland, W. S. (1993) *Visualizing Data*. Hobart Press, Summit, N.J.

Cleveland, W. S. and Devlin, S. J. (1988) Locally weighted regression: An approach to regression analysis by local fitting. *Journal of the American Statistical Association*, **83**, 596–610

Cleveland, W. S. and Grosse, E. H. (1991) Computational methods for local regression. *Statistics and Computing*, **1**, 47–62

Fan, J. (1993) Local linear regression smoothers and their minimax efficiencies. *Annals of Statistics*, **21**, 196–216.

Gasser, Th. and Müller, H.-J. (1979) Kernel estimation of regression functions. In: *Smoothing Techniques for Curve Estimation*, eds Th. Gasser and M. Rosenblatt, pp. 23–68. *Lecture Notes in Mathematics*, **757**. Springer, Heidelberg.

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

Hall, P. and Heyde, C. C. (1980) *Martingale Limit Theory and its Application*. Academic, New York.

Hall, P. and Marron, J. S. (1995) On the role of the ridge parameter in local linear smoothing. Manuscript.

Hall, P., Marron, J. S., Neumann, M. H. and Titterington, D. M. (1995) Curve estimation when the design density is low. *Annals of Statistics*, to appear.

Hall, P. and Turlach, B. A. (1995) Interpolation methods for adapting to sparse design in nonparametric regression. *Statistics Research Report SRR 021-95*, Centre for Mathematics and Its Applications, School of Mathematical Sciences, Australian National University.

Hastie, T. and Loader, C. (1993) Local regression: automatic kernel carpentry. *Statistical Science*, **8**, 120–43.

Jones, M. C. (1993) Simple boundary correction for kernel density estimation. *Statistics and Computing*, **3**, 135–46.

Marron, J. S. and Nolan, D. (1989) Canonical kernels for density estimation. *Statistics and Probability Letters*, **7**, 195–99.

Müller, H.-J. (1988) *Nonparametric Regression Analysis of Longitudinal Data. Lecture Notes in Statistics* **46**. Springer, New York.

Seifert, B. and Gasser, T. (1995) Variance properties of local polynomials and ensuing modifications (with discussion). *Computational Statistics*, to appear.

Seifert, B. and Gasser, T. (1996) Finite sample analysis of local polynomials: analysis and solutions. *Journal of the American Statistical Association*, **91**, 267–75.

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

Appendix: Outline proof of Theorem 2.1.

For the sake of brevity we treat only the case where σ is constant. Let χ denote the set of design points $X_1, \ldots, X_n$, and observe that

$$E(\hat{g} - g)^2 = V + B \qquad \text{(A.1)}$$

where $V = E\{\mathrm{var}(\hat{g}|\chi)\}$ and $B = E\{E(\hat{g}|\chi) - g\}^2$. We shall deal separately with these terms. Let C_1, $C_2, \ldots$ denote generic positive constants.

First we derive an asymptotic formula for V. Let $s_k^{(K)}$ and $s_k^{(L)}$ denote the versions of s_k for the kernels K and L, respectively. Define $t_k^{(M)} = \int u^k M(u)\mathrm{d}u$, $t = \int K^2$, $T_{xki}^{(M)} = \{(x - X_i)/h\}^k M\{(x - X_i)/h\}$ and $T_{xk}^{(M)} = \Sigma_{i=1}^n T_{xki}^{(M)} = h^{-k} s_k^{(M)}(x)$, for $M = K$ or L. Given $\epsilon > 0$, let $\mathcal{E}_{xk}(\epsilon)$ denote the event that

$$|(nh^{k+1})^{-1} s_k^{(M)}(x) - t_k^{(M)} f(x)| > \epsilon \quad \text{for } M = K \text{ or } L$$

and write $\tilde{\mathcal{E}}_{xk}$ for the complement of $\mathcal{E}_{xk}$. Under conditions (C_1) and (C_2), $E(T_{xk1}^{(M)}) = ht_k^{(M)} f(x) + o(h)$ and $E|T_{xk1}^{(M)}|^p =$

$O(h)$, uniformly in $x \in \mathcal{J}$, for $M = K$ or L, $0 \le k \le 2$ and each $p \ge 1$. From the last result we have, by Rosenthal's inequality (e.g. Hall and Heyde, 1980, p. 23), that $E|T_{xk}^{(M)} - ET_{xk}^{(M)}|^{2p} = O\{(nh)^p\}$ for all $p \ge 1$. Hence, by Markov's inequality, there exists a sequence $\epsilon_n \downarrow 0$ such that $P\{\mathcal{E}_{xk}(\epsilon_n^2)\} = O(n^{-\lambda})$ uniformly in $x \in \mathcal{J}$ and $0 \le k \le 2$, for all $\lambda > 0$.

Since $\sum w_i^{(M)} = s_0^{(M)} s_2^{(M)} - (s_1^{(M)})^2$ then if $\epsilon_n \to 0$ sufficiently slowly, there exists $n_1 \ge 1$ such that, if $n \ge n_1$ and the event

$$\tilde{\mathcal{E}}_x(\epsilon_n) = \bigcap_{0 \le k \le 2} \tilde{\mathcal{E}}_{xk}(\epsilon_n^2)$$

holds, then

$$\left| (n^2h^4)^{-1} \sum_{i=1}^n w_i^{(M)}(x) - t_0^{(M)} t_2^{(M)} f(x)^2 \right| \le \epsilon_n$$

uniformly in $x \in \mathcal{J}$, for $M = K$ or L. More simply

$$E\left\{ \sum_{i=1}^n (w_i^{(K)} + \eta w_i^{(L)})^2 \right\} \sim n^3 h^7 (t_2^{(K)})^2 t f^3$$

Write 'sup M' for the least upper bound to $M(x)$ over $-\infty < x < \infty$. Since

$$\sum_{i=1}^n (w_i^{(M)})^2 \le (\sup M) s_2 \{ s_0^{(M)} s_2^{(M)} - (s_1^{(M)})^2 \} \le C_1 n^3 \tag{A.2}$$

and

$$P\{\mathcal{E}_x(\epsilon_n)\} = O(n^{-\lambda}) \tag{A.3}$$

for all $\lambda > 0$, then

$$E\left[\sum_{i=1}^n \{ w_i^{(K)}(x) + \eta w_i^{(L)}(x) \}^2 I\{\mathcal{E}_x(\epsilon_n)\} \right] = O(n^{-\lambda})$$

Therefore

$$\begin{aligned} E\Bigg(& \left[\sum_{i=1}^n \{ w_i^{(K)}(x) + \eta w_i^{(L)}(x) \}^2 \right] \\ & \times \left[\sum_{i=1}^n \{ w_i^{(K)}(x) + \eta w_i^{(L)}(x) \} \right]^{-2} I\{\tilde{\mathcal{E}}_x(\epsilon_n)\} \Bigg) \\ & \sim \{ n^2 h^4 t_0^{(K)} t_2^{(K)} f(x)^2 \}^{-2} \{ n^3 h^7 (t_2^{(K)})^2 t f(x)^3 \} \\ & = (nh)^{-1} (t_0^{(K)})^{-2} t f(x)^{-1} \end{aligned}$$

uniformly in $x \in \mathcal{J}$. Furthermore

$$V = \sigma^2 E\left[\left\{ \sum_{i=1}^n (w_i^{(K)} + \eta w_i^{(L)})^2 \right\} \left\{ \sum_{i=1}^n (w_i^{(K)} + \eta w_i^{(L)}) \right\}^{-2} \right]$$

and so, if we prove that with

$$\begin{aligned} V_1(x) = E\Bigg(& \left[\sum_{i=1}^n \{ w_i^{(K)}(x) + \eta w_i^{(L)}(x) \}^2 \right] \\ & \times \left[\sum_{i=1}^n \{ w_i^{(K)}(x) + \eta w_i^{(L)}(x) \} \right]^{-2} I\{\mathcal{E}_x(\epsilon_n)\} \Bigg) \end{aligned}$$

we have

$$\sup_{\mathcal{J}} V_1 = O(n^{-\lambda}) \tag{A.4}$$

for all $\lambda > 0$, it will follow that, also uniformly in $x \in \mathcal{J}$

$$V \sim (nh)^{-1} \sigma^2 (t_0^{(K)})^{-2} t f^{-1} \tag{A.5}$$

In view of (A.2), (A.3) and the fact that $s_0^{(K)} s_2^{(K)} - (s_1^{(K)})^2 \ge 0$ and $\sup_{\mathcal{J}} s_2^{(M)} \le C_2 n$, (A.4) will follow via the Cauchy–Schwartz inequality if we prove that

$$E[\{ s_0^{(L)} s_2^{(L)} - (s_1^{(L)})^2 \}^{-2}] = O(n^{C_3}) \tag{A.6}$$

Put $Z_i = L\{(x - X_i)/h\}$ and define $\xi = \inf_{|x| \le 1} L(x/h)$. In this notation

$$\begin{aligned} s_0^{(L)} s_2^{(L)} - (s_1^{(L)})^2 &= \sum_{i \le i < j \le n} \sum (X_i - X_j)^2 Z_i Z_j \\ &\ge \xi^2 \sum_{1 \le i < j \le n} \sum (X_i - X_j)^2 \end{aligned} \tag{A.7}$$

Conditions (C_1) and (C_2) imply that $\xi^{-1} = O(n^{\zeta})$ for some $\zeta > 0$, and that

$$\sum_{1 \le i < j \le n} \sum (X_i - X_j)^2 \ge C_4 \sum_{1 \le i < j \le n} \sum (U_{(i)} - U_{(j)})^2 \tag{A.8}$$

where $U_{(1)} \le \cdots \le U_{(n)}$ are the order statistics of a sequence of independent random variables uniformly distributed on $\mathcal{I}$. Properties of spacings of order statistics may be used to prove that

$$E\left\{ \sum_{1 \le i < j \le n} \sum (U_{(i)} - U_{(j)})^2 \right\}^{-2} = O(n^4) \tag{A.9}$$

Result (A.6) follows on combining (A.7)–(A.9).

We conclude the proof of Theorem 2.1 by deriving the analogue of (A.5) for B. Observe that

$$\begin{aligned} E\{\hat{g}(x)|\chi\} - g(x) = & \left[\sum_{i=1}^n \{ w_i^{(K)}(x) + \eta w_i^{(L)}(x) \} \{ g(x_i) - g(x) \} \right] \\ & \times \left[\sum_{i=1}^n \{ w_i^{(K)}(x) + \eta w_i^{(L)}(x) \} \right]^{-1} \\ = & \frac{1}{2} \left[\sum_{i=1}^n \{ w_i^{(K)}(x) + \eta w_i^{(L)}(x) \} (x_i - x)^2 g_2(x, x_i) \right] \\ & \times \left[\sum_{i=1}^n \{ w_i^{(K)}(x) + \eta w_i^{(L)}(x) \} \right]^{-1} \end{aligned}$$

where $g_2(x, y) = 2\{g(y) - g(x) - (y - x)g'(x)\}/(y - x)^2 \approx g''(x)$. From this result, using a modified form of the argument employed to derive (A.5), we may show that

$$B(x) = \{ \tfrac{1}{2} h^2 (t_0^{(K)})^{-1} t_2^{(K)} g''(x) \}^2 + o(h^4)$$

uniformly in $x \in \mathcal{J}$, as $n \to \infty$. The theorem follows from this formula and (A.5).