An Effective Bandwidth Selector for Local Least Squares Regression

D. RUPPERT, S. J. SHEATHER, and M. P. WAND

Local least squares kernel regression provides an appealing solution to the nonparametric regression, or "scatterplot smoothing," problem, as demonstrated by Fan, for example. The practical implementation of any scatterplot smoother is greatly enhanced by the availability of a reliable rule for automatic selection of the smoothing parameter. In this article we apply the ideas of plug-in bandwidth selection to develop strategies for choosing the smoothing parameter of local linear squares kernel estimators. Our results are applicable to odd-degree local polynomial fits and can be extended to other settings, such as derivative estimation and multiple nonparametric regression. An implementation in the important case of local linear fits with univariate predictors is shown to perform well in practice. A by-product of our work is the development of a class of nonparametric variance estimators, based on local least squares ideas, and plug-in rules for their implementation.

KEY WORDS: Boundary effects; Kernel estimator; Local polynomial fitting; Nonparametric regression; Pilot estimation; Variance estimation

1. INTRODUCTION

Local least squares kernel regression has recently gained widespread acceptance as an attractive method for nonparametric estimation of the mean function from noisy regression data. The advantages of this approach include simplicity in terms of interpretability and mathematical analysis, ease of fast computation, and superior boundary behavior. (For recent contributions to the theory and computation of local least squares, kernel regression estimators see Fan 1992, 1993; Fan and Marron 1994; and Ruppert and Wand 1994.)

As with any nonparametric regression procedure, an important choice to be made is the amount of local averaging performed to obtain the regression estimate. For a kernel-type estimator, this is controlled by a parameter usually referred to as the bandwidth. When a single bandwidth is used for the entire range of the data, it is often called a *global* bandwidth. Bias considerations favor the choice of a relatively small bandwidth, whereas variance considerations favor the choice of a larger bandwidth. Pictorially, bandwidths that are too small produce estimates that are too wiggly, tending toward interpolation of the data, and bandwidths that are too large smooth out features in the true mean function. It is very useful for the analyst to have a data-driven bandwidth selector that estimates the correct amount of smoothing.

The goal of this article is to develop a reliable global bandwidth selector rule for local least squares regression. In related kernel estimation settings, such as density estimation, there recently has been considerable research devoted to the bandwidth selection problem. One of the main findings of this research is that traditional smoothing parameter selection rules, such as those based on cross-validation, exhibit very inferior asymptotic and practical performance (e.g. Härdle, Hall, and Marron 1988). On the other hand, "plug-in" bandwidth selection rules, which involve estimation of unknown functionals that appear in formulas for the asymptotically optimal bandwidth, have been shown to perform more reliably, both theoretically and in practice. (Recent references for global plug-in bandwidth selection are Chiu 1991, 1992; Gasser, Kneip, and Köhler 1991; Hall, Sheather, Jones, and Marron 1991; Jones, Marron, and Park 1991; Park and Marron 1990; Sheather 1992; and Sheather and Jones 1991.) Fan and Gijbels (1995) recently proposed a bandwidth selector for local least squares regression that combines both plug-in and cross-validation notions. These authors have also developed another approach to the problem based on an estimate of the average mean squared error (MSE) (Fan and Gijbels 1993).

In this article we demonstrate how one can apply the plug-in ideas to obtain an effective bandwidth selector for local least squares kernel regression. We develop and compare three plug-in bandwidth selectors for local linear regression, each of which is an adaptation of an existing plug-in bandwidth selector for kernel density estimation. The most sophisticated of our proposals is an adaptation of the "solve-the-equation" rule of Sheather and Jones (1991), which has been seen to perform quite well in simulation studies (see, for example, Jones, Marron, and Sheather 1992 and Park and Turlach 1992). The other two are simple direct plug-in rules based on zero and one functional estimation. Initial estimates for our plug-in procedures are a variant of the "blocking method" developed by Härdle and Marron (1993), with the number of blocks chosen by Mallows's C_p (Mallows 1973). The simplest of our proposals is, therefore, a variant of the selectors of Härdle and Marron (1993). Each of our proposals is seen to perform well in a small simulation study, especially the two more sophisticated rules. These two rules are also shown to have good theoretical properties and are simple to implement using fast-binning algorithms.

D. Ruppert is Professor, School of Operations Research and Industrial Engineering, Cornell University, Ithaca, NY 14853. S. J. Sheather is Associate Professor and M. P. Wand is Senior Lecturer, Australian Graduate School of Management, University of New South Wales, Kensington, NSW 2033, Australia. This research was partially supported by National Science Foundation Grants DMS-9002791 and DMS-9306196, National Institute of General Medical Sciences Grant GM-39015, and a grant from the Australian Research Council. The authors are grateful to two referee's and an associate editor for their helpful comments.

It is straightforward, at least in theory, to extend our ideas to more sophisticated settings, such as higher-degree polynomial fitting, derivative estimation, multiple nonparametric regression, and heteroscedastic models, although there are still some important practical issues that need to be investigated carefully. Another direction in which our proposals can be extended is toward local bandwidth selection, where different bandwidths are used depending on location. Once again this extension is straightforward, at least in principle. Further discussion on the extension to heteroscedastic models and local bandwidth selection is given in Section 6.

A noteworthy by-product of this research is the development of a local least squares estimate of the variance of the errors under the assumption of homoscedasticity. For this project, the motivation comes from the fact that the variance appears in the optimal bandwidth formulas. Nonetheless, our proposal, which can be viewed as an extension of the variance estimator of Hall and Marron (1990), is of interest in its own right and worthy of further study.

In Section 2 we present the relevant theory of local least squares kernel estimators, and in Section 3 give theory for the estimation of functionals that arise in the formula for the asymptotically optimal bandwidth. We present analogous theory for variance estimation in Section 4. We describe plug-in bandwidth selectors that follow from the theory of these three sections are described in Section 5 and discuss their fast computation in Section 6. In Section 7 we describe the theoretical performance of our plug-in bandwidth selectors, and in Section 8 present the results of a simulation study. We give conclusions in Section 9.

2. LOCAL LEAST SQUARES KERNEL REGRESSION

Let $(X_1, Y_1), \ldots, (X_n, Y_n)$ be a set of independent and identically distributed random pairs where the Y_i's are scalar response variables and the X_i's are predictor variables having common density f with support confined to a compact set $\mathcal{S} \subset \mathbb{R}$. We wish to estimate the conditional mean function $m(x) = E(Y|X = x)$ at each $x \in \mathcal{S}$. We can also express this problem in terms of the model

$$Y_i = m(X_i) + v(X_i)^{1/2}\varepsilon_i, \qquad i = 1, \ldots, n,$$

where $v(x) = \operatorname{var}(Y|X = x)$ is finite and the ε_i's are mutually independent and identically distributed random variables having zero mean, unit variance, and finite fourth moment. It is also assumed that the ε_i's are independent of the X_i's.

The local degree p least squares kernel estimator of $m(x)$ is given by $\hat{m}(x; h, p) = \hat{\beta}_0$, where

$$(\hat{\beta}_0, \ldots, \hat{\beta}_p)^T = \operatorname{argmin}_{\boldsymbol{\beta}} \sum_{i=1}^{n} \{Y_i - \beta_0 - \cdots - \beta_p(X_i - x)^p\}^2 \times K\{(X_i - x)/h\}/h \quad (1)$$

and $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_p)^T$. Here $h > 0$ is the bandwidth and the kernel K is a symmetric, compactly supported density. Standard weighted least squares theory leads to the explicit solution of (1) given by

$$\hat{m}(x; h, p) = \mathbf{e}_1^T(\mathbf{X}_{p,x}^T \mathbf{W}_x \mathbf{X}_{p,x})^{-1} \mathbf{X}_{p,x}^T \mathbf{W}_x \mathbf{Y},$$

$$\text{where} \quad \mathbf{X}_{p,x} = \begin{bmatrix} 1 & X_1 - x & \cdots & (X_1 - x)^p \\ \vdots & \vdots & \vdots & \vdots \\ 1 & X_n - x & \cdots & (X_n - x)^p \end{bmatrix},$$

$\mathbf{Y} = (Y_1, \ldots, Y_n)^T$, and $\mathbf{W}_x = \operatorname{diag}[K\{(X_1 - x)/h\}/h, \ldots, K_h\{(X_n - x)/h\}/h]$. Here and throughout, we let $\mathbf{e}_j$ denote a column vector having 1 as its jth entry and all other entries equal to zero. The length of $\mathbf{e}_j$ will be clear from the context.

To derive an asymptotically optimal bandwidth, we require an appropriate global loss criterion. The most convenient one is the conditional weighted mean integrated squared error (MISE) of $\hat{m}(\cdot; h, p)$, given by

$$\operatorname{MISE}\{\hat{m}(\cdot; h, p)|X_1, \ldots, X_n\} = E\left[\int \{\hat{m}(x; b) - m(x)\}^2 f(x)\, dx | X_1, \ldots, X_n\right].$$

Useful notations for a function L are $\mu_l(L) = \int u^l L(u)\, du$ and $R(L) = \int L(u)^2\, du$, assuming that the integrals converge. Also, $\int$ will be taken to mean integration over the entire real line. It follows from theorem 4.1 of Ruppert and Wand (1994) that for p odd,

$$\begin{aligned} &\operatorname{MISE}\{\hat{m}(x; h, p)|X_1, \ldots, X_n\} \\ &\quad = n^{-1}h^{-1}R(K_p)\int_{\mathcal{S}} v(x)\, dx \\ &\qquad + h^{2p+2}\{\mu_{p+1}(K)/(p+1)!\}^2 \int m^{(p+1)}(x)^2 f(x)\, dx \\ &\qquad + o_P(n^{-1}h^{-1} + h^{2p+2}), \end{aligned}$$

where K_p is a $(p+1)$th-order kernel (defined in the next section). The case where p is even leads to a more complicated approximation for $\operatorname{MISE}\{\hat{m}(x; h, p)\}$, so we consider only the case where p is odd. Thus the MISE-optimal bandwidth has the asymptotic approximation

$$h_{\text{MISE}} \simeq \left[\frac{(p+1)(p!)^2 R(K_p)\int_{\mathcal{S}} v(x)\, dx}{2\mu_{p+1}(K_p)^2 \int_{\mathcal{S}} m^{(p+1)}(x)^2 f(x)\, dx\, n}\right]^{1/(2p+3)} \quad (2)$$

Plug-in bandwidth strategies rely on replacing the unknown integrals in this approximation for h_{MISE} by estimators. The main unknown in this expression is the regression functional $\int m^{(p+1)}(x)^2 f(x)\, dx$. The next section is devoted to the kernel estimation of functionals of this type.

3. KERNEL ESTIMATION OF REGRESSION FUNCTIONALS

In this section we investigate kernel estimation of regression functionals of the form

$$\theta_{rs} = \int m^{(r)}(x) m^{(s)}(x) f(x)\, dx \qquad r, s \geq 0, \qquad r + s \text{ even},$$

because versions of such functionals appear in expressions for the optimal bandwidth for local least squares regression

(see Sec. 2). In addition, the optimal bandwidths for kernel estimation of regression functionals involve other regression functionals, which motivates us to study estimation of θ_{rs} for general r and s.

The natural kernel-type estimator of θ_{rs} is

$$\hat{\theta}_{rs}(g) = n^{-1}\sum_{i=1}^{n} \hat{m}_r(X_i; g)\hat{m}_s(X_i; g),$$

where

$$\hat{m}_r(x; g) = r!\mathbf{e}_{r+1}^T(\mathbf{X}_{p,z}^T\mathbf{W}_x\mathbf{X}_{p,x})^{-1}\mathbf{X}_{p,x}^T\mathbf{W}_x\mathbf{Y}$$

and $\mathbf{W}_x = \text{diag}[K\{(X_1 - x)/g\}/g, \ldots, K\{(X_n - x)/g\}/g]$ is based on the bandwidth $g > 0$. For simplicity, in this section we assume that p is an integer greater than r and s such that $p - r$ and $p - s$ are both odd.

Let $K_{r,p}(u) = r!\{|\mathbf{M}_{r,p}(u)|/|\mathbf{N}_p|\}K(u)$, where $\mathbf{N}_p$ is the $(p + 1) \times (p + 1)$ matrix having (i, j) entry equal to $\int u^{i+j-2}K(u)\,du$ and $\mathbf{M}_{r,p}(u)$ is the same as $\mathbf{N}_p$, except that the $(r + 1)$th row is replaced by $(1, u, \ldots, u^p)$. The kernel K_p is defined to be $K_{0,p}$. Finally, let $(L_1 * L_2)(x) = \int L_1(u)L_2(x - u)\,du$ denote the convolution of two real-valued functions L_1 and L_2. Assuming that m has $r + s$ continuous derivatives, and that $g \to 0$ and $ng^{r+s+1} \to \infty$ as $n \to \infty$, we have

$$\begin{aligned} &E\{\hat{\theta}_{rs}(g) - \theta_{rs}|X_1, \ldots, X_n\} \\ &\quad \simeq \frac{1}{(p+1)!}\{\mu_{p+1}(K_{s,p})g^{p-s+1}\theta_{r,p+1} \\ &\qquad\qquad + \mu_{p+1}(K_{r,p})g^{p-r+1}\theta_{s,p+1}\} \\ &\qquad + n^{-1}g^{-r-s-1}\left\{\int K_{r,p}(u)K_{s,p}(u)\,du\right\}\int_{\mathcal{S}} v(x)\,dx \end{aligned} \tag{3}$$

and

$$\begin{aligned} &\text{var}\{\hat{\theta}_{rs}(g)|X_1, \ldots, X_n\} \\ &\quad \simeq 2n^{-2}g^{-2r-2s-1}R(K_{r,p} * K_{s,p}) \\ &\qquad \times \int_{\mathcal{S}} v(x)\{v(x) + 2m(x)^2\}\,dx \\ &\qquad + 4n^{-1}\int (mf)^{(r+s)}(x)^2 v(x)f(x)^{-1}\,dx. \end{aligned} \tag{4}$$

A derivation of these results is given in the Appendix.

We now restrict attention to the important special case $r = s = 2$ and $p = 3$, because it gives rise to bandwidth selection rules for the local linear kernel estimator. For simplicity, we also focus on the case where the errors are homoscedastic, which corresponds to having $v(x) = \sigma^2$ for all x. Also, for simplicity, we take $\mathcal{S} = [a, b]$. Finally, we assume that K is a kernel for which $\mu_4(K_{2,3}) > 0$. This condition is satisfied by virtually all kernels used in practice. Combining the foregoing bias and variance expressions, we obtain the conditional MSE approximation,

$$\begin{aligned} &\text{MSE}\{\hat{\theta}_{22}(g)|X_1, \ldots, X_n\} \\ &\quad \simeq \left[\frac{1}{12}\mu_4(K_{2,3})g^2\theta_{24} + n^{-1}g^{-5}R(K_{2,3})\sigma^2(b - a)\right]^2 \\ &\qquad + 2\sigma^2 n^{-2}g^{-9}R(K_{2,3} * K_{2,3}) \\ &\qquad \times \int_a^b \{\sigma^2 + 2m(x)^2\}\,dx + 4\sigma^2 n^{-1} \\ &\qquad \times \int_a^b (mf)^{(4)}(x)f(x)^{-1}\,dx. \end{aligned}$$

The MSE-optimal bandwidth is found by minimizing the squared bias term to obtain

$$g_{\text{MSE}} \simeq C_2(K)\left[\frac{\sigma^2(b - a)}{|\theta_{24}|n}\right]^{1/7}, \tag{5}$$

where

$$\begin{aligned} C_2(K) &= C_2^I(K) \qquad \theta_{24} < 0, \\ &= C_2^{II}(K) \qquad \theta_{24} > 0, \end{aligned}$$

and

$$\begin{aligned} C_2^I(K) &= \left[\frac{12R(K_{2,3})}{\mu_4(K_{2,3})}\right]^{1/7}, \\ C_2^{II}(K) &= \left[\frac{30R(K_{2,3})}{\mu_4(K_{2,3})}\right]^{1/7} \end{aligned} \tag{6}$$

The optimal bandwidth for estimation of θ_{22} is, therefore, of order $n^{-1/7}$. The corresponding minimum MSE is of order $n^{-5/7}$ when $\theta_{24} < 0$ and $n^{-4/7}$ when $\theta_{24} > 0$.

Table 1. Kernel-Dependent Constants

Kernel	Epanechnikov	Biweight	Normal
$C_1(K)$	$15^{1/5}$	$35^{1/5}$	$\{1/(2\sqrt{\pi})\}^{1/5}$
$C_2^I(K)$	$315^{1/7}$	$(8505/13)^{1/7}$	$\{3/(8\sqrt{\pi})\}^{1/7}$
$C_2^{II}(K)$	$(1575/2)^{1/7}$	$(42525/26)^{1/7}$	$\{15/(16\sqrt{\pi})\}^{1/7}$
$C_4^I(K)$	$21^{1/7}$	$(243/13)^{1/7}$	$(3/4)^{1/7}$
$C_4^{II}(K)$	$(105/2)^{1/7}$	$(1215/26)^{1/7}$	$(15/8)^{1/7}$

4. LOCAL LEAST SQUARES VARIANCE ESTIMATION

The other unknown in the optimal bandwidth formula (2) (under homoscedasticity) is the variance σ^2. There is a relatively extensive literature devoted to estimation of σ^2 in the homoscedastic nonparametric regression context. For local least squares kernel regression, a natural approach is to extend the variance estimator of Hall and Marron (1990), which is based on a zero-degree least squares fit, to the general pth degree setting. The motivation of such estimators comes from the fact that the residual sum of squares satisfies

$$E\left[\sum_{i=1}^{n}\{Y_i - \hat{m}(X_i; \lambda, p)\}^2|X_1, \ldots, X_n\right] = \nu\sigma^2$$

whenever m is a polynomial of degree less than or equal to p. Here $\nu = n - 2\sum_i w_{ii} + \sum\sum_{ij} w_{ij}^2$, where

$$w_{ij} = \mathbf{e}_1^T(\mathbf{X}_{p,X_i}^T\mathbf{W}_{X_i}\mathbf{X}_{p,X_i})^{-1}\mathbf{X}_{p,X_i}^T\mathbf{W}_{X_i}\mathbf{e}_j$$

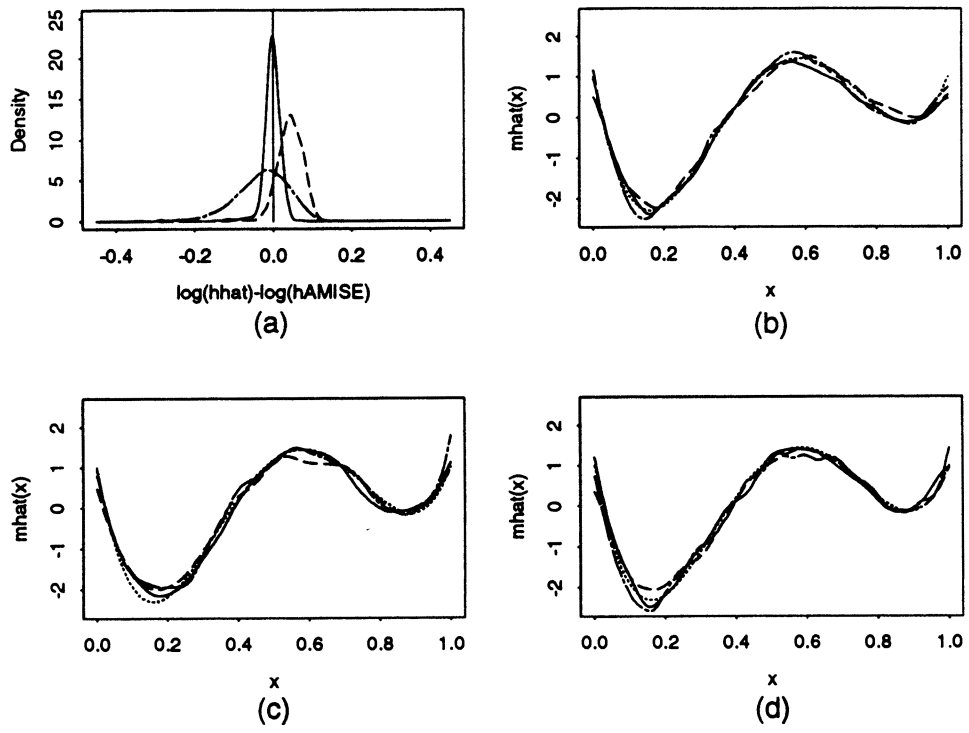

Figure 1. Graphical Summary of Simulation Results for Function 1 With n = 100. (a) Kernel density estimates based on log($\hat{h}$) − log(h_{AMISE}) values. Solid curve is for $\hat{h}_{ROT}$, dashed curve is for $\hat{h}_{DPI}$, and dot-dashed curve is for $\hat{h}_{STE}$. (b)–(d) Regression estimates for (b) ROT, (c) DPI, and (d) STE, based on bandwidth and sample near the median of the $\hat{h}$'s (solid curve), 10th percentile (dot-dashed curve), and 90th percentile (dashed curve). The dotted curve is the true regression function.

and $\mathbf{W}_x$ is based on the bandwidth $\lambda > 0$. This leads to the pth degree variance estimator,

$$\hat{\sigma}_p^2(\lambda) = \nu^{-1} \sum_{i=1}^{n} \{Y_i - \hat{m}(X_i; \lambda, p)\}^2.$$

This estimator is similar in nature to the local variance estimator proposed by Fan and Gijbels (1993, 1995).

Use of $\hat{\sigma}_p^2(\lambda)$ in practice requires selection of λ. To obtain a rule for the choice of λ, we appeal to results for the conditional MSE of $\hat{\sigma}_p^2(\lambda)$ when p is odd. Assume that m has p continuous derivatives and that f is compactly supported on $[a, b]$. Then, if $\lambda \to 0$ and $n\lambda \to \infty$ as $n \to \infty$, we have

$$E\{\hat{\sigma}_p^2(\lambda) - \sigma^2 | X_1, \ldots, X_n\} \simeq \lambda^{2p+2} \{\mu_{p+1}(K)/(p+1)!\}^2 \theta_{p+1,p+1} \quad (7)$$

and

$$\operatorname{var}\{\hat{\sigma}_p^2(\lambda) | X_1, \ldots, X_n\} \simeq n^{-1}\sigma^4 \operatorname{var}(\varepsilon^2) + 2(n^2\lambda)^{-1}\sigma^4(b-a)R(K_p * K_p - 2K_p). \quad (8)$$

The derivation of these results is outlined in the Appendix. It follows that the asymptotically MSE (AMSE)-optimal choice of λ is

$$\lambda_{\mathrm{AMSE}} = C_3(K) \left[\frac{\sigma^4(b-a)}{\theta_{p+1,p+1}^2 n^2} \right]^{1/(4p+5)} \quad (9)$$

where

$$C_3(K) = \left[\frac{\{(p+1)!\}^4 R(K_p * K_p - 2K_p)}{2(p+1)\mu_{p+1}(K_p)^4} \right]^{1/(4p+5)}$$

If $K(x) = (2\pi)^{-1/2} e^{-x^2/2}$ is the normal kernel, then it can be shown that

$$C_3(K) = \left\{ 4 \left(\frac{1}{2} + 2\sqrt{2} - \frac{4}{3}\sqrt{3} \right) \Big/ \sqrt{2\pi} \right\}^{1/9}$$

We believe that this is the first theoretical investigation into the properties of local polynomial variance estimators, apart from the zero-degree results of Hall and Marron (1990). Those authors showed that $\hat{\sigma}_0^2(\lambda)$ has certain minimax optimality properties. We speculate that in light of this finding and the optimality results for local polynomial fitting derived by Fan (1992, 1993) for estimation of the mean, higher-degree versions of $\hat{\sigma}_p^2(\lambda)$ can also be shown to possess optimality properties. Furthermore, preliminary investigations have lead us to believe that higher-degree versions of $\hat{\sigma}_p^2(\lambda)$ have substantial practical potential.

5. PLUG-IN BANDWIDTH SELECTION STRATEGIES

In this article our main goal is to obtain a ready-to-use plug-in bandwidth selector for local linear kernel estimation, under the assumption of homoscedastic errors. Extension to more complicated settings is possible using the theory of the previous three sections, but the practicalities are not addressed here.

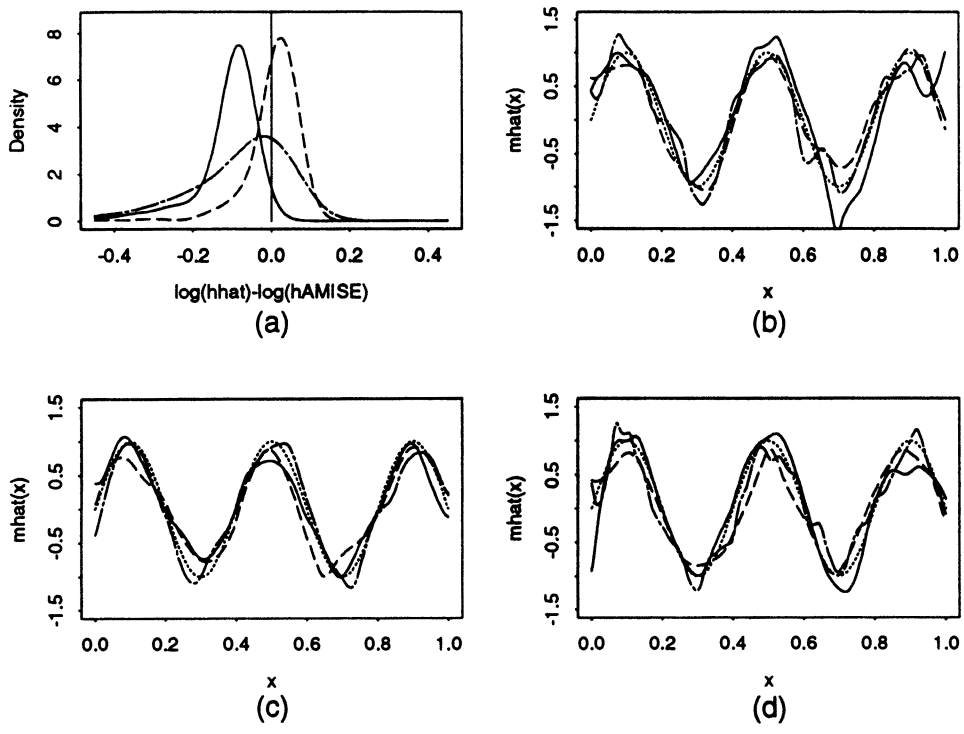

Figure 2. Graphical Summary of Simulation Results for Function 2 With n = 100. (a) Kernel density estimates based on log(ĥ) − log(h_{AMISE}) values. Solid curve is for $\hat{h}_{ROT}$, dashed curve is for $\hat{h}_{DPI}$, and dot-dashed curve is for $\hat{h}_{STE}$. (b)–(d) Regression estimates for (b) ROT, (c) DPI, and (d) STE, based on bandwidth and sample near the median of the ĥ's (solid curve), 10th percentile (dot-dashed curve), and 90th percentile (dashed curve). The dotted curve is the true regression function.

For the local linear kernel estimator, the MISE-optimal bandwidth is asymptotic to

$$h_{\mathrm{AMISE}} = C_1(K)\left[\frac{\sigma^2(b-a)}{\theta_{22}n}\right]^{1/5},$$

where $C_1(K) = [R(K)/\mu_2(K)^2]^{1/5}$, whereas the MSE-optimal bandwidth for estimation of θ_{22} is asymptotic to

$$g_{\mathrm{AMSE}} = C_2(K)[\sigma^2(b-a)/(|\theta_{24}|n)]^{1/7}.$$

There are a number of ways for using the h_{AMISE} and g_{AMSE} expressions to derive bandwidth selection strategies. The simplest is that where the θ_{22} in h_{AMISE} is simply replaced by a "rule-of-thumb" estimate of m, such as one based on a parametric fit. Such a rule would work reasonably well when the true regression function is close to the parametric fit, but it has no consistency properties and can perform poorly away from the parametric model. More sophisticated selectors with good consistency properties can be obtained by replacing θ_{22} with the kernel estimator $\hat{\theta}_{22}(g)$. Of course, this leads to a new bandwidth selection problem, because the optimal bandwidth g_{AMSE} depends on θ_{24}. This functional could also be estimated by another kernel estimation step and this process continued indefinitely, but at some stage a rule-of-thumb estimate of m will be needed.

We experimented with several strategies for obtaining an initial estimate of m. Our first suggestion was using an ordinary least squares quartic fit, because this is the lowest-degree polynomial that admits nonzero estimates of θ_{24}. But this proved to be inadequate for regression functions having many oscillations so, following Härdle and Marron (1993), we instead partitioned the range of the X data into N blocks and fit a quartic for each block. The partition can be formed by either dividing the range into equally-sized blocks or by dividing the data into equal-sized subsamples. The second option, which we use in our simulations, has the advantage of adapting better to nonuniform designs and decreasing the chance of overfitting. Let N be the number of subsamples and let $\mathcal{X}_j$ denote the jth subsample of the ordered X_i's. If N divides n and $t = n/N$, then $\mathcal{X}_j = \{X_{((j-1)t+1)}, \ldots, X_{(jt)}\}$. (If N does not divide n, then allocations to subsamples need to be slightly adjusted.) Let $\hat{m}_j^Q(x)$ be the least squares quartic fit obtained from data having X_i values in $\mathcal{X}_j$. For $\max(r,s) \le 4$, the "blocked quartic estimator" for θ_{rs} is

$$\hat{\theta}_{rs}^Q(N) = n^{-1}\sum_{i=1}^{n}\sum_{j=1}^{N}(\hat{m}_j^Q)^{(r)}(X_i)(\hat{m}_j^Q)^{(s)}(X_i)1_{\{X_i\in\mathcal{X}_j\}}.$$

Similarly, the blocked quartic estimator for σ^2 is

$$\hat{\sigma}_Q^2(N) = (n-5N)^{-1}\sum_{i=1}^{n}\sum_{j=1}^{N}\{Y_i - \hat{m}_j^Q(X_i)\}^2 1_{\{X_i\in\mathcal{X}_j\}}.$$

These estimators require a rule for choosing N. We found that Mallows's C_p (Mallows 1973) was a reasonable solution to this problem. For blocked quartic fits, this involves

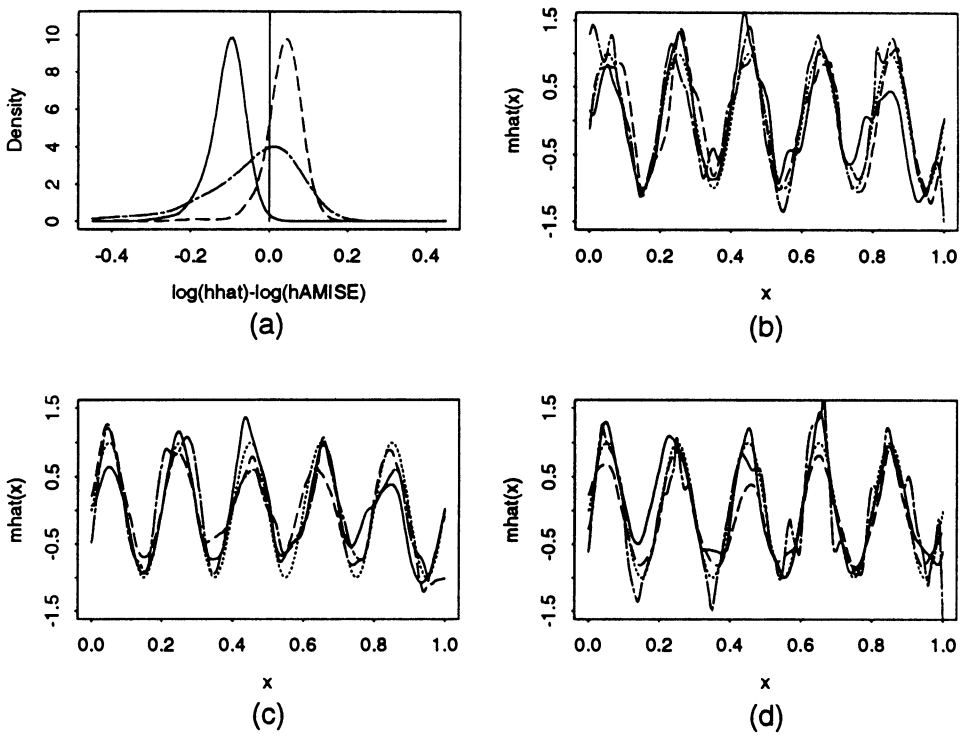

Figure 3. Graphical Summary of Simulation Results for Function 3 With n = 100. (a) Kernel density estimates based on log($\hat{h}$) − log(h_{AMISE}) values. Solid curve is for $\hat{h}_{ROT}$, dashed curve is for $\hat{h}_{DPI}$, and dot-dashed curve is for $\hat{h}_{STE}$. (b)–(d) Regression estimates for (b) ROT, (c) DPI, and (d) STE, based on bandwidth and sample near the median of the $\hat{h}$'s (solid curve), 10th percentile (dot-dashed curve), and 90th percentile (dashed curve). The dotted curve is the true regression function.

choosing $\hat{N}$ from the set $\{1, 2, \ldots, N_{\max}\}$ to minimize

$$C_p(N) = \mathrm{RSS}(N)/\{\mathrm{RSS}(N_{\max})/(n - 5N_{\max})\} - (n - 10N),$$

where $\mathrm{RSS}(N)$ is the residual sum of squares based on a blocked quartic fit over N blocks. This leaves us with the choice of $N_{\max}$. To reduce the chance of overfitting, it is sensible to take $N_{\max}$ to be of the form

$$N_{\max} = \max\{\min(\lfloor n/20 \rfloor, N^*), 1\} \tag{10}$$

for some positive integer N^*. For regression functions with few features, the choice of N^* is not very critical, and in this study we took $N^* = 5$. But if one feels that there are many oscillations in the regression function, then higher values of N^* could be considered.

Though the combination of blocked quartic fits and Mallows's C_p offers one simple and effective way of obtaining an initial estimate for m, many other possibilities could be used instead.

An optional adjustment to the kernel functional estimates is the truncation of data within $100\alpha\%$ of the boundaries, for some small value of α. The reason for this is that local polynomial kernel estimates of higher derivatives can be extremely variable near the boundary. This type of adjustment was also recommended by Gasser et al. (1991). In the case where the X_i's are supported on $[a, b]$, this involves replacement of $\hat{\theta}_{rs}(g)$ by

$$\hat{\theta}^{\alpha}_{rs}(g) = n^{-1} \sum_{i=1}^{n} m^{(r)}(X_i) m^{(s)}(X_i) 1_{\{(1-\alpha)a + \alpha b < X_i < \alpha a + (1-\alpha)b\}}.$$

Our experience has shown that taking $\alpha = .05$, say, tends to dramatically decrease the variability of the resulting bandwidth selector but can increase its bias when the true regression curve has many features near the boundary. For our rule $\hat{h}_{\mathrm{DPI}}$, given in the following paragraph, we use $\alpha = .05$. But for $\hat{h}_{\mathrm{STE}}$, we found that truncation led to severe bias problems, so we took $\alpha = 0$ for this rule. We are now in a position to give algorithms for three plug-in bandwidth selection strategies, in order of increasing level of sophistication:

- The rule-of-thumb bandwidth selector $\hat{h}_{\mathrm{ROT}}$
  1. Find $\hat{\theta}^Q_{22}(\hat{N})$ and $\hat{\sigma}^2_Q(\hat{N})$ based on a blocked quartic fit with $\hat{N}$ chosen by Mallows's C_p and $N_{\max}$ given by (10) with $N^* = 5$.
  2. The selected bandwidth is

$$\hat{h}_{\mathrm{ROT}} = C_1(K)\left[\frac{\hat{\sigma}^2_Q(\hat{N})(b-a)}{\hat{\theta}^Q_{22}(\hat{N})n}\right]^{1/5}$$

- The direct plug-in bandwidth selector $\hat{h}_{\mathrm{DPI}}$
  1. Find $\hat{\theta}^Q_{24}(\hat{N})$ and $\hat{\sigma}^2_Q(\hat{N})$ based on a blocked quartic fit with $\hat{N}$ chosen by Mallows's C_p and $N_{\max}$ given by (10) with $N^* = 5$.
  2. Estimate θ_{22} using $\hat{\theta}^{.05}_{22}(\hat{g}_{\mathrm{AMSE}})$, where

$$\hat{g}_{\mathrm{AMSE}} = C_2(K)\left[\frac{\hat{\sigma}^2_Q(\hat{N})(b-a)}{|\hat{\theta}^Q_{24}(\hat{N})|n}\right]^{1/7}$$

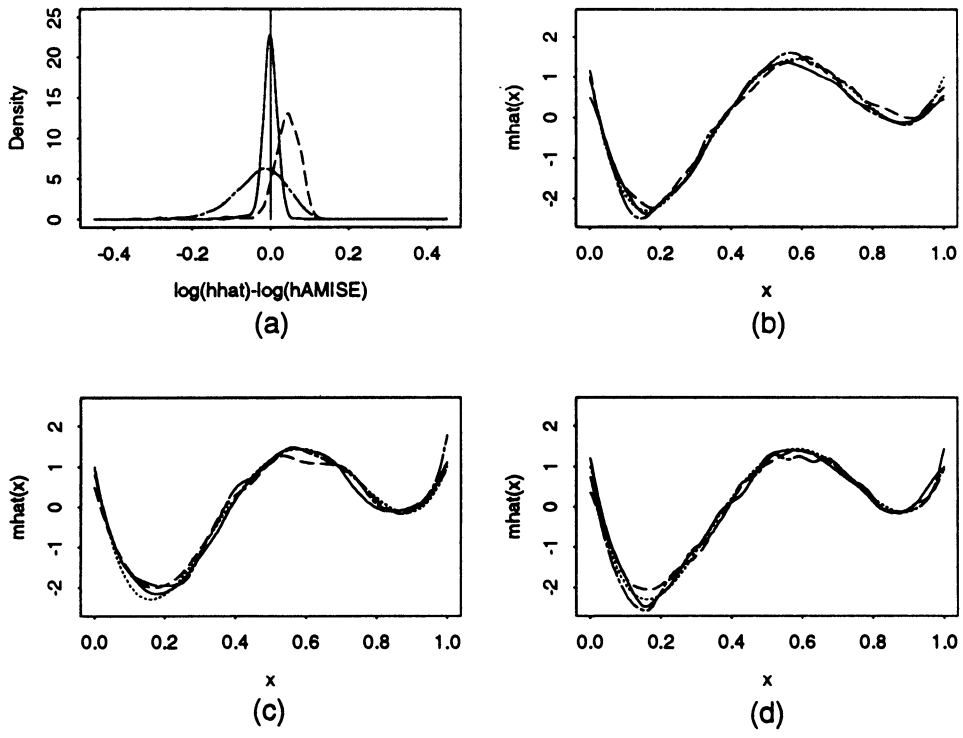

Figure 4. Graphical Summary of Simulation Results for Function 1 With n = 500. (a) Kernel density estimates based on log($\hat{h}$) − log(h_{AMISE}) values. Solid curve is for $\hat{h}_{ROT}$, dashed curve is for $\hat{h}_{DPI}$, and dot-dashed curve is for $\hat{h}_{STE}$. (b)–(d) Regression estimates for (b) ROT, (c) DPI, and (d) STE, based on bandwidth and sample near the median of the $\hat{h}$'s (solid curve), 10th percentile (dot-dashed curve), and 90th percentile (dashed curve). The dotted curve is the true regression function.

and estimate $\hat{\sigma}^2$ using $\hat{\sigma}_1^2(\hat{\lambda}_{\mathrm{AMSE}})$, where

$$\hat{\lambda}_{\mathrm{AMSE}} = C_3(K)\left[\frac{\hat{\sigma}_Q^4(\hat{N})(b-a)}{\hat{\theta}_{22}^{.05}(\hat{g}_{\mathrm{AMSE}})^2 n^2}\right]^{1/9}$$

3. The selected bandwidth is

$$\hat{h}_{\mathrm{DPI}} = C_1(K)\left[\frac{\hat{\sigma}_1^2(\hat{\lambda}_{\mathrm{AMSE}})(b-a)}{\hat{\theta}_{22}^{.05}(\hat{g}_{\mathrm{AMSE}})n}\right]^{1/5}$$

- The solve-the-equation bandwidth selector $\hat{h}_{\mathrm{STE}}$
  1. Find $\hat{\theta}_{22}^Q(\hat{N}), \hat{\theta}_{24}^Q(\hat{N})$, and $\hat{\sigma}_Q^2(\hat{N})$ based on a blocked quartic fit with $\hat{N}$ chosen by Mallows's C_p and $N_{\max}$ given by (10) with $N^* = 5$.
  2. Estimate σ^2 using $\hat{\sigma}_1^2(\hat{\lambda}_{\mathrm{AMSE}})$, where $\hat{\lambda}_{\mathrm{AMSE}}$ is defined in Step 2 of $\hat{h}_{\mathrm{DPI}}$. The value of $C_2(K)$ is taken to be one of the values in (6) according to the sign of $\hat{\theta}_{24}^Q(\hat{N})$.
  3. The selected bandwidth is $\hat{h}_{\mathrm{STE}}$, the solution to the equation

$$h = C_1(K)\left[\frac{\hat{\sigma}_1^2(\hat{\lambda}_{\mathrm{AMSE}})(b-a)}{\hat{\theta}_{22}(g(h))n}\right]^{1/5},$$

where

$$\begin{aligned} g(h) &= C_4(K)(\hat{\theta}_{22}^Q/|\hat{\theta}_{24}^Q|)^{1/7}h^{5/7}, \\ C_4(K) &= C_2(K)C_1(K)^{-5/7}, \end{aligned}$$

and $C_2(K)$ is taken to be one of the values in (6) according to the sign of $\hat{\theta}_{24}^Q(\hat{N})$.

The first two bandwidth selectors, $\hat{h}_{\mathrm{ROT}}$ and $\hat{h}_{\mathrm{DPI}}$, are based on straightforward plug-in ideas, using zero and one kernel functional estimation stage. The third selector, $\hat{h}_{\mathrm{STE}}$, is analogous to the solve-the-equation bandwidth selectors proposed by Park and Marron (1990) and Sheather and Jones (1991) in kernel density estimation and is based on the observation that

$$g_{\mathrm{AMSE}} = C_4(K)(\theta_{22}/|\theta_{24}|)^{1/7}h_{\mathrm{AMISE}}^{5/7}.$$

Table 1 gives values of the kernel-dependent constants required for the foregoing rules for some common kernels. Additional notation is $C_4^I(K) = C_2^I(K)C_1(K)^{-5/7}$ and $C_4^{II}(K) = C_2^{II}(K)C_1(K)^{-5/7}$. The Epanechnikov and Biweight kernels are given by $(3/4)(1-x^2)$ and $(15/16) \times (1-x^2)^2$ for $|x| \le 1$ and zero otherwise. The normal kernel is the standard normal density.

There are several directions in which the methodology presented in this section can be extended. Two worth further discussion are heteroscedastic models and local bandwidth selection.

If the homoscedasticity assumption seems inappropriate, then the $\sigma^2(b-a)$ appearing in the asymptotic formulas should be replaced by $\int_a^b v(x)\,dx$. In practice we would need to estimate this integral, the most obvious estimate being $\int_a^b \hat{v}(x;\lambda)$, where $\hat{v}(\cdot;\lambda)$ is the "method-of-moments" estimator for v based on local linear kernel estimators with bandwidth λ. Theory for the optimal choice of λ, analogous

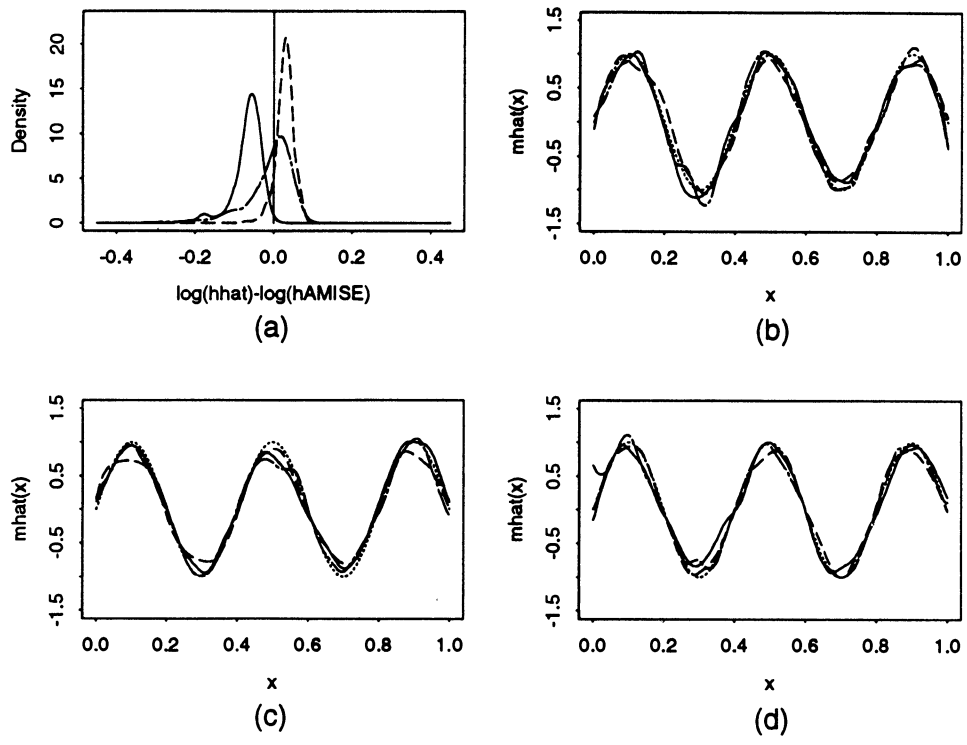

Figure 5. Graphical Summary of Simulation Results for Function 2 With n = 500. (a) Kernel density estimates based on log($\hat{h}$) − log(h_{AMISE}) values. Solid curve is for $\hat{h}_{ROT}$, dashed curve is for $\hat{h}_{DPI}$, and dot-dashed curve is for $\hat{h}_{STE}$. (b)–(d) Regression estimates for (b) ROT, (c) DPI, and (d) STE, based on bandwidth and sample near the median of the $\hat{h}$'s (solid curve), 10th percentile (dot-dashed curve), and 90th percentile (dashed curve). The dotted curve is the true regression function.

to (9), would need to be developed for this extension to be accomplished. This remains an open problem.

Local bandwidth selection rules can be derived by straightforward extension of the plug-in ideas described earlier. The simplest way is to partition the X space and apply the methodology to each subset of the scatterplot. In many cases, simple quartic fits should suffice as initial estimates rather than blocked quartic fits. This procedure gives a bandwidth for each partition, but the change in bandwidth between partitions can lead to roughness of the regression estimate. To overcome this problem, Fan and Gijbels (1995) suggested presmoothing the "bandwidth step-function" to give a smoothly changing set of local bandwidths. But choice of the most appropriate partition is a new problem introduced by this extension.

6. COMPUTATIONAL ISSUES

The local linear kernel estimator admits the explicit expression

$$\hat{m}(x;h,1) = \frac{\hat{s}_2(x;h)\hat{t}_0(x;h) - \hat{s}_1(x;h)\hat{t}_1(x;h)}{\hat{s}_0(x;h)\hat{s}_2(x;h) - \hat{s}_1(x;h)^2},$$

where

$$\hat{s}_k(x;h) = \sum_{i=1}^{n} (X_i - x)^k K_h(X_i - x)$$

and

$$\hat{t}_k(x;h) = \sum_{i=1}^{n} (X_i - x)^k K_h(X_i - x) Y_i.$$

Let $[a,b]$ be an interval containing each of the X_i. Fan and Marron (1994) described fast computation of $\hat{m}(x;h,1)$ over an equally spaced grid $a = g_1 < g_2 < \cdots < g_M = b$. The essential idea involves binning the (X_i, Y_i)'s to obtain grid counts $(c_1, d_1), \ldots, (c_M, d_M)$ that represent the contributions of the data at each grid point. There are several strategies for obtaining grid counts. One that has particularly good properties is "linear binning" (see Hall and Wand 1993), for which

$$c_l = \sum_{i=1}^{n} W_l(X_i) \quad \text{and} \quad d_l = \sum_{i=1}^{n} W_l(X_i) Y_i.$$

Here $W_l(x) = (1 - |x - g_l|/\delta)_+$ and $\delta = (g_M - g_1)/(M - 1)$. The approximation to $\hat{s}_k(g_j;h)$ is

$$\begin{aligned} \tilde{s}_k(g_j;h,M) &= \sum_{l=1}^{M} (g_l - g_j)^k K_h(g_l - g_j) c_l \\ &= \sum_{l=1-M}^{M-1} c_{j-l} \kappa_l^{(k)}, \qquad j = 1, \ldots, M, \end{aligned}$$

where $\kappa_l^{(k)} = (l\delta)^k K_h(l\delta)$. The formula for the $\tilde{t}_k(g_j;h)$ values is analogous. Binned computation of the $\hat{s}_k$ and $\hat{t}_k$ has the advantage of requiring only $O(M)$ kernel evaluations, as well as being the discrete convolution of the count vectors with the $\kappa_l^{(k)}$. This allows very fast computation of $\hat{m}$ over the grid points.

For kernel estimation of θ_{22}, a cubic polynomial fit is used to estimate m''. This estimate of m'' can be written

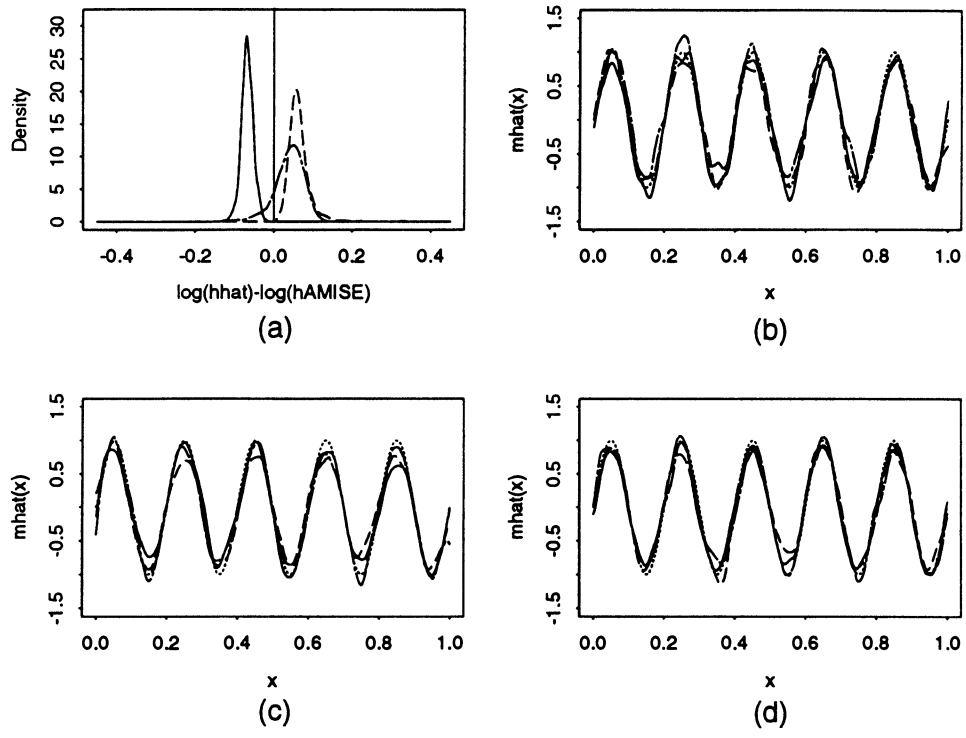

Figure 6. Graphical Summary of Simulation Results for Function 3 With n = 500. (a) Kernel density estimates based on log($\hat{h}$) − log(h_{AMISE}) values. Solid curve is for $\hat{h}_{ROT}$, dashed curve is for $\hat{h}_{DPI}$, and dot-dashed curve is for $\hat{h}_{STE}$. (b)–(d) Regression estimates for (b) ROT, (c) DPI, and (d) STE, based on bandwidth and sample near the median of the $\hat{h}$'s (solid curve), 10th percentile (dot-dashed curve), and 90th percentile (dashed curve). The dotted curve is the true regression function.

as the solution to a 4×4 linear system involving $\tilde{s}_k$'s and $\hat{t}_k$'s. Let $\tilde{m}_2(g_j; g, 3)$ be the estimate of $m''(g_j)$ obtained from binned approximations $\tilde{s}(g_j; g)$ and $\tilde{t}(g_j; g)$. Then the appropriate binned estimate of θ_{22} is

$$\tilde{\theta}_{22}(g, M) = n^{-1} \sum_{l=1}^{M} \tilde{m}_2(g_l; g, 3)^2 c_l.$$

Similar ideas can be used for fast implementation of $\hat{\sigma}_p^2(\lambda)$. For example, the residual sum of squares can be approximated by

$$\sum_{i=1}^{n} Y_i^2 - 2 \sum_{l=1}^{M} \tilde{m}(g_l; \lambda, p) d_l + \sum_{l=1}^{M} \tilde{m}(g_l; \lambda, p)^2 c_l.$$

The accuracy of the binned approximation increases as M is increased. Following the advice of Fan and Marron (1994), we use $M = 400$ in our examples. Theoretical support for this choice was provided by work of Hall and Wand (1993).

7. THEORETICAL PERFORMANCE

The rule-of-thumb bandwidth selector is based on an inconsistent estimator of θ_{22} and thus has no consistency properties itself. On the other hand, the kernel estimation of θ_{22} in $\hat{h}_{\mathrm{DPI}}$ and $\hat{h}_{\mathrm{STE}}$ implies that these selectors converge to the MISE-optimal bandwidth at a rate determined by the quality of the functional estimation. (The closeness of h_{AMISE} and h_{MISE} can also have an effect.)

To obtain the asymptotic behavior of the $\hat{h}_{\mathrm{DPI}}$, first note that $h_{\mathrm{MISE}} = h_{\mathrm{AMISE}} + O_P(n^{-3/5})$. This can be derived by a straightforward extension of the results of Fan (1992) and Ruppert and Wand (1994). Also, noting that $\hat{\sigma}_1^2(\hat{\lambda}_{\mathrm{AMSE}}) - \sigma^2 = O_P(n^{-1/2})$ and the formal approximation

$$\hat{\theta}_{22}(g)^{-1/5} - \theta_{22}^{-1/5} \simeq -\frac{1}{5} \theta_{22}^{-6/5} \{\hat{\theta}_{22}(g) - \theta_{22}\},$$

it follows that the relative error of $\hat{h}_{\mathrm{DPI}}$ is dominated by

$$(\hat{h}_{\mathrm{DPI}} - h_{\mathrm{MISE}})/h_{\mathrm{MISE}} \simeq -\frac{1}{5} \theta_{22}^{-1} \{\hat{\theta}_{22}(g) - \theta_{22}\}.$$

Let $g = Gn^{-1/7}$ for some constant $G > 0$. Then (5) entails that, conditional on $X_1, \ldots, X_n$,

$$n^{2/7}(\hat{h}_{\mathrm{DPI}} - h_{\mathrm{MISE}})/h_{\mathrm{MISE}} \to_P D, \tag{11}$$

where

$$D = -\frac{1}{5} \theta_{22}^{-1} \left\{ \frac{1}{2} \mu_4(K_{2,3}) \theta_{24} G^2 + \sigma^2 (b - a) R(K_{2,3}) G^{-5} \right\}.$$

A rigorous proof of (11) can be obtained using techniques similar to those used in the proof of theorem 1 of Park and Marron (1992). In terms of relative rate of convergence to h_{MISE}, (11) shows that $\hat{h}_{\mathrm{DPI}}$ is an $O_P(n^{-2/7})$ bandwidth selector for any choice of G. Observe that our choice of G is an attempt to make D equal to zero, although

Table 2. Numerical Summary of Simulation Study for n = 100

Function	1	2	3
h_{AMISE}	4.01e − 2	2.97e − 2	1.71e − 2
$\hat{h}_{ROT}$	3.89e − 2[a](5.48e − 3[b])	2.34e − 2(3.52e − 3)	1.34e − 2(1.30e − 3)
	3.96e − 2[c]	2.42e − 2	1.36e − 2
$\hat{h}_{DPI}$	4.33e − 2(6.29e − 3)	3.05e − 2(3.93e − 3)	1.87e − 2(1.85e − 3)
	4.36e − 2	3.10e − 2	1.88e − 2
$\hat{h}_{STE}$	3.49e − 2(8.87e − 3)	2.57e − 2(6.38e − 3)	1.60e − 3(3.61e − 3)
	3.56e − 2	2.66e − 2	1.66e − 2
ISE_{ROT}	1.14e − 1(1.53e − 1)	5.93e − 2(1.60e − 1)	6.05e − 1(10.9)
	8.70e − 2 [2]	4.22e − 2 [2.5]	8.46e − 2 [3]
ISE_{DPI}	1.18e − 1(3.12e − 1)	4.85e − 2(8.63e − 2)	1.67e − 1(1.50)
	8.40e − 2 [1]	3.78e − 2 [1]	6.98e − 2 [1]
ISE_{STE}	1.52e − 1(4.27e − 1)	7.52e − 2(4.06e − 1)	1.17e − 1(0.25)
	9.83e − 2 [3]	4.27e − 2 [2.5]	7.68e − 2 [2]

NOTE: Average [a] (standard deviation [b]), and median [c] of selected bandwidths and ISE for each strategy. The ISE ranking of each selector is shown in square brackets.

this goal is not achieved asymptotically, because $\hat{h}_{\mathrm{DPI}}$ uses rule-of-thumb estimates of σ^2 and θ_{24}. If θ_{24} is negative and is estimated by an $o_P(n^{-1/14})$ consistent estimator (Park and Marron 1992; Sheather and Jones 1991), then an $O_P(n^{-5/14})$ rule results. An $O_P(n^{-1/2})$ rule can be obtained by extending the h_{MISE} approximation to two terms and using higher-degree fits to estimate unknown regression functionals, analogous to work of Hall et al. (1991), or by using bandwidth factorization as suggested by Jones et al. (1991). But simulation studies in the density estimation context have indicated that these adjustments do not markedly improve the practical performance of the selector, so our current preference is to use the simple $O_P(n^{-2/7})$ version of $\hat{h}_{\mathrm{DPI}}$. Similar arguments can be used to establish that $\hat{h}_{\mathrm{STE}}$ is also an $O_P(n^{-2/7})$ bandwidth selector. Observe that the $O_P(n^{-2/7})$ relative rate is a big improvement on cross-validation selectors that have an $O_P(n^{-1/10})$ rate (Härdle et al. 1988).

8. PRACTICAL PERFORMANCE

We conducted a simulation study to evaluate and compare each of the bandwidth selectors described in Section 5. Such a study is necessarily restrictive, because there are many possibilities for the choice of regression function, design density, error density, sample size, and noise level. But we feel that our results are very useful for gauging the practical performance of the proposed bandwidth selectors.

For our study, the standard deviation of the errors was set to be $\sigma = \frac{1}{4}(\max m - \min m)$. The X_i's were generated from the uniform distribution on [0, 1]. Normal errors were used throughout. The example regression functions were

1. $m(x) = 1 - 48x + 218x^2 - 315x^3 + 145x^4$
2. $m(x) = \sin(5\pi x)$
3. $m(x) = \sin(10\pi x)$.

We used sample sizes $n = 100$ and $n = 500$. The number of replications in the simulation was 500.

Figures 1–6 give a graphical summary of the results. In each case the plot (a) shows kernel density estimates based on $\log(\hat{h}) - \log(h_{\mathrm{AMISE}})$ values for each rule. It would be more appropriate to replace h_{AMISE} by h_{MISE}, but this is not easily computable in random design regression contexts and depends on the particular realization of the design variables. The remaining plots show regression estimates based on $\hat{h}$ values that are near the 10th, 50th, and 90th percentiles of the $\hat{h}$ sample for each strategy—(b) $\hat{h}_{\mathrm{ROT}}$, (c) $\hat{h}_{\mathrm{DPI}}$, and (d) $\hat{h}_{\mathrm{STE}}$.

Taking the sample sizes and noise levels into account, all three rules appear to perform very well for the test functions considered in this study. But $\hat{h}_{\mathrm{ROT}}$ tends to undersmooth somewhat for the more wiggly functions, due to its heavier

Table 3. Numerical Summary of Simulation Study for n = 500

Function	1	2	3
h_{AMISE}	2.91e − 2	2.15e − 2	1.24e − 2
$\hat{h}_{ROT}$	2.89e − 2[a](1.68e − 3[b])	1.86e − 2(1.47e − 3)	1.06e − 2(3.61e − 4)
	2.90e − 2[c]	1.88e − 2	1.06e − 2
$\hat{h}_{DPI}$	3.22e − 2(2.18e − 3)	2.30e − 2(1.18e − 3)	1.44e − 2(1.00e − 3)
	3.22e − 2	2.31e − 2	1.42e − 2
$\hat{h}_{STE}$	2.75e − 2(4.01e − 3)	2.12e − 2(2.63e − 3)	1.37e − 2(1.70e − 3)
	2.75e − 2	2.19e − 2	1.38e − 2
ISE_{ROT}	2.40e − 2(9.50e − 3)	1.01e − 2(2.52e − 3)	1.71e − 2(4.19e − 3)
	2.24 [2]	9.68e − 3 [3]	1.62e − 2 [3]
ISE_{DPI}	2.34e − 2(9.36e − 3)	9.60e − 3(2.60e − 3)	1.64e − 2(4.48e − 3)
	2.17 [1]	9.14e − 3 [1]	1.57e − 2 [2]
ISE_{STE}	2.53e − 2(9.95e − 3)	9.95e − 3(2.68e − 3)	1.66e − 2(6.91e − 3)
	2.40 [3]	9.50e − 3 [2]	1.57e − 2 [1]

NOTE: Average [a] (standard deviation [b]), and median [c] of selected bandwidths and ISE for each strategy. The ISE ranking of each selector is shown in square brackets.

dependence on the blocked quartic fits. We also observed a problem with $\hat{h}_{\mathrm{STE}}$ in that it very occasionally selects a bandwidth that is much larger than the optimum. In our study this occurred twice in the 3,000 cases.

Tables 2 and 3 summarize the results numerically. Averages, medians, and standard deviations of each of the three data-based bandwidths and their corresponding integrated squared errors (ISE) are given. Because of occasional boundary variance problems, the ISE values were computed over [.1, .9].

To test for differences between the ISE's of the three bandwidth selectors, realizations of $\mathrm{ISE}_{\mathrm{ROT}}$, $\mathrm{ISE}_{\mathrm{DPI}}$, and $\mathrm{ISE}_{\mathrm{STE}}$ were retained for each sample from the three curves. Paired Wilcoxon tests were performed to determine whether the median ISE's were significantly different. Bandwidth selectors shared the same ISE ranking when the paired Wilcoxon test showed no difference at the (5/3)% level. Otherwise, separate rankings were assigned with "1" signifying the best performer and "3" the worst. Overall, $\hat{h}_{\mathrm{DPI}}$ performed best, ranking first in all settings but one.

9. CONCLUSION

We have proposed three plug-in bandwidth selection strategies for local linear regression, by adapting ideas used in kernel density estimation. Our comparison through Monte Carlo suggests that $\hat{h}_{\mathrm{DPI}}$ and $\hat{h}_{\mathrm{STE}}$ perform very well in practice—although a more comprehensive simulation study would be required to confirm this and to compare it with other proposals. Moreover, the fast and simple $\hat{h}_{\mathrm{ROT}}$ seems to perform adequately for a wide range of situations. The good performance of $\hat{h}_{\mathrm{DPI}}$ is particularly appealing, because it is based on very simple ideas; is relatively straightforward to implement, requiring no minimization or root-finding; and has good, well-understood theoretical properties. On the other hand, $\hat{h}_{\mathrm{STE}}$ seems to perform about as well as $\hat{h}_{\mathrm{DPI}}$, but has the extra complication of requiring a root-finding step. We speculate that the improved performance of solve-the-equation approaches over direct plug-in approaches in density estimation settings is due to the inadequacy of the normal scale initial estimate, which does not apply here. On balance, we thus recommend $\hat{h}_{\mathrm{DPI}}$ as an effective bandwidth selector for local linear regression and anticipate that its analogs in more complex settings, such as local bandwidth choice, higher-degree fitting, derivative estimation and multivariate designs, will also prove effective.

APPENDIX: DERIVATION OF CONDITIONAL MEAN SQUARED ERROR RESULTS

Derivation of (3) and (4)

To keep the notation less cumbersome, we suppress the p in $\mathbf{X}_{p,x}$. Let $\mathbf{M} = [m(X_1), \ldots, m(X_n)]^T$ and $\mathbf{V} = \mathrm{diag}\{v(X_1), \ldots, v(X_n)\}$. Noting that $E(\mathbf{Y}\mathbf{Y}^T | X_1, \ldots, X_n) = \mathbf{M}\mathbf{M}^T + \mathbf{V}$, and using the approximations of the proof of theorem 4.1 of Ruppert and Wand (1994) and the law of large numbers, we have

$$
\begin{aligned}
&E\{\hat{\theta}_{rs}(g) | X_1, \ldots, X_n\} \\
&\quad = n^{-1} \sum_{i=1}^{n} \{r! \mathbf{e}_{r+1}^T (\mathbf{X}_{X_i}^T \mathbf{W}_{X_i} \mathbf{X}_{X_i})^{-1} \mathbf{X}_{X_i}^T \mathbf{W}_{X_i} \mathbf{M}\} \{s! \mathbf{e}_{s+1}^T (\mathbf{X}_{X_i}^T \mathbf{W}_{X_i} \mathbf{X}_{X_i})^{-1} \mathbf{X}_{X_i}^T \mathbf{W}_{X_i} \mathbf{M}\} \\
&\qquad + n^{-1} r! s! \sum_{i=1}^{n} \mathbf{e}_{r+1}^T (\mathbf{X}_{X_i}^T \mathbf{W}_{X_i} \mathbf{X}_{X_i})^{-1} \mathbf{X}_{X_i}^T \mathbf{W}_{X_i} \mathbf{V} \mathbf{W}_{X_i} \mathbf{X}_{X_i} (\mathbf{X}_{X_i}^T \mathbf{W}_{X_i} \mathbf{X}_{X_i})^{-1} \mathbf{e}_{s+1} \\
&\quad \simeq n^{-1} \sum_{i=1}^{n} \left[m^{(r)}(X_i) + \mu_{p+1}(K_{r,p}) \left\{ \frac{m^{(p+1)}(X_i)}{(p+1)!} \right\} g^{p-r+1} \right] \left[m^{(s)}(X_i) + \mu_{p+1}(K_{s,p}) \left\{ \frac{m^{(p+1)}(X_i)}{(p+1)!} \right\} g^{p-s+1} \right] \\
&\qquad + n^{-1} g^{-r-s-1} \left\{ \int K_{r,p}(u) K_{s,p}(u)\, du \right\} \left[n^{-1} \sum_{i=1}^{n} v(X_i) f(X_i)^{-1} \right] \\
&\quad \simeq \theta_{rs} + \frac{\mu_{p+1}(K_{s,p})}{(p+1)!} g^{p-s+1} \theta_{r,p+1} + \frac{\mu_{p+1}(K_{r,p})}{(p+1)!} g^{p-r+1} \theta_{s,p+1} \\
&\qquad + n^{-1} g^{-r-s-1} \left[\int K_{r,p}(u) K_{s,p}(u)\, du \right] \int_S v(x)\, dx,
\end{aligned}
$$

which immediately leads to (3).

For the conditional variance, note that

$$
\begin{aligned}
\hat{\theta}_{rs}(g) &= n^{-1} \sum_{i=1}^{n} \sum_{j=1}^{n} L_{rs}^*(X_i, X_j) Y_i Y_j \\
&= n^{-1} \sum_{i=1}^{n} \sum_{j=1}^{n} L_{rs}(X_i, X_j) Y_i Y_j, \qquad \text{(A.1)}
\end{aligned}
$$

where

$$
L_{rs}(X_i, X_j) = \sum_{k=1}^{n} \{r! \mathbf{e}_{r+1}^T (\mathbf{X}_{X_k}^T \mathbf{W}_{X_k} \mathbf{X}_{X_k})^{-1} \mathbf{X}_{X_k}^T \mathbf{W}_{X_k} \mathbf{e}_i\} \times \{s! \mathbf{e}_{s+1}^T (\mathbf{X}_{X_k}^T \mathbf{W}_{X_k} \mathbf{X}_{X_k})^{-1} \mathbf{X}_{X_k}^T \mathbf{W}_{X_k} \mathbf{e}_j\}
$$

and $L_{rs}(X_i, X_j) = [L_{rs}^*(X_i, X_j) + L_{rs}^*(X_j, X_i)]/2$. We introduce L_{rs} because it is symmetric. L_{rs}^* is not symmetric, but can be seen to be "asymptotically symmetric."

We seek an approximation to $L^*_{rs}(X_i, X_j)$. In the proof of theorem 4.2 of Ruppert and Wand (1994), it is shown that for $p - r$ odd,

$$r!\mathbf{e}_{r+1}^T(n^{-1}\mathbf{X}_{X_k}^T\mathbf{W}_{X_k}\mathbf{X}_{X_k})^{-1} \simeq g^{-r}f(X_k)^{-1}\mathbf{e}_{r+1}^T\mathbf{N}_p^{-1}\mathbf{A}^{-1},$$

where $\mathbf{N}_p$ is the $(p+1)\times(p+1)$ matrix having (l,l') entry equal to $\int u^{l+l'-2}K(u)\,du$ and $\mathbf{A} = \operatorname{diag}(1, g, \ldots, g^p)$. From this approximation and a cofactor argument analogous to that used in the proof of theorem 4.1 of Ruppert and Wand (1994), we obtain

$$\begin{aligned} &r!\mathbf{e}_{r+1}^T(\mathbf{X}_{X_k}^T\mathbf{W}_{X_k}\mathbf{X}_{X_k})^{-1}\mathbf{X}_{X_k}^T\mathbf{W}_{X_k}\mathbf{e}_i \\ &\qquad \simeq n^{-1}g^{-r-1}f(X_k)^{-1}K_{r,p}\{(X_i - X_k)/g\}. \end{aligned}$$

This leads to

$$\begin{aligned} L^*_{rs}(X_i, X_j) &\simeq n^{-2}g^{-r-s-2}\sum_{k=1}^{n} f(X_k)^{-2}K_{r,p}\{(X_i - X_k)/g\} \\ &\quad \times K_{s,p}\{(X_j - X_k)/g\} \\ &\simeq n^{-1}g^{-r-s-2}\int f(z)^{-1}K_{r,p}\{(X_i - z)/g\} \\ &\quad \times K_{s,p}\{(X_j - z)/g\}\,dz. \end{aligned} \tag{A.2}$$

From (A.1) and the symmetry of L_{rs}, we have

$$\begin{aligned} &\operatorname{var}\{\hat{\theta}_{rs}(g)|X_1, \ldots, X_n\} \\ &= n^{-2}\sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{k=1}^{n}\sum_{l=1}^{n} L_{rs}(X_i, X_j)L_{rs}(X_k, X_l)\operatorname{cov}(Y_iY_j, Y_kY_l|X_1, \ldots, X_n) \\ &= n^{-2}\sum_{i=1}^{n} L_{rs}(X_i, X_i)^2\operatorname{var}(Y_i^2|X_1, \ldots, X_n) + 2n^{-2}\sum\sum_{i\neq j} L_{rs}(X_i, X_j)^2\operatorname{var}(Y_iY_j|X_1, \ldots, X_n) \\ &\quad + 2n^{-2}\sum\sum_{i\neq j} L_{rs}(X_i, X_i)L_{rs}(X_i, X_j)\operatorname{cov}(Y_i^2, Y_iY_j|X_1, \ldots, X_n) \\ &\quad + 4n^{-2}\sum\sum\sum_{i\neq j\neq k} L_{rs}(X_i, X_j)L_{rs}(X_i, X_k)\operatorname{cov}(Y_iY_j, Y_iY_k|X_1, \ldots, X_n). \end{aligned} \tag{A.3}$$

Let

$$\begin{aligned} \tau(X_i, X_j) &\equiv \operatorname{var}(Y_iY_j|X_1, \ldots, X_n) \\ &= v(X_i)v(X_j) + v(X_i)m(X_j)^2 \\ &\quad + v(X_j)m(X_i)^2 \end{aligned}$$

and $T_{rs} \equiv K_{r,p} * K_{s,p}$. Also, note that $K_{r,p}$ and $K_{s,p}$ are either both symmetric or both skew-symmetric when $r + s$ is even.

The second term of (A.3) equals

$$\begin{aligned} &\frac{1}{2n^2}\sum\sum_{i\neq j}[L_{rs}(X_i, X_j)^2 + 2L_{rs}(X_i, X_j)L_{rs}(X_j, X_k) \\ &\qquad + L_{rs}(X_j, X_i)] \\ &\quad \times \operatorname{var}(Y_iY_j|X_1, \ldots, X_n) = B_1 + B_2 + B_3, \text{ say.} \end{aligned}$$

Using (A.2), B_1 can be approximated by

$$\begin{aligned} &n^{-4}g^{-2r-2s-4}\int\int f(w)^{-1}f(z)^{-1}\sum\sum_{i\neq j} K_{r,p}\{(X_i - w)/g\}K_{s,p}\{(X_j - w)/g\} \\ &\quad \times K_{r,p}\{(X_i - z)/g\}K_{s,p}\{(X_j - z)/g\}\tau(X_i, X_j)\,dw\,dz \\ &\simeq 2n^{-2}g^{-2r-2s-4}\int\int\int\int f(w)^{-1}f(z)^{-1}K_{r,p}\{(x - w)/g\}K_{s,p}\{(y - w)/g\} \\ &\quad \times K_{r,p}\{(x - z)/g\}K_{s,p}\{(Y - z)/g\}\tau(x, y)f(x)f(y)\,dx\,dy\,dw\,dz \\ &\simeq n^{-2}g^{-2r-2s-2}\int\int f(x - gu)^{-1}f(x - gv)^{-1}\int K_{r,p}(u)K_{s,p}\{(x - y)/g - u\}\,du \\ &\quad \times \int K_{r,p}(v)K_{s,p}\{(x - y)/g - v\}\,dv\tau(x, y)f(x)f(y)\,dx\,dy \\ &\simeq n^{-2}g^{-2r-2s-2}\int\int f(x)^{-1}f(y)\tau(x, y)T_{rs}\{(x - y)/g\}^2\,dx\,dy \\ &\simeq n^{-2}g^{-2r-2s-1}R(T_{rs})\int_S v(x)\{v(x) + 2m(x)^2\}\,dx. \end{aligned}$$

By a similar argument, each of B_2 and B_3 can be approximated by the same quantity.

Now we consider the fourth term of (A.3). Note that

$$\begin{aligned} L_{rs}(X_i, X_j)L_{rs}(X_i, X_k) &= [L^*_{rs}(X_i, X_j)L^*_{rs}(X_i, X_k) \\ &\quad + L^*_{rs}(X_j, X_i)L^*_{rs}(X_i, X_k) \\ &\quad + L^*_{rs}(X_i, X_j)L^*_{rs}(X_k, X_i) \\ &\quad + L^*_{rs}(X_j, X_i)L^*_{rs}(X_k, X_i)]. \end{aligned}$$

Substituting this into the fourth term of (A.3), the fourth term can be written as $C_1 + C_2 + C_3 + C_4$, say. Each of C_i's can be approximated by the same quantity. We demonstrate this approximation for C_1. Observing that $\operatorname{cov}(Y_iY_j, Y_iY_k|X_1, \ldots, X_n) = v(X_i)m(X_j)m(X_k)$, C_1 is approximately

$$\begin{aligned} & n^{-4}g^{-2r-2s-4}\int\int f(w)^{-1}f(z)^{-1}\sum\sum\sum_{i\neq j\neq k} K_{r,p}\{(X_i-w)/g\}K_{s,p}\{(X_j-w)/g\} \\ &\quad \times K_{r,p}\{(X_i-z)/g\}K_{s,p}\{(X_k-z)/g\}v(X_i)m(X_i)m(X_k)\,dw\,dz \\ &\simeq n^{-1}g^{-2r-2s-4}\int\int\int\int\int f(w)^{-1}f(z)^{-1}K_{r,p}\{(x-w)/g\}K_{s,p}\{(y-w)/g\} \\ &\quad \times K_{r,p}\{(x-z)/g\}K_{s,p}\{(t-z)/g\}f(x)f(y)f(t)v(x)m(y)m(t)\,dw\,dz\,dx\,dy\,dt \\ &= n^{-1}g^{-2r-2s-2}\int\int\int f(x-gu)^{-1}f(x-gv)^{-1}\int K_{r,p}(u)K_{s,p}\{(x-y)/g-u\}\,du \\ &\quad \times \int K_{r,p}(v)K_{s,p}\{(x-t)/g-v\}\,dv f(x)f(y)f(t)v(x)m(y)m(t)\,dx\,dy\,dt \\ &= n^{-1}g^{-2r-2s-2}\int\int\int f(x)^{-1}f(y)f(t)v(x)m(y)m(t)T_{rs}\{(x-y)/g\}T_{rs}\{(x-t)/g\}\,dx\,dy\,dt \\ &= n^{-1}g^{-2r-2s}\int\int\int T_{rs}(u)T_{rs}(v)f(x)^{-1}f(x-gu)f(x-gv)v(x)m(x-gu)m(x-gv)\,dx\,du\,dv \\ &= n^{-1}g^{-2r-2s}\int v(x)f(x)^{-1}\left\{\int (mf)(x-gu)T_{rs}(u)\,du\right\}^2 dx \\ &\simeq n^{-1}\int_S (mf)^{(r+s)}(x)^2 v(x)f(x)^{-1}\,dx. \end{aligned}$$

The last step follows from Taylor expansion and the result

$$\begin{aligned} \int u^j T_{rs}(u)\,du &= 0 \qquad j = 0, \ldots, r+s-1, \\ &= (r+s)! \qquad j = r+s. \end{aligned}$$

This result can be easily derived by noting the moment properties of $K_{r,p}$ given by Ruppert and Wand (1994). Similar arguments can be used to show that the first and third terms of (A.3) lead to terms of lower order, so the stated result for the conditional variance of $\hat{\theta}_{rs}(g)$ follows immediately.

Derivation of (7) and (8)

Results (7) and (8) can be easily derived by extending the proof of Hall and Marron (1990) to the case of general odd p, using the approximations of Ruppert and Wand (1994). First, note that

$$\hat{\sigma}^2_p(\lambda) = \nu^{-1}\sum_{i=1}^n \left(Y_i - \sum_{j=1}^n w_{ij}Y_j\right)^2.$$

From Hall and Marron (1990), $E\{\hat{\sigma}^2_p(\lambda) - \sigma^2|X_1, \ldots, X_n\} = \nu^{-1}\sum_{i=1}^n \delta_i^2$ and

$$\operatorname{var}\{\hat{\sigma}^2_p(\lambda)|X_1, \ldots, X_n\} = \nu^{-2}\left\{\sum_{j=1}^n E(\Delta_j^2) + 2\sigma^4\sum\sum_{j\neq k} t_{jk}^2\right\},$$

where

$$\delta_i = m(X_i) - \sum_{j=1}^n w_{ij}m(X_j),$$

$$t_{jk} = \sum_{i=1}^n w_{ij}w_{ik} - 2w_{jk},$$

and

$$\Delta_j = \left(\delta_j - \sum_{i=1}^n \delta_i w_{ij}\right)\sigma\varepsilon_j + \left(1 - 2w_{jj} + \sum_{i=1}^n w_{ij}^2\right)\sigma^2(\varepsilon_j^2 - 1).$$

Arguments analogous to those used to prove (A.2) lead to the approximations

$$w_{ij} \simeq K_p\{(X_i - X_j)/\lambda\}/\{n\lambda f(X_i)\}$$

and

$$t_{jk} \simeq (K_p * K_p - 2K_p)\{(X_j - X_k)/\lambda\}/\{n\lambda f(X_j)\}.$$

Thus, noting that $\nu \simeq n$ and appealing to theorem 4.1 of Ruppert and Wand (1994), we obtain

$$\begin{aligned} & E\{\hat{\sigma}^2_p(\lambda) - \sigma^2|X_1, \ldots, X_n\} \\ &\quad \simeq n^{-1}\{\lambda^{p+1}\mu_{p+1}(K_p)/(p+1)!\}^2\sum_{i=1}^n m^{(p+1)}(X_i)^2 \\ &\quad \simeq \lambda^{2p+2}\{\mu_{p+1}(K_p)/(p+1)!\}^2\theta_{p+1,p+1}. \end{aligned}$$

Also,

$$\sum\sum_{j\neq k} t_{jk}^2 \simeq \lambda^{-1}(b-a)R(K_p * K_p - 2K_p),$$

which leads to

$$\operatorname{var}\{\hat{\sigma}_p^2 | X_1, \ldots, X_n\} \simeq n^{-1}\sigma^4 \operatorname{var}(\varepsilon^2) + 2n^{-2}\lambda^{-1}(b-a)\sigma^4 R(K_p * K_p - 2K_p).$$

[Received July 1993. Revised July 1994.]

REFERENCES

Chiu, S.-T. (1991), "Bandwidth Selection for Kernel Density Estimation," *The Annals of Statistics,* 19, 1883–1905.

——— (1992), "An Automatic Bandwidth Selector for Kernel Density Estimation," *Biometrika,* 79, 771–782.

Fan, J. (1992), "Design-Adaptive Nonparametric Regression," *Journal of the American Statistical Association,* 87, 998–1004.

——— (1993), "Local Linear Regression Smoothers and Their Minimax Efficiency," *The Annals of Statistics,* 21, 196–216.

Fan, J., and Gijbels, I. (1993), "Bandwidth and Adaptive Order Selection for Local Polynomial Fitting in Function Estimation," Institute of Statistics Mimeo Series #2080, University of North Carolina, Chapel Hill.

——— (1995), "Data-Driven Bandwidth Selection in Local Polynomial Fitting: Variable Bandwidth and Spatial Adaptation," *Journal of the Royal Statistical Society,* Ser. B, 57, 371–394.

Fan, J., and Marron, J. S. (1994). "Fast Implementations of Nonparametric Curve Estimators," *Journal of Computational and Graphical Statistics,* 3, 35–56.

Gasser, T., Kneip, A., and Köhler, W. (1991), "A Fast and Flexible Method for Automatic Smoothing," *Journal of the American Statistical Association,* 86, 643–52.

Hall, P., and Marron, J. S. (1990), "On Variance Estimation in Nonparametric Regression," *Biometrika,* 77, 415–419.

Hall, P., Sheather, S. J., Jones, M. C., and Marron, J. S. (1991), "On Optimal Data-Based Bandwidth Selection in Kernel Density Estimation," *Biometrika,* 78, 521–530.

Hall, P., and Wand, M. P. (1993), "On the Accuracy of Binned Approximations to Kernel Estimators," Working Paper Series 93-003, University of New South Wales, Australian Graduate School of Management.

Härdle, W., Hall, P., and Marron, J. S. (1988), "How Far are Automatically Chosen Regression Smoothing Parameters From Their Optimum?" (with discussion), *Journal of the American Statistical Association,* 83, 86–99.

Härdle, W., and Marron, J. S. (1993), "Fast and Simple Scatterplot Smoothing," CORE Discussion Paper No. 9143, Universite Catholique de Louvain.

Jones, M. C., Marron, J. S., and Park, B. U. (1991), "A Simple Root-n Bandwidth Selector," *The Annals of Statistics,* 4, 1919–1932.

Jones, M. C., Marron, J. S., and Sheather, S. J. (1992), "Progress in Data-Based Bandwidth Selection for Kernel Density Estimation," Working Paper Series 92-014, University of New South Wales, Australian Graduate School of Management.

Mallows, C. L. (1973), "Some Comments on C_p," *Technometrics,* 15, 661–675.

Park, B. U., and Marron, J. S. (1990), "Comparison of Data-Driven Bandwidth Selectors," *Journal of the American Statistical Association,* 85, 66–72.

——— (1992), "On the Use of Pilot Estimators in Bandwidth Selection," *Journal of Nonparametric Statistics,* 1, 231–240.

Park, B. U., and Turlach, B. A. (1992), "Practical Performance of Several Data-Driven Bandwidth Selectors," *Computational Statistics,* 7, 251–270.

Ruppert, D., and Wand, M. P. (1994), "Multivariate Locally Weighted Least Squares Regression," *The Annals of Statistics,* 22, 1346–1370.

Sheather, S. J. (1992), "The Performance of Six Popular Bandwidth Selection Methods on Some Real Data Sets," *Computational Statistics,* 7, 225–250.

Sheather, S. J., and Jones, M. C. (1991), "A Reliable Data-Based Bandwidth Selection Method for Kernel Density Estimation," *Journal of the Royal Statistical Society,* Ser. B, 53, 683–690.