Local Polynomial Variance-Function Estimation

David RUPPERT
School of Operations Research and Industrial Engineering
Cornell University
Ithaca, New York 14853

M. P. WAND
Australian Graduate School of Management
University of New South Wales
Sydney 2052
Australia

Ulla HOLST and Ola HÖSSJER
Department of Mathematical Statistics
Lund University
S-221 00 Lund
Sweden

The conditional variance function in a heteroscedastic, nonparametric regression model is estimated by linear smoothing of squared residuals. Attention is focused on local polynomial smoothers. Both the mean and variance functions are assumed to be smooth, but neither is assumed to be in a parametric family. The biasing effect of preliminary estimation of the mean is studied, and a degrees-of-freedom correction of bias is proposed. The corrected method is shown to be adaptive in the sense that the variance function can be estimated with the same asymptotic mean and variance as if the mean function were known. A proposal is made for using standard bandwidth selectors for estimating both the mean and variance functions. The proposal is illustrated with data from the LIDAR method of measuring atmospheric pollutants and from turbulence-model computations.

KEY WORDS: Bandwidth; Heteroscedasticity; Kernel smoothing; Nonparametric regression; Smoother matrix.

In regression analysis, it is often the case that the homoscedasticity assumption is violated. An example of this is given in Figure 1(a). The data are taken from Holst, Hössjer, Björklund, Ragnarson, and Edner (1996), who used local polynomial regression for evaluation of the concentration of atmospheric atomic mercury measured with LIDAR technique (LIght Detection And Ranging; see Sigrist 1994). The concentration is proportional to the derivative of the mean function, but because of the severe heteroscedasticity, the variance function must be estimated to obtain a satisfactory bandwidth for the derivative and further to estimate the variance of the total amount of pollutants in a certain area. Holst et al. (1996) used a parametric model for the variance function.

In other examples, the variance function itself is of interest in its own right. For example, one of the authors (DR) is collaborating with mechanical engineers at Cornell on the analysis of data from the Monte Carlo simulation of turbulence by the pdf (probability density function) method (Pope 1985). In this work, one has available the spatial position, velocity, and other properties of simulated particles. One, of course, needs to estimate conditional expectations such as mean velocity as a function of position. When studying turbulence, however, the variance of velocity and its derivatives as functions of position are also essential; see Section 4.4.

In this article we extend local polynomial regression to estimation of the variance function. As we show in Section 1, our proposal can be generalized to any linear smoother (e.g., smoothing splines, running means). Nevertheless, we focus on local polynomials because of their intuitiveness and simplicity. Our theoretical analyses show that the attractive properties of odd-degree local polynomial smoothers, such as design adaptivity and automatic boundary correction, carry over to variance-function estimation. "Design adaptivity" (Fan 1992) refers to local polynomial estimation's elimination of bias and extra variability due to unequally spaced predictor variables.

The literature on nonparametric variance-function estimation is rather sparse. Carroll (1982) developed kernel estimators in the context of linear regression, and Müller and Stadtmüller (1987) and Hall and Carroll (1989) proposed and analyzed kernel-type variance-function estimators assuming a nonparametric mean function. Fan and Gijbels (1995) proposed a type of local polynomial variance-function estimator as part of their bandwidth-selection procedure. These works and several others are discussed in more detail in Section 1.5.

Our main theoretical contributions are deeper results on the bias and variance of the estimated variance function. These results are important because they address the major practical problem of choosing bandwidths for estimating the mean and variance functions. Moreover, to the best of our knowledge, we are the first to consider estimating derivatives of the variance function, a topic with applications in engineering.

In Section 1, we formulate a general class of nonparametric variance-function estimators, with local polynomial

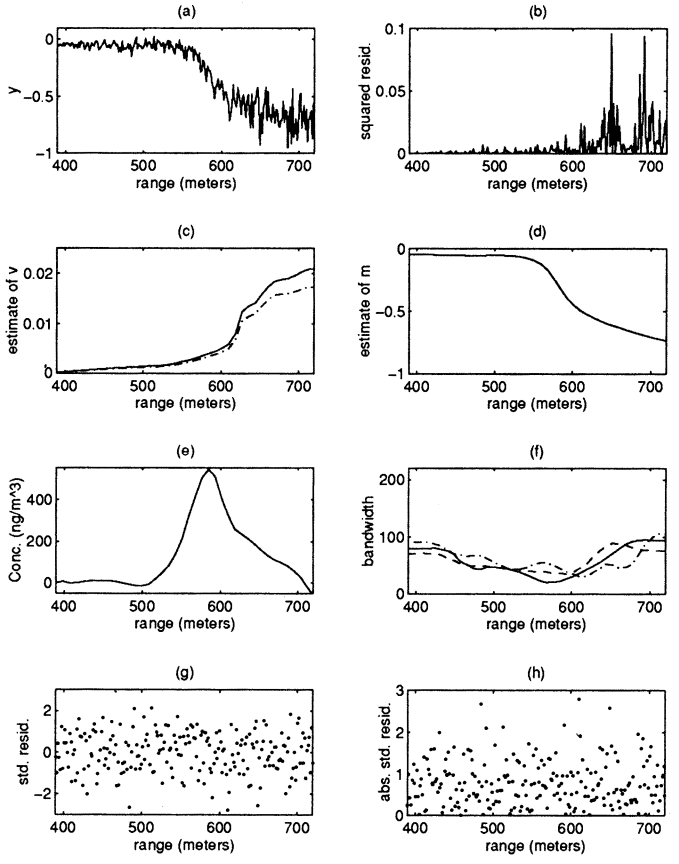

Figure 1. LIDAR Data: (a) Raw Data (221 observations); (b) Squared Residuals From a Preliminary Local Quadratic Estimate of the Mean Function With a Span of 5%; (c) Local Linear Estimate of the Variance Function v, Corrected for Estimation of the Mean (solid) and Uncorrected (dotted and dashed); (d) Local Quadratic Estimate of the Mean Function m; (e) Local Quadratic Estimate of Concentration = Cm′; (f) Bandwidths for Estimation of m (solid), Estimation of v (dotted and dashed), and Estimation of m′ (dashed); (g) Standardized Residuals = $\{Y_i - \hat{m}(X_i)\}\hat{v} \neq^{1/2} (X_i)$; (h) Absolute Standardized Residuals.

variance estimators as a special case. Section 2 investigates the theoretical properties of these estimators and applies these results to bandwidth selection. Computational methods are in Section 3. Section 4 illustrates the methodology.

The variance-function estimator in Section 1 was proposed independently by Mathur (1995), but the asymptotic theory, computational implementation, and bandwidth selectors proposed here were not given by Mathur.

1. FORMULATION

1.1 A General Class of Variance-Function Estimators

The local polynomial estimates of variance that we consider in this article can be defined for general linear smoothers, so we start at this level of generality.

Let $(X_1, Y_1), \ldots, (X_n, Y_n)$ be a sample of random pairs that are assumed to satisfy the heteroscedastic nonparametric regression model

$$Y_i = m(X_i) + \varepsilon_i, \qquad \operatorname{var}(\varepsilon_i) = v(X_i), \quad i = 1, \ldots, n, \quad (1)$$

where the *errors* $\varepsilon_1, \ldots, \varepsilon_n$ are independent zero-mean random variables satisfying $E(\varepsilon_i^4) < \infty$. We call m the *mean function* and v the *variance function*. We will also let m and v denote the column vectors containing values of $m(X_i)$ and $v(X_i), 1 \le i \le n$, respectively. Finally, Y will be used to denote the $n \times 1$ vector of Y_i values.

Suppose that $\hat{m} = [\hat{m}(X_1), \ldots, \hat{m}(X_n)]'$ is a linear smooth of the (X_i, Y_i)'s. By this, we mean that $\hat{m} = SY$ for some $n \times n$ matrix S, often referred to as the *smoother matrix*. Examples of linear smoothers include smoothing splines, regression splines, and local polynomials (e.g., see Hastie and Tibshirani 1990). It is assumed that S preserves constant vectors in the sense that $S\mathbf{1} = \mathbf{1}$, where $\mathbf{1}$ denotes a vector of ones.

Let S_1 be the smoother matrix corresponding to an initial smooth of the data, and put $r = (I - S_1)Y$, the vector of residuals. Then a natural means of estimating $v = [v(X_1), \ldots, v(X_n)]'$ is to smooth the squared residuals to obtain $S_2 r^2$. Here S_2 is another smoother matrix and r^2 contains the squares of the entries of r. It seems reasonable that our estimator should be unbiased when the errors are homoscedastic—that is, $v = \sigma^2\mathbf{1}$ for $\sigma^2 > 0$—and the bias of $\hat{m}$ from the initial smoother S_1 can be ignored. Under homoscedasticity,

$$E(S_2 r^2 | X_1, \ldots, X_n) = S_2[\{E(S_1 Y | X_1, \ldots, X_n) - m\}^2 + \sigma^2(\mathbf{1} + \Delta)],$$

where $\Delta = \operatorname{diag}(S_1 S_1' - 2S_1)$ and $\operatorname{diag}(A)$ denotes the column vector containing the diagonal entries of the square matrix A. Because $E(S_2 r^2 | X_1, \ldots, X_n) = \sigma^2(\mathbf{1} + S_2\Delta)$ when $S_1 Y$ is conditionally unbiased, this motivates the estimator

$$\hat{v} = (S_2 r^2)/(\mathbf{1} + S_2\Delta). \quad (2)$$

The convention here and throughout is that the vector multiplication and division are elementwise.

As a referee has pointed out, an alternative to (2) is to "studentize" r_i by dividing by $\sqrt{1 + \Delta_{ii}}$. This eliminates the bias due to estimating the mean. The squared studentized residual vector, call it $\tilde{r}^2$, can then be smoothed so that

$$\hat{v} = S_2 \tilde{r}^2 = S_2(r^2/(\mathbf{1} + \Delta)). \quad (3)$$

In examples, we found little difference between (2) and (3). The advantage of (3) is a savings in computational effort in that one need not smooth Δ. We have found, however, that most of the computational effort is in bandwidth selection, not the subsequent smoothing, so this advantage is minimal. We decided to use (2) because of some analogies between it and estimators for parametric models. These analogies are discussed next.

1.2 Relationships With Parametric Modeling

One can view variance-function estimators given by (2) as generalizations of those commonly used when the mean or variance function is modeled parametrically. For example, if the mean is modeled linearly, $Y_i = (X\beta)_i + \varepsilon_i$, $\operatorname{var}(\varepsilon_i) = v(X_i), i = 1, \ldots, n$, where X is an $n \times p$ design matrix and β is a $p \times 1$ matrix of coefficients, then one should replace S_1 by the "hat" matrix $R = X(X'X)^{-1}X'$. Using the symmetry and idempotency of R, we obtain the variance-function estimator $\hat{v} = S_2\{(R - I)Y\}^2/[\mathbf{1} - S_2\{\operatorname{diag}(R)\}]$.

On the other hand, if the homoscedastic nonparametric regression model $Y_i = m(X_i) + \varepsilon$, $\text{var}(\varepsilon_i) = \sigma^2, i = 1, \ldots, n$, is assumed, then one should simply average the squared residuals by taking $S_2 = n^{-1}\mathbf{1}\mathbf{1}'$. This results in $\hat{\sigma}^2 = \{Y'(S_1 - I)'(S_1 - I)Y\}/\{n + \text{tr}(S_1 S_1' - 2S_1)\}$, which includes variance estimators for nonparametric regression considered by, for example, Buckley, Eagleson, and Silverman (1988) and Cleveland and Devlin (1988).

For the homoscedastic linear regression model, the estimator reduces to the familiar $\hat{\sigma}^2 = Y'(I - R)Y/(n - p)$.

1.3 Local Polynomial Variance-Function Estimation

The class of linear smoothers that we concentrate on is made up of those commonly referred to as *local polynomial smoothers;* for example, see Wand and Jones (1995) or Fan and Gijbels (1996) for an introduction. These smoothers were introduced into modern statistical practice in an important article by Cleveland (1979), and they became an important modern theoretical tool starting with Stone (1977). Local polynomial regression, however, has a long history (Cleveland and Loader 1996). Interesting examples of local polynomial regression in applied statistics were given by Cleveland and Devlin (1988).

To estimate $m(x)$ at a fixed x, we fit a pth-degree polynomial to the data by weighted least squares, with the weight given to (X_i, Y_i) decreasing to 0 as the distance from X_i to x increases. As usual in the literature, we will use the weight $K\{(X_i - x)/h\}$, where K is a pdf and h is a bandwidth. Then $\hat{m}(x)$ is the predicted value at x of this weighted least squares fit. For both notational and computational reasons, we work with deviations from x—that is, $(X_i - x)$. Then $\hat{m}(x)$ is the intercept of the fit. Thus, using the matrix solution to a weighted least squares problem, the (i, j) entry of the pth-degree local polynomial smoother matrix, $S_{p,h}$, is

$$(S_{p,h})_{ij} = e_1'\{X_p(X_i)'W_h(X_i)X_p(X_i)\}^{-1}X_p(X_i)'W_h(X_i)e_j, \quad (4)$$

where e_i is the column vector with 1 in the ith position and zeros elsewhere,

$$X_p(x) = \begin{bmatrix} 1 & X_1 - x & \cdots & (X_1 - x)^p \\ \vdots & \vdots & \ddots & \vdots \\ 1 & X_n - x & \cdots & (X_n - x)^p \end{bmatrix}$$

and

$$W_h(x) = \text{diag}_{1 \le i \le n} K\left(\frac{X_i - x}{h}\right),$$

where $\text{diag}_{1 \le i \le n} a_i$ denotes the $n \times n$ diagonal matrix with $a_1, \ldots, a_n$ on the diagonal. The premultiplication by e_1' picks off the estimated intercept from the weighted least squares estimate.

Using this notation, one can define the local polynomial estimate of $v(x)$ to be

$$\hat{v}(x) = \hat{v}(x; p_1, h_1, p_2, h_2) = \frac{e_1'\{X_{p_2}(x)'W_{h_2}(x)X_{p_2}(x)\}^{-1}X_{p_2}(x)'W_{h_2}(x)r^2}{1 + e_1'\{X_{p_2}(x)'W_{h_2}(x)X_{p_2}(x)\}^{-1}X_{p_2}(x)'W_{h_2}(x)\Delta},$$

where $r = (I - S_{p_1,h_1})Y$ and $\Delta = \text{diag}(S_{p_1,h_1}S_{p_1,h_1}' - 2S_{p_1,h_1})$.

For estimation of v at the observations, this definition is easily seen to be a member of the class of variance estimators described by (2), with $S_1 = S_{p_1,h_1}$ and $S_2 = S_{p_2,h_2}$.

1.4 Estimation of Derivatives of the Variance Function

As mentioned in the introduction, some applications require that derivatives of v be estimated. For example, the first two derivatives of v are used in the study of turbulence; see Section 4.4. As discussed by Ruppert and Wand (1994), local polynomial estimation of the kth derivative of m is straightforward in principle, and there is no problem extending derivative estimation to v. In practice, however, accurate estimation of a derivative may require large sample sizes, especially if $k > 1$, and appropriate values of the degree of the polynomial and the bandwidth depend on k and must be chosen carefully. One needs to use $p_2 \ge k$, and then for the second smoother matrix, S_2, one merely replaces e_1' in (4) by $k!e_{k+1}'$. The theory in the next section extends easily to derivative estimation, but for simplicity we only consider the case of estimating v itself.

1.5 Other Work on Nonparametric Variance-Function Estimation

Works on nonparametric variance estimation can be categorized according to the following criteria:

1. The mean function may be parametrically or nonparametrically modeled.
2. The variance function may be considered constant (homoscedastic) or nonconstant (heteroscedastic).
3. In a first stage of estimation, the variance at X_i, $v(X_i)$, may be estimated by a residual from a preliminary fit. Alternatively, one may use a squared "pseudo-residual," which is a weighted average of a fixed (independent of n) number of the Y_i's. The weights sum to 0 to eliminate a constant mean and the squared weights sum to 1 so that the squared pseudo-residual estimates the local variance. The term "pseudo-residual" comes from Müller and Stadtmüller (1987).
4. In a second estimation stage, the squared residuals or pseudo-residuals may be kernel-smoothed or smoothed by local polynomial regression.
5. The bandwidth for smoothing the residuals or pseudo-residuals may be chosen subjectively or by a data-based method.

Regarding Criterion 1, an important early article by Carroll (1982) used a parametric model for m. Hall and Carroll (1989) considered both parametric and nonparametric models for m. All other works that we are aware of model m nonparametrically.

Buckley et al. (1988) modeled homoscedastic data and Silverman (1985) modeled heteroscedastic data by smoothing squared residuals from a spline fit. Rice (1984) assumed homoscedasticity and used the simplest pseudo-residual $(Y_i - Y_{i-1})/\sqrt{2}$, where the (X_i, Y_i) pairs have been sorted by the X_i's. Gasser, Sroka, and Jennen-Steinmetz (1986),

who concentrated on homoscedasticity but also considered heteroscedasticity, fitted a straight line to (X_{i-1}, Y_{i-1}) and (X_{i+1}, Y_{i+1}) and used as a pseudo-residual the deviation, suitably normalized, between Y_i and the line at X_i. Müller and Stadtmüller (1987) discussed more general pseudo-residuals.

Assuming homoscedasticity, Hall, Kay, and Titterington (1990) found asymptotically optimal estimators based on pseudo-residuals, which they called "difference-based" estimators. These authors mentioned that estimators based on residuals are more efficient than those based on pseudo-residuals, but they argued for using pseudo-residuals because using residuals requires choosing a bandwidth for estimation of the mean. This requirement of a bandwidth for the mean is not a serious problem, however, as we shall show. Moreover, because pseudo-residuals are based on a fixed number of X_i's, they are correlated, even asymptotically, which complicates their analysis. In contrast, if $\hat{m}$ is consistent, then the residuals are asymptotically uncorrelated. As we shall argue in Section 2.3, standard bandwidth selectors developed for *independent* observations can be used for smoothing squared residuals. This is not true of squared pseudo-residuals.

Most works suggest smoothing the squared residuals or pseudo-residuals by kernel smoothing. We advocate local polynomial methods because of their automatic boundary bias correction and adaptivity to unequally spaced designs. The general class of variance-function estimation introduced in Section 1.1 includes smoothing squared residuals by kernels, local polynomial regression, smoothing splines, or any other linear smoother, but this class does not include estimators that smooth pseudo-residuals.

As far as we are aware, there have been no prior proposals for bandwidth selection when estimating a variance function. Based on the theory developed in Section 2.2, in Section 2.3 we make a broad proposal: Take a favorite bandwidth selector for estimating the mean function and apply it to smoothing the squared residuals. Thus, in regard to Criterion 5, this is the only work that we are aware of with a data-based method.

2. THEORY

In this section we first obtain exact matrix algebraic expressions for the conditional mean and covariance of $\hat{v}$ for the general class of variance-function estimators introduced in Section 1.1. With local polynomials, these results yield meaningful asymptotic approximations, which are useful for choosing bandwidths or assessing the variability of the estimates.

We retain the convention that multiplication and division of column vectors is elementwise. For square matrices A and B we avoid confusion between usual matrix multiplication and element-wise multiplication by using the notation $A \odot B$ for the latter (this is sometimes called the *Hadamard product* of A and B). We let $\chi = \{X_1, \ldots, X_n\}$ to abbreviate expectations that are conditional on the predictors. Moreover, $\operatorname{cov}(U|W)$ denotes the conditional covariance matrix of U given W whenever U and W are random vectors.

2.1 General Variance-Function Estimators

The following matrices are useful for a concise representation of the bias and covariance of $\hat{v}$:

$$V = \operatorname{diag}(v), \quad G = \operatorname{diag}_{1\le i\le n}\{E(\varepsilon_i^3)\},$$

$$T = \operatorname{diag}_{1\le i\le n}\{E(\varepsilon_i^4)\}.$$

Theorem 1. Let $b_1 = (S_1 - I)m$ denote the bias vector of the smooth S_1. Then

$$E(\hat{v} - v|\chi) = \frac{(S_2 - I)v + S_2\{b_1^2 + \operatorname{diag}(S_1 V S_1' - 2S_1 V)\} - (S_2\Delta)v}{\mathbf{1} + S_2\Delta} \tag{5}$$

and

$$\begin{aligned}\operatorname{cov}(\hat{v}|\chi) = {} & S_2[\{(S_1 - I)\odot(S_1 - I)\}(T - 3V^2) \\ & \times \{(S_1 - I)\odot(S_1 - I)\}' \\ & + 2(\operatorname{diag} b_1)(S_1 - I) \\ & \times G\{(S_1 - I)\odot(S_1 - I)\}' \\ & + 2\{(S_1 - I)\odot(S_1 - I)\} \\ & \times G(S_1 - I)'(\operatorname{diag} b_1) \\ & + 2\{(S_1 - I)V(S_1 - I)'\} \\ & \odot \{(S_1 - I)V(S_1 - I)'\} \\ & + 4\{(S_1 - I)V(S_1 - I)'\} \\ & \odot (b_1 b_1')]S_2'/\{(\mathbf{1} + S_2\Delta)(\mathbf{1} + S_2\Delta)'\}.\end{aligned}$$

The proof is given in the Appendix.

The expression for $\operatorname{cov}(\hat{v}|\chi)$ simplifies considerably if normality of the errors can be assumed.

Corollary 1.1. If the errors ε_i are normally distributed, then

$$\operatorname{cov}(\hat{v}|\chi) = \frac{2S_2[\{(S_1 - I)V(S_1 - I)'\} \odot \{(S_1 - I)V(S_1 - I)' + 2b_1 b_1'\}]S_2'}{(\mathbf{1} + S_2\Delta)(\mathbf{1} + S_2\Delta)'}.$$

Remark 1. The conditional mean average squared error (MASE) of $\hat{v}$ is defined as

$$\operatorname{MASE}(\hat{v}) = n^{-1}E\left[\sum_{i=1}^{n}\{\hat{v}(X_i) - v(X_i)\}^2|\chi\right].$$

Noting that $\operatorname{MASE}(\hat{v}) = n^{-1}\{\|E(\hat{v}|\chi) - v\|^2 + \operatorname{tr}\operatorname{cov}(\hat{v}|\chi)\}$, where $\|x\|^2 = x'x$, one can use the preceding results to find exact expressions for $\operatorname{MASE}(\hat{v})$ for *any* pair of smoother matrices S_1 and S_2.

2.2 Asymptotics for Local Polynomial Variance-Function Estimators

In practice, the X_i's can be either fixed or random, and in the latter case they need be neither independent nor identically distributed. In fact, all the results in Section 2.1 are conditional on the X_i's and so do not depend on their distribution. Asymptotics, however, require some assumptions about the behavior of the X_i's as $n \to \infty$.

The simplest assumption, and the one we will use in this section, is that the X_i's are iid. Let f denote the common density of $X_1, \ldots, X_n$ and the function η be given by $\eta(X_i) = \text{var}(\varepsilon_i^2), i = 1, \ldots, n$. Define the function $K_{(p)}(u) = \{|M_p(u)|/|N_p|\}K(u)$, where 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 $M_p(u)$ is the same as N_p with the first column replaced by $(1, u, \ldots, u^p)$. $K_{(p)}$ is a pth-order kernel (Ruppert and Wand 1994).

Theorem 2. Suppose that x is an interior point of the support of f, m has $p_1 + 2$ continuous derivatives, v has $p_2 + 2$ continuous derivatives, and f and η are differentiable in a neighborhood of x, and that $h_1, h_2 \to 0, nh_1, nh_2 \to \infty$, and

$$\{h_1^{2(p_1+1)} + (nh_1)^{-1}\} = o(h_2^{p_2+1}) \tag{6}$$

as $n \to \infty$. Then, for p_2 odd,

$$E\{\hat{v}(x) - v(x)|\chi\} = \left\{\int u^{p_2+1}K_{(p_2)}(u)\,du\right\} \times \left\{\frac{v^{(p_2+1)}(x)}{(p_2+1)!}\right\} h_2^{p_2+1} + o_P(h_2^{p_2+1})$$

and, for p_2 even,

$$E\{\hat{v}(x) - v(x)|\chi\} = \left\{\int u^{p_2+2}K_{(p_2)}(u)\,du\right\} \times \left\{\frac{v^{(p_2+1)}(x)f'(x)}{f(x)(p_2+1)!} + \frac{v^{(p_2+2)}(x)}{(p_2+2)!}\right\} h_2^{p_2+2} + o_P(h_2^{p_2+2}).$$

In either case

$$\text{var}\{\hat{v}(x)|\chi\} = \left\{\int K_{(p_2)}(u)^2\,du\right\} \times \{n^{-1}h_2^{-1}\eta(x)/f(x)\} + o_P\{(nh_2)^{-1}\}.$$

Once again, we defer the proof to the Appendix.

Remark 2. The leading terms depend only on the bandwidth h_2, indicating that the initial bandwidth h_1 has only a second-order effect on the asymptotic performance of $\hat{v}(x)$. If $p_1 = p_2$ and if h_1 and h_2 are chosen optimally for estimation of m and v, then $h_1^{2(p_1+1)}$ and $(nh_1)^{-1}$ will be of the same order as $n \to \infty$ and both will be $o_P(h_2^{p_2+1})$ so that (6) is satisfied.

Remark 3. Comparison with theorem 4.1 of Ruppert and Wand (1994) shows that the leading bias and variance terms for our local polynomial variance estimator are analogous to those for the local polynomial estimator of the mean function. The only difference is that the asymptotic bias depends on derivatives of v rather than m, and the asymptotic variance of $\hat{v}(x)$ is proportional to the variance of the squared errors, rather than the Y_i's.

Here is an important result: Asymptotically, $\hat{v}$ behaves like a local polynomial smooth of the (unobservable) ε_i^2's; that is, v can be estimated as well as if m were known so that there is no loss in asymptotic efficiency due to estimating m. This result tells us that the estimate of the variance function based on squared residuals is "adaptive" in the sense of Bickel (1982). The result also has important implications for bandwidth selection because it justifies applying standard bandwidth selectors developed to estimate the mean function to smoothing the squared residuals to estimate the variance function. Thus, new bandwidth selectors for the variance function are not needed; see the next section.

Remark 4. One could also rework the steps used to prove Theorem 2 for the situation in which x is converging to the boundary of the support of f to show that, for odd p, the local polynomial variance estimator induces an automatic "boundary kernel-type" correction. This attractive feature has been pointed out in the mean estimation context by, for example, Fan and Gijbels (1992), Hastie and Loader (1993), and Ruppert and Wand (1994).

Remark 5. The automatic boundary correction and design adaptivity of odd-degree p has led some authors—for example, Fan and Gijbels (1995)—to state that p should be taken to be odd. We do *not* make such a general recommendation. As Cleveland and Loader (1996, sec. 10) pointed out, the superior boundary bias correction and design adaptivity of odd-degree p compared to even-degree are asymptotic properties, so their relevance to statistical practice must be assessed by finite-sample results. As both Ruppert and Wand (1994) and Cleveland and Loader (1996) argued, increasing an even-degree p by 1 to get an odd degree will substantially increase variance at the boundaries, even though there is no increase in asymptotic variance in the interior. Our experience with real and simulated data is that local linear regression ($p = 1$) is, in fact, usually superior to kernel regression ($p = 0$). When $p > 1$, however, the case for odd-degree p is not so clear. Cleveland and Devlin (1988) used $p = 2$ quite successfully in some examples. Ruppert (1995) had a detailed simulation study of $m(x) = x + \exp(-16x^2)$ with the X_i's iid Uniform$(-2, 2)$. This function has substantial curvature changing from convex to concave and then back to convex. One might expect a local cubic fit to outperform a local quadratic fit because a quadratic polynomial has constant curvature. Local quadratic fitting, however, performs as well in the interior as local cubics and outperforms local cubics near the boundaries, where the higher boundary variance of local cubics becomes a serious problem.

2.3 Bandwidth Choice and Choosing the Degrees of the Local Polynomial Fits

An important practical problem is the choice of the bandwidths. One may use either local bandwidths, where h_1 and h_2 are functions of x, or global bandwidths that do not depend on x. For concreteness, let us assume that the bandwidths are local. Ideally, one would choose both h_1 and h_2 to minimize the mean squared error (MSE) of $\hat{v}$ at the point x. This is difficult to do in practice, however, because the effects of h_1 on the MSE of $\hat{v}$ are of second order and therefore difficult to estimate.

Using Theorem 2 and Remark 3, we suggest an alternative strategy that will produce asymptotically optimal bandwidths. First, use a local bandwidth selector to find asymp-

totically optimal h_1 for estimation of $m(x)$. One could, for example, use the bandwidth selector of Fan and Gijbels (1995), though in the example of Section 4 we use the empirical bias bandwidth selection (EBBS) method of Ruppert (1995). Next, treat the squared residuals as if they were the squared ε's, and apply the same bandwidth selector used for estimation of the mean function to the squared residuals. If one uses $p_1 \geq p_2$, then (6) will be satisfied.

As a rule of thumb, when estimating m and v, not their higher derivatives, we recommend $p_1 = 2$ (or perhaps 3) and $p_2 = 1$. Generally, v does not have strong curvature, and a local linear fit for $\hat{v}$ suffices. Often m has sufficient curvature that a local linear fit to m will have enough bias to significantly inflate the squared residuals, resulting in upward bias in v. This bias can even overwhelm $\hat{v}$ if v is small; that is, $\hat{v}$ might be mostly an estimate of squared bias in regions where m has strong curvature if $\hat{m}$ is local linear. Using a local quadratic fit to m can be a big help in such cases.

2.4 Extension to Multivariate Predictors

As Cleveland and Devlin (1988) demonstrated in their examples, local polynomial regression can be quite successful with two or more predictors. In principle, extension of the formulation and theory of the general class of variance estimators to multivariate predictor variables is straightforward. The expressions for $\hat{v}$ at (2) are the same except that the rows of the smoother matrices S_1 and S_2 correspond to X_i's that live in higher-dimensional space rather than on the real line. Theorem 1 continues to hold in the multivariate case.

3. PIECEWISE POLYNOMIAL BINNING

This section contains a computational method that is particularly well suited for larger datasets—for example, the turbulence dataset of Section 4.4, which has 20,000 observations. We only describe the implementation for univariate X_i.

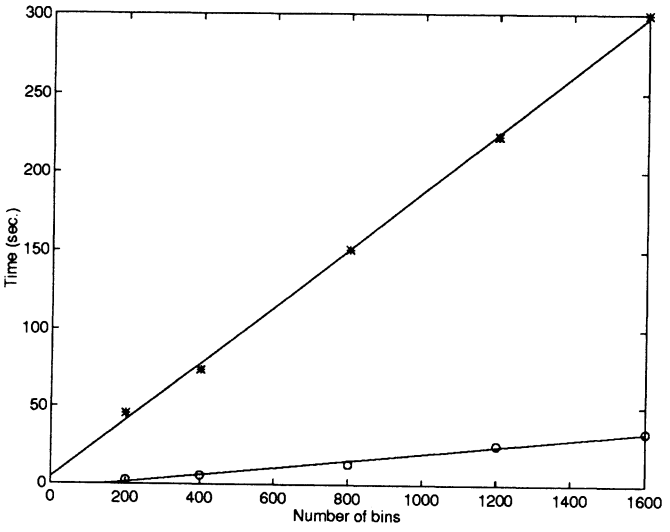

Figure 2. Binning Times (circles) and Smoothing Times (asterisks) as a Function of the Number of Bins, With Linear Regression Fits.

First, the data are binned according to their x values into n_{bin} disjoint subsets with roughly equal number of observations per subset. (Alternatively, one could use equal length bins.) For the jth bin, $j = 1, \ldots, n_{\text{bin}}$, let $\bar{x}_j$ be the mean of the X_i's in that bin. Fit a p_bth-degree polynomial to the data in the jth bin. Let $\bar{y}_j$ be the fitted value of this model at $\bar{x}_j$, and let $\bar{v}_j$ be the residual mean square from the model. Using the residual mean square induces the proper degrees-of-freedom correction of the bias induced by using $\hat{m}$ in place of m when computing the residuals. Therefore, if m is a p_bth-degree polynomial and if v is constant on the jth bin, then $\bar{y}_j$ and $\bar{v}_j$ are unbiased estimators of $m(\bar{x}_j)$ and $v(\bar{x}_j)$.

Because the bins are nonoverlapping, $\{\bar{y}_1, \ldots, \bar{y}_{n_{\text{bin}}}\}$ are mutually independent, as are $\bar{v}_1, \ldots, \bar{v}_{n_{\text{bin}}}\}$. To estimate m, apply any linear smoother and bandwidth selector combination desired to the data $(\bar{x}_i, \bar{y}_i)$ and do the same to $(\bar{x}_i, \bar{v}_i)$ to estimate v. No degrees-of-freedom correction of bias due to estimation of the mean is needed here, because the correction was made at the binning stage.

The idea is to choose n_{bin} and p_b so that $\bar{y}_i$ and $\bar{v}_i$ from the binning stage are very undersmoothed estimators of $m(\bar{x}_i)$ and $v(\bar{x}_i)$, respectively. Thus, the number of observations per bin should be small, though it must of course be at least $p_b + 2$ so that the residual degrees of freedom are positive and should be at least twice this minimum for good efficiency of $\hat{v}$. The correct degree of smoothing is done at the smoothing stage that follows binning.

Using $p_b = 1$ will give accuracy similar to the popular linear binning technique, but $p_b > 1$ will be more accurate than binning techniques now in the literature and will allow a smaller value of n_{bin}. Piecewise polynomial binning has the attractive property of reproducing polynomials. More precisely, if the (X_i, Y_i) fall exactly on a pth-degree polynomial and $p_b \geq p$, then the $(\bar{x}_j, \bar{y}_j)$ fall on the same polynomial. Therefore, if one smooths the binned data by local polynomial regression of degree at least p, then the smoothed fit will also fall on the same polynomial as the original data.

Binning is needed if one uses a computationally intensive bandwidth selector such as EBBS (Ruppert 1995) used in our examples or cross-validation that requires that the smoothing be done for many values of the bandwidth. The binning is done only once, and then all calculations needed to compute the bandwidth need only be done with the much smaller binned dataset. For example, in the dataset of Section 4.4, there are 20,000 observations. Using a SPARC 20 and a MATLAB program written by the authors, binning takes 2.5, 5.8, 13.1, 25.5, and 34.0 seconds (clock, not central processing unit time) with n_{bin} equal to 200, 400, 800, 1,200, and 1,600, respectively; that is, the time is roughly proportional n_{bin}. The remaining calculations needed to obtain $\hat{m}$ and $\hat{v}$ on an 80-point grid, including calculating their bandwidths, used 45, 74, 151, 223, and 300 seconds for the same values of n_{bin}. These times are plotted in Figure 2 with linear regression fits. Extrapolating the linear fit to the smoothing times gives a time of 3,670 seconds (about 61 minutes) for 20,000 bins or, equivalently in computational effort, the raw data. Thus, binning 20,000 raw data points

to 200 binned data points converts a computation time from about one hour to one of less than a minute.

Piecewise polynomial binning requires little additional programming and seems well worth the extra effort for large datasets, say $n > 1{,}000$. We do not use binning in our example (LIDAR) with 221 observations.

An alternative form of binning presented by Turlach and Wand (1996) requires equally spaced bin centers and, at the expense of somewhat more programming effort, further reduces computational time.

4. EXAMPLES

4.1 Bandwidths

As we have argued, the theory in Section 2 suggests that any bandwidth suitable for estimating m or its derivatives is suitable for smoothing the squared residuals to estimate v or its derivatives. In our examples we use the EBBS bandwidth selector proposed by Ruppert (1995). EBBS (empirical bias bandwidth selection) has the following advantages:

1. The bandwidth is local—that is, can depend on x. EBBS minimizes an estimate of MSE at x. The estimate of MSE takes into account boundary effects.
2. Estimation of derivatives can be accommodated.
3. Both odd- and even-degree polynomials can be used. In contrast, plug-in bandwidths such as those of Ruppert, Sheather, and Wand (1995) and Fan and Gijbels (1995) restrict to odd-degree because the bias of even-degree local polynomial fitting is more complex.
4. Exact formulas rather than asymptotic approximations are used as much as possible. In particular, there is no assumption that the X_i's have a probability density—exact formulas use only their sample values.

Suppose that for some $r \geq 0$ we wish to estimate $m^{(r)}(x)$ for all x on some grid. Let $\text{MSE}(h;x)$ be the MSE of $\hat{m}^{(r)}(x)$ using bandwidth h. MSE, bias, and variance are taken to mean conditional given $X_1, \ldots, X_n$. To estimate $\text{MSE}(h;x)$, we estimate the variance and bias separately. These are denoted by $\text{var}(h;x)$ and $\text{bias}(h;x)$. An exact formula for $\text{var}(h;x)$ exists (Ruppert and Wand 1994), and we substitute an estimate $\hat{v}(x)$ for v in this formula. Asymptotics (Ruppert and Wand 1994) suggest modeling the bias with the equation

$$E\hat{m}^{(r)}(x;h) = b_0 + b_{p+1-r}h^{p+1-r} + \cdots + b_{p+1-r+t}h^{p+t-r} \quad (7)$$

for some $t \geq 1$, with $t = 1$ or 2 recommended. Here we write $\hat{m}^{(r)}(x;h)$ instead of $\hat{m}^{(r)}(x)$ to denote dependence on the bandwidth h. To estimate $\text{bias}(h_0;x)$ at some h_0, we calculate $\hat{m}^{(r)}(x;h)$ for values $\{h_j\}_{j=1}^M, M \geq t+1$, in a neighborhood of h_0. Then we fit model (7) by least squares to the "data" $\{(h_j, \hat{m}^{(r)}(x;h_j))\}_{j=1}^M$ and use $\hat{b}_0, \ldots, \hat{b}_{p+t-r}$ to estimate $\text{bias}(h_0;x)$. Using $\widehat{\text{MSE}}(h_0;x) = \widehat{\text{var}}(h_0;x) + \text{bias}^2(h_0;x)$, we can estimate MSE at any fixed values of h and x. Model (7) is our only use of asymptotics, and we estimate the coefficients in (7) directly rather than by plugging estimates into formulas for the asymptotic values of these coefficients.

For fixed x we estimate MSE at a grid of h values, say 12 values between $\text{span}(x, .1)$ and $\text{span}(x, 1)$. Here $\text{span}(x, q)$ is the smallest value of h such that at least $100q\%$ of $X_1, \ldots, X_n$ are with h units of x. [The idea of a span is borrowed from Cleveland's (1979) LOWESS.] We then let $\tilde{h}(x)$ minimize $\widehat{\text{MSE}}(h;x)$ over this grid of h values. Thus, $\tilde{h}(x)$ is a local bandwidth that attempts to minimize MSE at each x on a grid. In many, if not most, datasets, $\tilde{h}(x)$ will be rather variable, so we suggest kernel smoothing of $\tilde{h}(x)$ over x. In our examples, we use a triangular kernel with bandwidth giving a span equal to a user-chosen tuning parameter, BANDSPAN, on the grid of x values, where $\tilde{h}(x)$ has been calculated. Experimentation by Ruppert (1995) suggested that the value of BANDSPAN is not too critical, and BANDSPAN of 4 to 8 when using a 50- to 100-point grid of x values can be recommended. Let $\hat{h}(x)$ be the smooth of $\tilde{h}(h)$. We compute $\hat{m}^{(r)}(x;\hat{h}(x))$ on the same grid of x values as where $\hat{h}(x)$ has been found. For calculation of residuals, we use cubic spline interpolation of $\hat{m}^{(r)}(x;\hat{h}(x))$ from the x grid to $X_1, \ldots, X_n$. Unless n is small, say less than 100, direct computation of $\hat{m}^{(r)}(x;\hat{h}(x))$ for $x = X_1, \ldots, X_n$ would be quite slow because computation of $\hat{h}(x)$ is moderately intensive. Interpolation is another technique borrowed from Cleveland (1979).

To estimate both m and v using EBBS, we recommend a three-step algorithm:

1. Estimate m using a "small" fixed span.
2. Estimate v by smoothing squared residuals. Assume that the errors have a constant kurtosis so that the variance of the squared errors is proportional to their mean squared. This assumption allows one to avoid estimating the variance function of the squared errors.
3. Estimate m using EBBS and the estimated variance function.

Steps 2 and 3 could be iterated, though we have found that this is generally not necessary. One can check that the span in Step 1 is sufficiently small by seeing whether it is smaller than the span chosen in Step 3 for $\hat{m}$. For more details of the EBBS method, see Ruppert (1995).

There are many other possible bandwidth selectors—for example, generalized cross-validation and related estimators; see Hastie and Tibshirani (1990). Many of these bandwidth selectors are somewhat simpler than EBBS and could be used with the local polynomial variance-function estimators we are proposing. We certainly do not recommend against them, especially if simplicity is of primary concern. Most of the simpler bandwidth selectors produce a global smoothing parameter, however, rather than a locally adaptive bandwidth. Moreover, most selectors target the bandwidth optimal for estimating m (or σ^2), not a derivative. For these two reasons, we have focused attention on EBBS.

4.2 A Simulation Study

We performed a small simulation to assess the sensitivity of $\hat{v}$ to the initial bandwidth used to estimate m. Asymp-

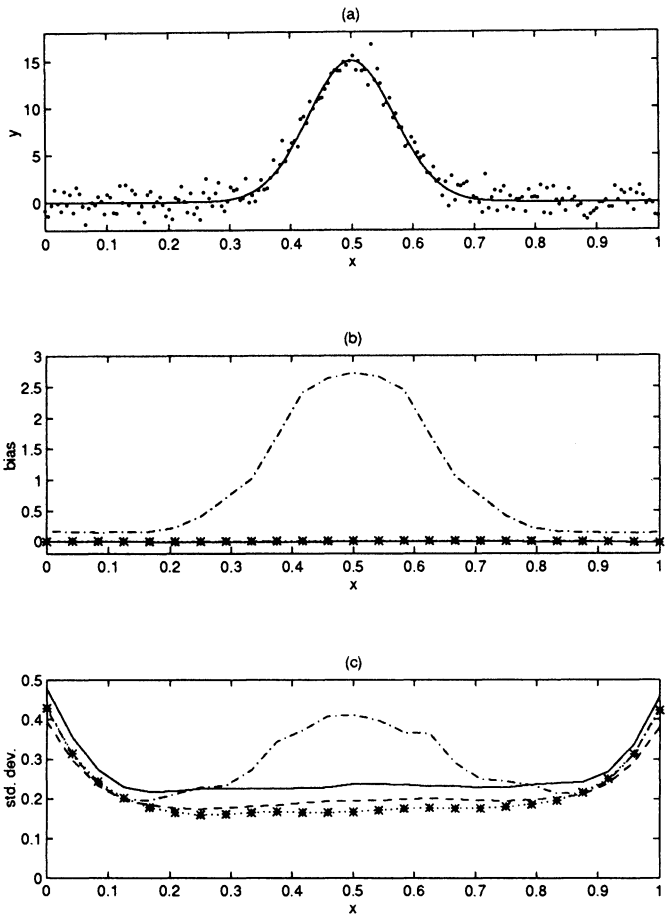

Figure 3. Simulation: (a) Typical Dataset; (b) Bias of $\hat{v}$ When Initial Estimate of m Uses Span = .05 (solid), Span = .15 (dashed), Span = .45 (dot and dashed), and EBBS (dotted with asterisks) (the solid, dashed, and dotted lines are virtually indistinguishable); (c) Standard Deviation of $\hat{v}$ With Line Types as in (b).

totics suggest that the sensitivity will be small unless the initial bandwidth is too large, causing bias. This is precisely what we found in this finite-sample study.

There were 500 replicate datasets, each of $n = 200$ observations with the X_i's equally spaced on [0, 1]. Given X_i, we generated Y_i by $Y_i = 25\exp\{-100(X_i - .5)^2\} + \varepsilon_i$ with ε_i standard normal. Note that $v \equiv 1$. The reason for this choice of v is that we are studying bias in $\hat{v}$ due to bias in $\hat{m}$, not bias in $\hat{v}$ due to curvature in v. The later type of bias is analogous to bias in $\hat{m}$, which has been well studied. Figure 3(a) shows a typical dataset.

There were four choices of the initial bandwidth for estimating m, span = .05, span = .15, span = .45, and the EBBS bandwidth using the algorithm of the previous section with span = .05 in Step 1. The EBBS bandwidth for $\hat{m}$ minimized the estimated MSE over the range corresponding to span = .05 to span = 1. In all four cases, the preliminary estimator of m was local quadratic, and the estimator of v was local linear smoothing of squared residuals with correction for bias due to estimation of the mean as discussed in Section 1.1. EBBS was used in all cases to choose the bandwidth for the squared residuals, with the bandwidth restricted to the range span = .05 to span = 1.

The estimate $\hat{v}$ was calculated at 25 grid points equally spaced between 0 and 1. Bias and standard deviation of $\hat{v}$ were estimated at each of these grid points using the 500 replicates. These are plotted in Figure 3, (b) and (c). The dotted curves of the EBBS estimates have asterisks at the 25 grid points.

From (b) and (c) we see that bias is very small compared to the standard deviation of the estimate unless the bandwidth for estimation of m is too large, say span = .45. In that case, the bias in $\hat{m}$ inflates the squared residuals and masquerades as extra variability. The standard deviation of $\hat{v}$ is smaller using an EBBS bandwidth for m than for any of the fixed-span estimates used in the study.

4.3 LIDAR Data

The LIDAR technique has proven to be an efficient tool in monitoring the distribution of meteorological parameters and several atmospheric species of importance; see Sigrist (1994).

The received signal power $P(\lambda, x)$ as function of range x and wavelength λ is described by the deterministic single-scattering LIDAR equation

$$P(\lambda, x) = \frac{k(\lambda, x)}{x^2}\beta(\lambda, x) \times \exp\left\{-2\int_0^x (\sigma(\lambda)N(s) + \alpha(\lambda, s))\, ds\right\},$$

where $k(\lambda, x)$ is an instrument factor, $\beta(\lambda, x)$ the backscattering coefficient, $N(s)$ the concentration of the studied species at distance s, $\sigma(\lambda)$ the absorption cross-section for wavelength λ, and $\alpha(\lambda, s)$ the attenuation due to general scattering and absorption of the aerosol.

The DIAL technique (DIfferential Absorption Lidar) employs two different wavelengths, one in resonance with an absorption line of the species of interest and the other off resonance, denoted λ_{on} and λ_{off}, respectively. After differentiating the logarithm of the ratio between the two signals, we get

$$\int_0^x N(s)\, ds = -\frac{1}{2\sigma_{\text{diff}}}\ln\frac{P(\lambda_{\text{on}}, x)}{P(\lambda_{\text{off}}, x)} + \frac{1}{2\sigma_{\text{diff}}}\ln\frac{\beta(\lambda_{\text{on}}, x)}{\beta(\lambda_{\text{off}}, x)} - \frac{\alpha_{\text{diff}}}{\sigma_{\text{diff}}},$$

where $\alpha_{\text{diff}} = \int_0^r (\alpha(\lambda_{\text{on}}, s) - \alpha(\lambda_{\text{off}}, s)), ds$ is the differential attenuation due to the general scattering and absorption in the atmosphere, and $\sigma_{\text{diff}} = \sigma(\lambda_{\text{on}}) - \sigma(\lambda_{\text{off}})$ is the differential absorption cross-section. The instrument factors for the two wavelengths are assumed to be the same.

The basic idea of the DIAL concept is that the second and third term can be neglected if the two wavelengths are close together. In its simplest form, the DIAL equation is reduced to

$$N(x) = -\frac{1}{2\sigma_{\text{diff}}}\frac{\partial}{\partial r}\ln\frac{P(\lambda_{\text{on}}, x)}{P(\lambda_{\text{off}}, x)}.$$

One restriction with the DIAL technique is that the return signals are quite often weak, which makes a concentration profile from a single-pulse pair very noisy. Temporal averaging with up to several hundred shots is therefore generally used to improve the signal-to-noise ratio.

The available observations corresponding to the different measuring distances, $x_1, \ldots, x_n$, are averages of a number of shots M_s,

$$P(\lambda_{\text{on}}, x_i) = \frac{1}{M_s} \sum_{j=1}^{M_s} P^{(j)}(\lambda_{\text{on}}, x_i)$$

$$P(\lambda_{\text{off}}, x_i) = \frac{1}{M_s} \sum_{j=1}^{M_s} P^{(j)}(\lambda_{\text{off}}, x_i),$$

where M_s is approximately 100. Now let

$$Y(x_i) = \ln \frac{P(\lambda_{\text{on}}, x_i)}{P(\lambda_{\text{off}}, x_i)}.$$

This means that the concentration of mercury at range x is proportional to the first derivative of the regression function with proportional constant $C = -1/2\sigma_{\text{diff}} = -(1/16)10^6$ ng/m^2.

The results are illustrated in Figure 1. In (a) we have the raw data. First we used a local quadratic estimate of the mean with a fixed span of .05. In this example and the next, we use the Epanechnikov kernel, which is $K(x) = (3/4)(1-x^2)I\{|x| \leq 1\}$. Squared residuals from this fit are in (b).

In (c) we have a local linear smooth of these squared residuals using EBBS and computed on a 50-point equally spaced grid with BANDSPAN = 4. In EBBS, the tuning parameters (see Sec. 5.1) were BANDSPAN $= 4, t = 2$, and $M = 5$. The solid curve in (c) is the corrected estimate, and the dotted curve is the uncorrected estimate. The correction is not sizable, but it does increase the estimated variance as expected.

In (d) and (e) we have a local quadratic estimate of m and concentration $(=Cm')$ on a 50-point x grid, respectively, using EBBS with the estimate $\hat{v}$ in (c). The bandwidths for estimating m (solid), v (dotted), and m' (dashed) are shown in (f).

The squared residuals in Figure 1(b) suggest that v might be bimodal. Our local bandwidth selector, however, chooses bandwidths large enough to smooth away the bimodality, perhaps suggesting that the apparent bimodality is merely a chance phenomenon and, in fact, v is monotonically increasing. Of course, our methodology is not designed to test for bimodality, and if bimodality were an important issue here, we would want to use a technique specifically designed to test for it.

Finally, in (g) and (h) we have the standardized residuals and their absolute values, respectively, where the ith standardized residual is $\{Y_i - \hat{m}(X_i)\}\hat{v}^{-1/2}(X_i)$. Plotting absolute residuals is useful for detecting heteroscedasticity (Carroll and Ruppert 1988). Neither plot shows any apparent pattern, suggesting that the residuals have a constant mean of 0 and a constant variance so that the estimates of m and v are satisfactory.

4.4 Turbulence Data

In this example we look at an especially difficult problem because v'' must be estimated at the boundary. In this study, spatial position is reduced to one dimension because the quantities of interest depend on space in only one direction. We have bivariate data (X_i, Y_i), where X_i is position and Y_i is velocity of a particle.

These data are part of a "feasibility study" by mechanical engineers at Cornell to see whether certain quantities of interest can be accurately estimated by the Monte Carlo pdf model of velocity. The data do *not* come from an actual simulation of the pdf model. Instead, the mean and variance functions, m and v, were found by Taylor series approximations to the deterministic Reynolds-stress model.

The Lagrangian pdf, $f_L(Y, X; t)$, is the joint density function of velocity and position at time t. This pdf evolves according to a partial differential equation; see equation (46) of Dreeben and Pope (1995). Taking first and second moments with respect to velocity or position of all terms in this equation gives equations for the first two moments of velocity and position. These equations can be solved, at least approximately, though not in closed form. Numerical solutions give m and v, which are shown in Figure 4.

To generate data, 20,000 values, $\{X_i: i = 1, \ldots, 20{,}000\}$, were taken uniformly distributed on [0, .1] and at each X_i, Y_i were generated from Model (1) with ε_i normally distributed.

The idea is that these data will be similar to what would be obtained if a stochastic simulation of the turbulence model were programmed and run.

The engineers wanted to know if the second derivative of v at the left boundary—for example, $v''(0)$, could be estimated accurately in the pdf method. This quantity is of special interest because it is a boundary condition on turbulent dissipation. The pdf method of simulations steps through time, with the boundary condition at one time step being estimated with data from the previous time step. The left boundary corresponds to a real physical boundary, so it is not possible to have x negative; this makes estimation of $v''(0)$ difficult. Although v is only an approximation to the "true" variance function, it is the "population" variance function that generated these data. If $v''(0)$ can be estimated accurately here, the engineers feel that the second derivative of the "true" v can be accurately estimated later with data from a stochastic simulation of the pdf model.

We implemented the piecewise polynomial binning described in Section 3 with $n_{\text{bin}} = 200$ (100 observations/bin) and $p_b = 2$ (piecewise quadratic binning). The residual mean squares are plotted in Figure 4(a) as a function of $\bar{x}$. Figure 4(b) is a plot of a local quadratic smooth of the data in (a) (solid) and the v (dashed).

In (d) we have $\hat{v}''$ (solid) using local cubic smoothing as discussed in Section 1.4 and with the EBBS bandwidth shown in (e). Because only $v''(0)$ is of interest, we restrict x to the range [0, .003], which includes 0 and the two x grid points to the right of 0. At this scale, the piecewise linear nature of the plot is obvious. If desired, the estimate of v'' on the 80-point grid could be cubicly interpolated to a finer grid before plotting. The true v'' (dashed) is also shown in (d).

As can be seen in Figure 4, $v(0) = 0$. In fact, $v'(0)$ is also 0. As a referee has mentioned, the prior knowledge

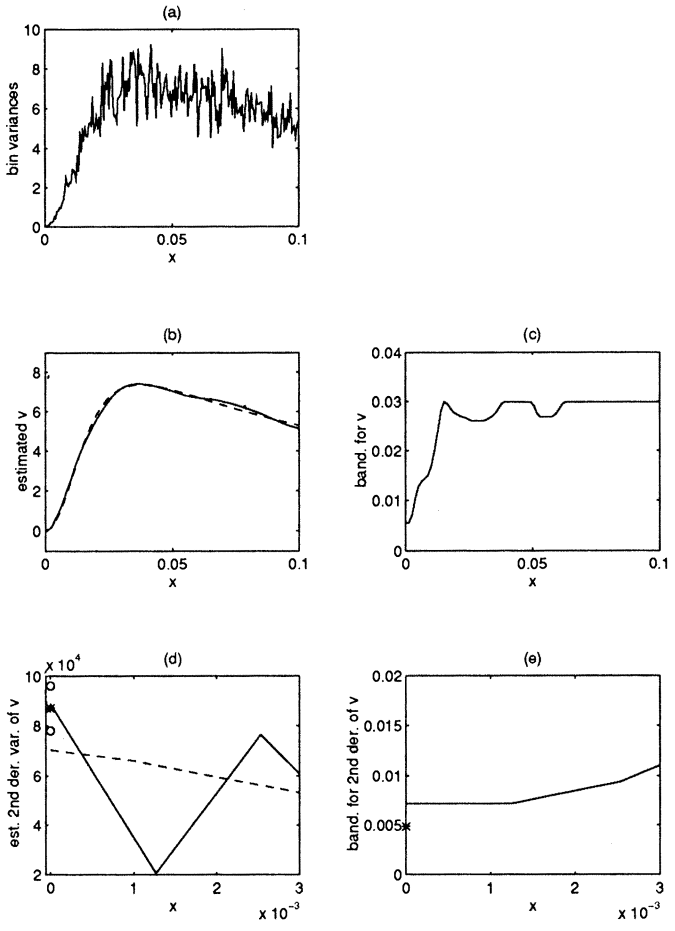

Figure 4. Turbulence Data: (a) Residual Mean Squares Plotted Against Bin Means of x; (b) Local Quadratic Smooth of Data in (a) (solid) and True v (dashed); (c) Local EBBS Bandwidth Used in (b); (d) Local Cubic Estimate of v'' Without Constraints (solid), With Intercept and Linear Coefficient Constrained to Be 0 (asterisk), and True v'' (dashed) [also, constrained estimate plus and minus two standard errors (open circles)]; (e) Local EBBS Bandwidths Used in (d), Unconstrained Estimation (solid) and Constrained Estimation (asterisk).

that $v(0) = v'(0) = 0$ can be used to improve the accuracy of $\hat{v}''(0)$ by dropping the intercept and linear term of the local polynomial fit at 0 to the squared residuals. Because dropping these terms simply changes one linear model into another, the EBBS bandwidth selector applies with the appropriate modification of the exact formula for $\text{var}\{\hat{v}''(0;h)\}$. In (d) the asterisk is $\hat{v}''(0)$ from a local fit with only the quadratic and cubic terms, which we will call the constrained estimator. Using the constrained estimator can considerably improve $\hat{v}''(0)$ because the standard error decreases by a factor of 2.25. In this particular sample, however, the constrained estimate did not change much from the usual estimator. We experimented somewhat with the tuning parameters and found that the constrained estimator was much more stable when the tuning parameters were varied. The constrained estimator plus and minus two standard deviations is shown in (d) as open circles. The bias in the constrained estimator is evident; bias, of course, is unavoidable in nonparametric estimation. The EBBS bandwidth for the constrained estimator is shown as an asterisk in (e). This bandwidth is a bit too large because a somewhat smaller bandwidth will have less bias without increasing the standard error too much. Thus, EBBS has chosen a reasonably good bandwidth but not the best possible. Because EBBS is the only bandwidth proposal applicable to this situation, however, by default it is the best available technology. Moreover, estimation of a best local bandwidth for estimating a second derivative is an inherently difficult problem, so EBBS might be doing about as well as possible here.

ACKNOWLEDGMENTS

We thank Stephen Pope and Tom Dreeben for supplying the turbulence data and for helpful discussions and Pär Ragnarson and Hans Edner for supplying the LIDAR data. We also thank the referees and associate editor for very useful suggestions.

APPENDIX: PROOFS OF THEOREMS

A.1 Proof of Theorem 1

First note that

$$\hat{v} = \frac{S_2 \text{ diag}\{(S_1 - I)YY'(S_1 - I)'\}}{\mathbf{1} + S_2\Delta}.$$

For the bias we have

$$\begin{aligned} E(\hat{v}|\chi) &= \frac{S_2 \text{ diag}\{(S_1 - I)(mm' + V)(S_1 - I)'\}}{\mathbf{1} + S_2\Delta} \\ &= \frac{S_2\{ \text{ diag}(b_1 b_1') + v + \text{ diag}(S_1 V S_1' - 2S_1 V)\}}{\mathbf{1} + S_2\Delta}. \end{aligned}$$

Direct algebra then leads to the stated result.

The result for $\text{cov}(\hat{v}|\chi)$ depends heavily on the following lemma.

Lemma 1. Let Y be a random vector having all entries independent. Define $m = E(Y), V = \text{diag}[E\{(Y - m)^2\}], G = \text{diag}[E\{(Y - m)^3\}]$, and $T = \text{diag}[E\{(Y - m)^4\}]$. Then for any square constant matrix A having the same number of rows as Y,

$$\begin{aligned} \text{cov}\{(AY)^2\} &= (A \odot A)(T - 3V^2)(A \odot A)' \\ &\quad + 2\{\text{diag}(Am)AG(A \odot A)' \\ &\quad + (A \odot A)GA' \text{ diag}(Am)\} + 2(AVA') \\ &\quad \odot (AVA') + 4(AVA') \odot \{(Am)(Am)'\}. \end{aligned}$$

Proof. We will use the tensor notation and results of McCullagh (1987). Let a_{ij} denote the (i,j) entry of A. A generalized cumulant of the set of random variables $(Y_1, \ldots, Y_n)$ is an ordinary cumulant of random variables formed by taking products from this set. Generalized cumulants will be denoted using partitioned superscript notation. For example, $\kappa^i = \text{cum}(Y_i) = E(Y_i), \kappa^{i,j} = \text{cum}(Y_i, Y_j) = \text{cov}(Y_i, Y_j)$, and $\kappa^{i,j,kl} = \text{cum}(Y_i, Y_j, Y_k Y_l)$. There are many formulas relating generalized cumulants to ordinary cumulants and to moments; see McCullagh (1987).

The (m,n) entry of $\text{cov}\{(AY)^2\}$ is easily shown to be

$$\text{cov}\{(AY)^2\}_{mn} = \sum_i \sum_j \sum_k \sum_l a_{mi} a_{mj} a_{nk} a_{nl} \kappa^{ij,kl}.$$

One of the fundamental identities for generalized cumulants, given on page 58 of McCullagh (1987), states that

$\kappa^{ij,kl} = \kappa^{i,j,k,l} + \kappa^i\kappa^{j,k,l} + \kappa^j\kappa^{i,k,l} + \kappa^k\kappa^{i,j,l} + \kappa^l\kappa^{i,j,k} + \kappa^{i,k}\kappa^{j,l} + \kappa^{i,l}\kappa^{j,k} + \kappa^i\kappa^k\kappa^{j,l} + \kappa^i\kappa^l\kappa^{j,k} + \kappa^j\kappa^k\kappa^{i,l} + \kappa^j\kappa^l\kappa^{i,k}$. This implies that, because of the mutual independence of the Y_i's,

$$\begin{aligned}\operatorname{cov}\{(AY)^2\}_{mn} &= \sum_i a_{mi}^2 a_{ni}^2 \kappa^{i,i,i,i} \\ &+ 2\sum_i\sum_j (a_{mi}a_{mj}a_{nj}^2\kappa^i\kappa^{j,j,j} + a_{mi}^2 a_{ni}a_{nj}\kappa^j\kappa^{i,i,i}) \\ &+ 2\sum_i\sum_j a_{mi}a_{mj}a_{ni}a_{nj}\kappa^{i,i}\kappa^{j,j} \\ &+ 4\sum_i\sum_j\sum_k a_{mi}a_{mk}a_{nj}a_{nk}\kappa^i\kappa^j\kappa^{k,k}.\end{aligned}$$

It is easily verified that the stated result follows from this.

The following lemma shows how covariance matrices are affected by elementwise multiplication. Its proof is quite trivial and is omitted.

Lemma 2. If a is a constant vector having the same length as Y, then $\operatorname{cov}(aY) = (aa') \odot \operatorname{cov}(Y)$. The result for $\operatorname{cov}(\hat{v}|\chi)$ follows immediately from Lemmas 1 and 2 and the well-known result: $\operatorname{cov}(AY) = A \operatorname{cov}(Y)A'$.

A.2 Proof of Theorem 2

For an rth differentiable function g we let $g^{(r)} = [g^{(r)}(X_1), \ldots, g^{(r)}(X_n)]'$. We also use the convention that if U_n and W_n are n-dimensional random vectors and if c_n is a sequence of random variables, then $U_n = W_n + o_p(c_n)$ means that for each fixed $i, |U_n(i) - W_N(i)| = o_p(c_n)$ as $n \to \infty$ and similarly for $O_P(\cdot)$. The main stepping-stone for getting from Theorem 1 to Theorem 2 is Lemma 3.

Lemma 3. Suppose that the function g has $p+2$ continuous derivatives, that f is differentiable, and that $X_1, \ldots, X_n$ are each in the interior of the support of f. Assume that $h = h_n \to 0$ and $nh \to \infty$ as $n \to \infty$. Then

1. $$S_{p,h}g = \begin{cases} g + h^{p+1}\left\{\int u^{p+1}K_{(p)}(u)\,du\right\} \\ \quad \times \frac{g^{(p+1)}}{(p+1)!} + o_P(h^{p+1}) & p \text{ odd} \\ g + h^{p+2}\left\{\int u^{p+2}K_{(p)}(u)\,du\right\} \\ \quad \times \left\{\frac{g^{(p+1)}f'}{f(p+1)!} + \frac{g^{(p+2)}}{(p+2)!}\right\} + o_P(h^{p+2}) & p \text{ even,} \end{cases}$$

2. $$\operatorname{diag}\{S_{p,h}(\operatorname{diag} g)S_{p,h}'\} = (nh)^{-1}\left\{\int K_{(p)}(u)^2\,du\right\}(g/f) + o_P\{(nh)^{-1}\},$$

3. $\operatorname{diag}(S_{p,h}) = O_P\{(nh)^{-1}\}$,

and

4. $S_{p,h}(\operatorname{diag} g)S_{p,h}' = O_P\{(nh)^{-1}\}$.

Proof. Results (1) and (2) are direct consequences of theorem 4.1 of Ruppert and Wand (1994). Arguments similar to the ones employed there can be used to establish results (3) and (4).

Theorem 2 can be derived from Theorem 1 by repeated application of Lemma 3. For the conditional bias, Lemma 3 shows that the dominating term of (5) is $(S_2 - I)v$. Because the location of the X_i is arbitrary, the required result follows immediately.

The conditional variance result requires a little more algebra but is otherwise just as straightforward to derive. When the numerator of (6) is expanded out, the dominating terms are seen to be $S_2\{(T - 3V^2) + 2V^2\}S_2' = S_2 \operatorname{diag}(\eta)S_2'$, where $\eta = [\eta(X_1), \ldots, \eta(X_n)]'$. Application of (2) of Lemma 3 then leads to the desired result.

[Received September 1995. Revised January 1997.]

REFERENCES

Bickel, P. J. (1982), "On Adaptive Estimation," *The Annals of Statistics*, 10, 647–671.

Buckley, M. J., Eagleson, G. K., and Silverman, B. W. (1988), "The Estimation of Residual Variance in Nonparametric Regression," *Biometrika*, 75, 189–200.

Carroll, R. J. (1982), "Adapting for Heteroscedasticity in Linear Models," *The Annals of Statistics*, 10, 1224–1233.

Carroll, R. J., and Ruppert, D. (1988), *Transformation and Weighting in Regression*, New York: Chapman & Hall.

Cleveland, W. (1979), "Robust Locally Weighted Regression and Smoothing Scatterplots," *Journal of the American Statistical Association*, 74, 829–836.

Cleveland, W., and Devlin, S. (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 Loader, C. (1996), "Smoothing by Local Regression: Principles and Methods," in *Statistical Theory and Computational Aspects of Smoothing*, eds. W. Härdle and M. G. Schimek, Heidelberg: Physica Verlag, pp. 101–149.

Dreeben, T. D., and Pope, S. B. (1995), "Pdf and Reynolds-stress Modeling of Near-wall Turbulent Flows." unpublished manuscript.

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

Fan, J., and Gijbels, I. (1992), "Variable Bandwidth and Local Linear Regression Smoothers," *The Annals of Statistics*, 20, 2008–2038.

——— (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.

——— (1996), *Local Polynomial Modeling and Its Applications*, London: Chapman & Hall.

Gasser, T., Sroka, L., and Jennen-Steinmetz, C. (1986), "Residual Variance and Residual Pattern in Nonlinear Regression," *Biometrika*, 73, 625–634.

Hall, P., and Carroll, R. J. (1989), "Variance Function Estimation in Regression: The Effect of Estimating the Mean," *Journal of the Royal Statistical Society*, Ser. B, 51, 3–14.

Hall, P., Kay, J. W., and Titterington, D. M. (1990), "Asymptotically Optimal Difference-Based Estimation of Variance in Nonparametric Regression," *Biometrika*, 77, 521–528.

Hastie, T. J., and Loader, C. (1993), "Local Regression: Automatic Kernel Carpentry" (with discussion), *Statistical Science*, 8, 120–143.

Hastie, T. J., and Tibshirani, R. J. (1990), *Generalized Additive Models*, New York: Chapman & Hall.

Holst, U., Hössjer, O., Björklund, C., Ragnarson, P., and Edner, H. (1996), "Locally Weighted Least Squares Kernel Regression and Statistical Evaluation of LIDAR Measurements," *Environmetrics*, 7, 401–416.

Mathur, A. (1995), "On Estimation of Residual Variance Function," paper presented at the Joint Statistical Meetings, Orlando, Florida, August 13–17.

McCullagh, P. (1987), *Tensor Methods in Statistics*, London: Chapman & Hall.

Müller, H. G., and Stadtmüller, U. (1987), "Estimation of Heteroscedasticity in Regression Analysis," *The Annals of Statistics*, 15, 610–625.

Pope, S. B. (1985), "Pdf Methods for Turbulent Reactive Flows," *Progress in Energy and Combustion Science*, 11, 119–192.

Rice, J. (1984), "Bandwidth Choice for Nonparametric Regression," *The Annals of Statistics*, 12, 1215–1230.

Ruppert, D. (1995), "Empirical-Bias Bandwidths for Local Polynomial Nonparametric Regression and Density Estimation," unpublished

manuscript (TR 1137 at http://www.orie.cornell.edu/trlist/trlist.html).

Ruppert, D., Sheather, S. J., and Wand, M. P. (1995), "An Effective Bandwidth Selector for Local Least Squares Regression," *Journal of the American Statistical Association*, 90, 1257–1270.

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

Sigrist, M. (ed.) (1994), *Air Monitoring by Spectroscopic Techniques* (Chemical Analysis Series, Vol. 127), New York: Wiley.

Silverman, B. (1985), "Some Aspects of the Spline Smoothing Approach to Non-parametric Regression Curve Fitting" (with discussion), *Journal of the Royal Statistical Society*, Ser. B, 47, 1–52.

Stone, C. J. (1977), "Consistent Nonparametric Regression," *The Annals of Statistics*, 5, 595–620.

Turlach, B. A., and Wand, M. P. (1996), "Fast Computation of Auxiliary Quantities in Local Polynomial Smoothing," *Journal of Computational and Graphical Statistics*, 5, 337–350.

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