Local Polynomial Variance Function Estimation

David Ruppert

School of Operations Research and
Industrial Engineering,
Cornell University,
Ithaca, NY 14852, U.S.A.

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,
Box 118, S-221 00 Lund, SWEDEN

29th August, 1995

Abstract

The conditional variance function in a heteroscedastic, nonparametric regression model is estimated by linear smoothing of squared residuals. Attention is focussed 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 effect of preliminary estimation of the mean is studied, and a "degrees of freedom" 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.

1. Introduction

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 et. al (1995), where local polynomial regression is used for evaluation of the concentration of atmospheric atomic mercury measured with LIDAR technique (LIght Detection And Ranging, cf. Sigrist (1994)). In this example 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. In Holst et. al (1995) a parametric model is used 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 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 quantities such as mean velocity as a function of position. However, in the study of turbulence the variance of velocity and its derivatives as a function of position are also essential; see Section 7.2.

In this article we extend local polynomial regression ideas to estimation of the variance function. As we show in Section 2, our proposal can be generalised 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.

The literature on nonparametric variance function estimation is rather sparse. Carroll (1982) developed kernel estimators in the context of linear regression, while Müller and Stadtmüller (1987) and Hall and Carroll (1989) proposed and analysed kernel-type variance function estimators in the presence of a nonparametric mean function. Fan and Gijbels (1995) proposed a type of local polynomial variance function estimator as part of their bandwidth selection procedure.

In Section 2 we formulate a general class of nonparametic variance function estimators, and obtain local polynomial variance estimators as a special case. Section 3 investigates the theoretical properties of these estimators. Computational issues are described in Section 4 and extension to multivariate predictors in Section 5. Section 6 contains some illustrations of the methodology.

The variance function estimator in Section 2 was proposed independently by Mathur (1995), but the asymptotic theory, computational implementation, and bandwidth selectors proposed here are not in Mathur.

2. Formulation

2.1. A general class of variance function estimators

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

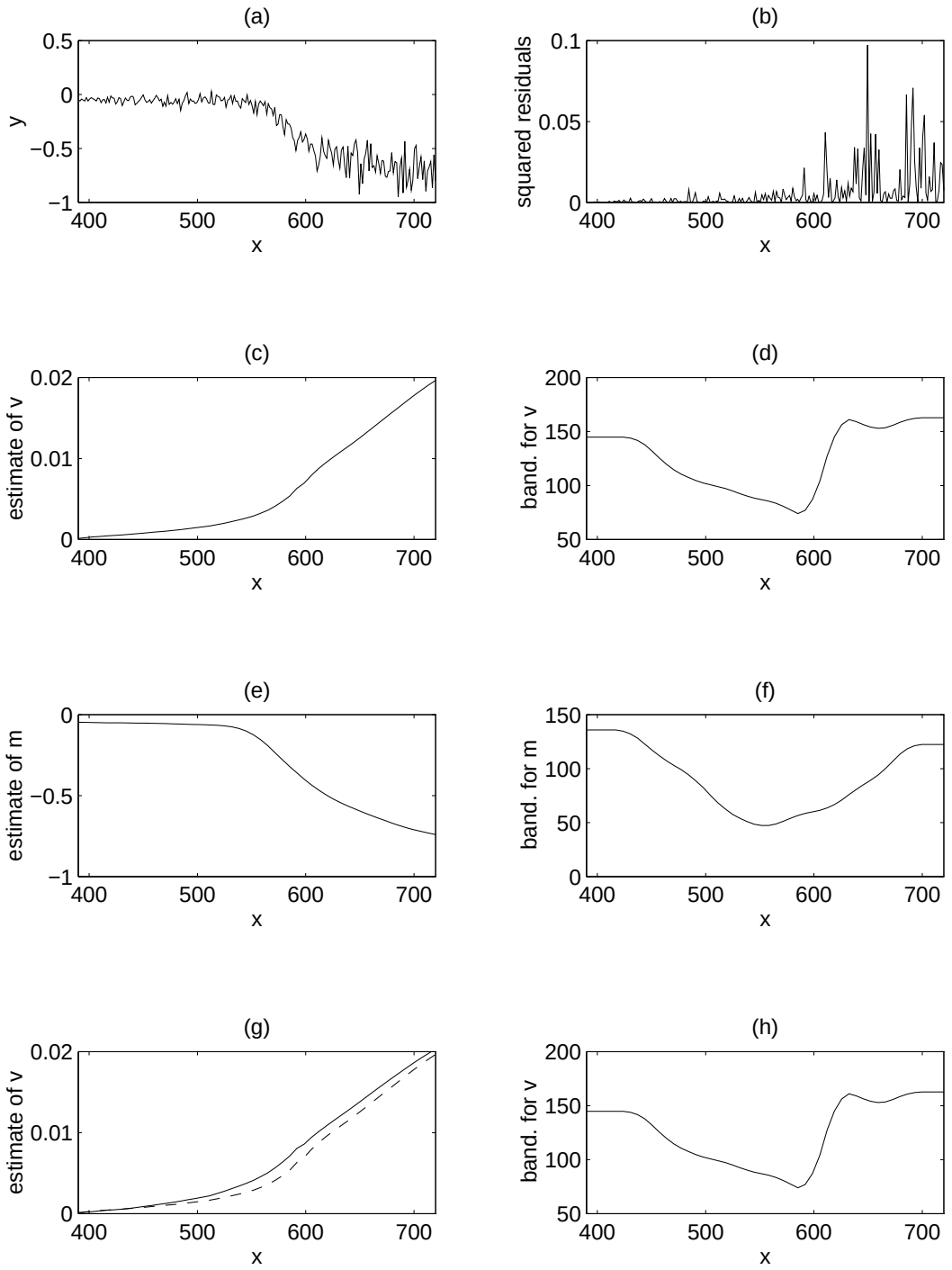

Figure 1: *LIDAR data. (a) Raw data (221 observations). (b) Squared residuals from a preliminary local linear estimate of the mean function with a global bandwidth of 30. (c)–(d) Local linear smooth of (b) and bandwidth. (e)–(f) Local linear smooth of (a) and bandwidth using the variance estimate in panel (c). (g)–(h) Local linear smooth and bandwidth using squared residuals from the fit in (e). In panel (g), the dashed curve is the raw smooth and the solid curve is "degrees of freedom" corrected.*

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:

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

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 $v(X_i)$, $1 \leq i \leq n$, respectively. Finally, Y will be used to the denote the $n \times 1$ vector of Y_i values.

Let

$$\widehat{m} = [\widehat{m}(X_1), \ldots, \widehat{m}(X_n)]^T$$

be a linear smooth of the (X_i, Y_i)'s. By this we mean that

$$\widehat{m} = SY$$

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 (see e.g. 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

$$r = (S_1 - I)Y,$$

the vector of residuals. Then a natural means of estimating $v = [v(X_1), \ldots, v(X_n)]^T$ 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 the initial smooth 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

$$\Delta = \text{diagonal}(S_1 S_1^T - 2S_1)$$

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

$$\widehat{v} = (S_2 r^2) \big/ (\mathbf{1} + S_2\Delta). \tag{2}$$

The convention here and throughout is that the vector multiplication and division are element-wise.

2.2. Relationships with parametric modelling

One can view the class of variance function estimators given by (2) as a generalisation of those commonly used whether either the mean or variance function are modelled parametrically. For example, if the mean is modelled linearly:

$$Y_i = (X\beta)_i + \varepsilon_i, \quad \text{var}(\varepsilon_i) = v(X_i), \quad i = 1, \ldots, n,$$

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 $R = X(X^TX)^{-1}X^T$. Using the symmetry and idempotency of R we obtain the variance function estimator

$$\widehat{v} = S_2\{(R-I)Y\}^2 \Big/ [\mathbf{1} - S_2\{\text{diagonal}(R)\}].$$

On the other hand, if the homoscedastic nonparametric regression model

$$Y_i = m(X_i) + \varepsilon, \quad \text{var}(\varepsilon_i) = \sigma^2, \quad i = 1, \ldots, n$$

is assumed then one should simply average the squared residuals by taking $S_2 = n^{-1}\mathbf{1}\mathbf{1}^T$. This results in

$$\hat{\sigma}^2 = \{Y^T(S_1 - I)^T(S_1 - I)Y\} \Big/ \{n + \text{tr}(S_1S_1^T - 2S_1)\},$$

which includes variance estimators for nonparametric regression considered by, for example, Buckley, Eagleson and Silverman (1988).

For the homoscedastic linear regression model the estimator reduces to the familiar

$$\hat{\sigma}^2 = Y^T(I - R)Y/(n - p).$$

2.3. Local polynomial variance function estimation

The class of linear smoothers that we concentrate on in this paper are those commonly referred to as *local polynomial smoothers* (see e.g Wand and Jones, 1995). The (i, j) entry of the pth degree local polynomial smoother matrix, $S_{p,h}$, is

$$(S_{p,h})_{ij} = e_1^T\{X_p(X_i)^T W_h(X_i) X_p(X_i)\}^{-1} X_p(X_i)^T W_h(X_i) e_j \tag{3}$$

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

$$X_p(x) = \begin{bmatrix} 1 & X_1 - x & \cdots & (X_1 - x)^p \\ \vdots & \vdots & \ddots & \vdots \\ 1 & X_n - x & \cdots & (X_n - x)^p \end{bmatrix} \text{ and } W_h(x) = \underset{1 \le i \le n}{\text{diag}} \; K\left(\frac{X_i - x}{h}\right)$$

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. Typically K is a smooth bell-shaped function such as the standard normal density, called the *kernel*, and $h = h(x)$ is a scaling parameter, usually referred to as the *bandwidth* at the point x.

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

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

where

$$r = (I - S_{p_1,h_1})Y \qquad \text{and} \qquad \Delta = \text{diagonal}(S_{p_1,h_1}S_{p_1,h_1}^T - 2S_{p_1,h_1}).$$

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 $S_2 = S_{p_2,h_2}$.

2.4. Estimation of derivatives of v

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 7.2. As discussed in Ruppert and Wand (1994), local polynomial estimation of the kth derivative of m is straightforward, and there is no problem extending derivative estimation to v. One needs to use $p_2 \geq k$ and then for the second smoother matrix, S_2, one merely replaces e_1^T in (3) by $k!e_{k+1}^T$. The theory in the next section extends easily to derivative estimation, but for simplicity we only consider the case of estimating v itself.

3. Theory

In this section we start by showing that it is possible to obtain exact matrix algebraic expressions for the conditional mean and covariance of $\widehat{v}$ for the general class of variance function estimators introduced in Section 2.1. In the local polynomial case one can use these results to obtain meaningful asymptotic approximations.

We retain the convention that multiplication and division of column vectors is element-wise. 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 $\mathcal{X} = \{X_1, \ldots, X_n\}$ to abbreviate expectations that are conditional on the predictors. Also, $\mathcal{C}(U|W)$ denotes the conditional covariance matrix of U given W whenever U and W are random vectors.

3.1. General variance function estimators

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

$$V = \text{diag}(v), \quad G = \underset{1 \leq i \leq n}{\text{diag}} \{E(\varepsilon_i^3)\} \quad \text{and} \quad T = \underset{1 \leq i \leq n}{\text{diag}} \{E(\varepsilon_i^4)\}.$$

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

$$E(\widehat{v} - v|\mathcal{X}) = \frac{(S_2 - I)v + S_2\{b_1^2 + \text{diagonal}(S_1 V S_1^T - 2S_1 V)\} - (S_2\Delta)v}{\mathbf{1} + S_2\Delta} \tag{4}$$

and

$$\begin{aligned}
\mathcal{C}(\widehat{v}|\mathcal{X}) \;=\; & S_2[\{(S_1 - I) \odot (S_1 - I)\}(T - 3V^2)\{(S_1 - I) \odot (S_1 - I)\}^T \\
& +2(\text{diag } b_1)(S_1 - I)G\{(S_1 - I) \odot (S_1 - I)\}^T \\
& +2\{(S_1 - I) \odot (S_1 - I)\}G(S_1 - I)^T(\text{diag } b_1) \\
& +2\{(S_1 - I)V(S_1 - I)^T\} \odot \{(S_1 - I)V(S_1 - I)^T\} \\
& +4\{(S_1 - I)V(S_1 - I)^T\} \odot (b_1 b_1^T)]S_2^T / \{(\mathbf{1} + S_2\Delta)(\mathbf{1} + S_2\Delta)^T\}.
\end{aligned}$$

The proof is given in the Appendix.

The expression for $\mathcal{C}(\widehat{v}|\mathcal{X})$ simplifies considerably if normality of the errors can be assumed:

COROLLARY 1.1. *If the errors ε_i are normally distributed then*

$$\mathcal{C}(\widehat{v}|\mathcal{X}) = \frac{2S_2[\{(S_1-I)V(S_1-I)^T\} \odot \{(S_1-I)V(S_1-I)^T + 2b_1b_1^T\}]S_2^T}{(\mathbf{1}+S_2\Delta)(\mathbf{1}+S_2\Delta)^T}.$$

REMARK 1. The conditional Mean Average Squared Error (MASE) of $\widehat{v}$ is defined as

$$\text{MASE}(\widehat{v}) = n^{-1}E\left[\sum_{i=1}^{n}\{\widehat{v}(X_i) - v(X_i)\}^2|\mathcal{X}\right].$$

Noting that

$$\text{MASE}(\widehat{v}) = n^{-1}\{\|E(\widehat{v}|\mathcal{X}) - v\|^2 + \text{tr}\,\mathcal{C}(\widehat{v}|\mathcal{X})\},$$

where $\|x\|^2 = x^Tx$, one can use the above results to find exact expressions for MASE($\widehat{v}$) for *any* pair of smoother matrices S_1 and S_2.

3.2. Local polynomial variance function estimators

Let f denote the common density of $X_1, \ldots, X_n$ and the function η be given by:

$$\eta(X_i) = \text{var}(\varepsilon_i^2), \quad i = 1, \ldots, n.$$

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 neighbourhood of x, and that $h_1, h_2 \to 0$, $nh_1, nh_2 \to \infty$, and*

$$\left\{h_1^{2(p_1+1)} + (nh_1)^{-1}\right\} = o(h_2^{p_2+1}) \tag{5}$$

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

$$E\{\widehat{v}(x) - v(x)|\mathcal{X}\} = \left\{\int u^{p_2+1}K_{(p_2)}(u)\,du\right\}\left\{\frac{v^{(p_2+1)}(x)}{(p+1)!}\right\}h_2^{p_2+1} + o_P(h_2^{p_2+1})$$

and, for p even

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

In either case

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

Once again, we defer to 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 $\widehat{v}(x)$. If $p_1 = p_2$ and if h_1 is chosen optimally for estimation of m, 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 (5) 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 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 $\widehat{v}(x)$ is proportional to the variance of the squared errors, rather than the Y_i's. Asymptotically, $\widehat{v}$ behaves like a local polynomial smooth of the (unobservable) ε_i^2's, i.e., v can be estimated as well as if m were known, so that there is no loss in asymptotic efficiency due to estimating m. For this reason, the estimate of the variance function based on squared residuals is "adaptive" in the sense of Bickel (1982).

REMARK 4. Once could also rework the steps used to prove Theorem 2 for the situation where 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).

3.3. Bandwidth choice

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's assume that the bandwidths are local. Ideally, one would choose both h_1 and h_2 to minimize the MSE of $\hat{v}$ at the point x. However, this is difficult to do in practice, since 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 asymptotically 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 6 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 selection to estimation of the mean function of the squared residuals. If one uses $p_1 = p_2$, then (5) will be satisfied.

4. COMPUTATION

Direct computation of $\widehat{v}$ over a grid can be quite expensive, especially if the sample size is large. For example, if one decides to compute $\widehat{v}$ at the observations then one must deal with the fact that diagonal$(S_1 S_1^T)$ requires $O(n^2)$ operations for exact computation.

A simple way to overcome computational problems such as this is to use binned approximations. Turlach and Wand (1995) explain how one can apply binning to the type of

quantities that arise in variance function estimation. Let $g_1 < \ldots < g_M$ be an equally-spaced grid over the range of the X_i's and let $\delta = (g_M - g_1)/(M-1)$ be the gap between successive grid points. The grid count (c_ℓ, d_ℓ^Y) at grid point g_ℓ, with respect to linear binning, is given by

$$c_\ell = \sum_{i=1}^{n}(1 - |\delta^{-1}X_i - g_\ell|)_+ \quad \text{and} \quad d_\ell^Y = \sum_{i=1}^{n}(1 - |\delta^{-1}X_i - g_\ell|)_+ Y_i$$

where $x_+ = \max(0, x)$. Set

$$\widetilde{X}_p(x) = \begin{bmatrix} 1 & g_1 - x & \cdots & (g_1 - x)^p \\ \vdots & \vdots & \ddots & \vdots \\ 1 & g_M - x & \cdots & (g_M - x)^p \end{bmatrix}, \; \widetilde{W}_h(x) = \operatorname*{diag}_{1 \le \ell \le M} K\left(\frac{g_\ell - x}{h}\right),$$

$$C = \operatorname*{diag}_{1 \le \ell \le M}(c_\ell), \quad c = (c_1, \ldots, c_M)^T \quad \text{and} \quad d^Y = (d_1^Y, \ldots, d_M^Y)^T.$$

Then the binned analogue of the smoother matrix is $\widetilde{S}_{p,h}$ where

$$(\widetilde{S}_{p,h})_{\ell\ell'} = e_1^T\{\widetilde{X}_p(g_\ell)^T\widetilde{W}_h(g_\ell)C\widetilde{X}_p(g_\ell)\}^{-1}\widetilde{X}_p(g_\ell)^T\widetilde{W}_h(g_\ell)e_{\ell'},$$

since it can be easily shown that $\widetilde{S}$ maps the Y grid counts in d^Y to the vector of binned smooths at the grid points. The vector of binned estimates of v at the grid points is then

$$\tilde{v} = (\widetilde{S}_{p_2,h_2}\widetilde{r^2})\big/(\mathbf{1} + \widetilde{S}_{p_2,h_2}\tilde{\Delta})$$

where

$$\widetilde{r^2} = d^{Y^2} - 2(\widetilde{S}_{p_1,h_1}d^Y)d^Y + (\widetilde{S}_{p_1,h_1}d^Y)^2 c \quad \text{and} \quad \tilde{\Delta}_\ell = \{U_\ell - 2(\widetilde{S}_{p_1,h_1})_{\ell\ell}\}c_\ell.$$

Here U_ℓ, $1 \le \ell \le M$ denotes the binned approximation to diagonal$(S_{p_1,h_1}S_{p_1,h_1}^T)$ over the grid. Its fast computation is described in Section 4 of Turlach and Wand (1995).

5. Extension to Multivariate 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) is 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.

In the case of local polynomial smoothing, extra notation is required to handle multivariate X_i. See, for example, Ruppert and Wand (1994). If the X_i's are d-dimensional then the kernel K should be a d-variate function and scalings by a positive definite $d \times d$ bandwidth matrix H should be permitted. The weight associated with multivariate local polynomial smooth is then

$$W_H(x) = \operatorname*{diag}_{1 \le i \le n} K\{H^{-1/2}(X_i - x)\}.$$

For the local linear multivariate variance estimator, with bandwidth matrices H_1 and H_2, one can show that, for x in the interior of the support of f,

$$E\{\hat{v}(x) - v(x)|\mathcal{X}\} = \tfrac{1}{2}\left\{\int u^2 K(u)\,du\right\} \mathrm{tr}\{H_2 \mathcal{H}_v(x)\} + o_P\{\mathrm{tr}(H_2)\}$$

and that

$$\mathrm{var}\{\hat{v}(x)|\mathcal{X}\} = n^{-1}|H_2|^{-1/2}\left\{\int K(u)^2\,du\right\}\eta(x)/f(x) + o_P\{n^{-1}|H_2|^{-1/2}\}$$

where $\mathcal{H}_v$ is the Hessian matrix of v. This is analogous to the result for estimation of m in the multivariate context, given by Theorem 2.1 of Ruppert and Wand (1994).

6. Piecewise Polynomial Binning

In this section, we describe an alternative method that is particularly well suited for larger data sets, e.g., the turbulence data set of Section 7.2 which has 20,000 observations. We only describe the implementation for univariate X_i.

First the data are binned according to their x-values into n_{bin} disjoint subsets with roughly equal number of observations per subset (not equal lengths). For the jth bin, $j = 1, \ldots, n_{\mathrm{bin}}$, let $\bar{x}_j$ be the mean of the X_i's in that bin. Fit a p_1th 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. Therefore, if m is a p_1th 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, $\{\bar{y}_1, \ldots, \bar{y}_{n_{\mathrm{bin}}}\}$ are mutually independent as are $\bar{v}_1, \ldots, \bar{v}_{n_{\mathrm{bin}}}\}$. 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 is needed here, since the correction was made at the binning stage.

The idea is to choose n_{bin} and p_1 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_1+2 so that the residual degrees of freedom is 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.

Using $p_1 = 1$ will give accuracy similar to the popular linear binning technique, while $p_1 > 1$ will be more accurate than binning techniques now in the literature and will allow a smaller value of n_{bin}.

7. Examples

7.1. LIDAR data

We now return to the LIDAR data described in Section 1. First we used a local linear estimate of the mean with a global bandwidth chosen subjectively to equal 30. Squared

residuals from this fit are in Figure 1(b). In Figure 1(c) we have a local linear smooth of these squared residuals using the EBBS local bandwidth of Ruppert (1995) and computed on a 50-point equally spaced grid. The bandwidth itself is in Figure 1(d)—the EBBS method allows smoothing of the bandwidth with two tuning parameters, MSESPAN and BANDSPAN, which are both equal to 4 here. This means that an initial bandwidth at each point of the grid is based on a 9-point moving average of an estimated MSE and the final bandwidth is a 9-point moving average of the initial bandwidth; see Ruppert (1995) for details. The bandwidth selector assumes that the ratio of the standard deviation to the mean of the squared residuals does not depend on x, so the variance function of the squared residuals need not be separately estimated.

In Figure 1(e) we have a local linear smooth of the data in Figure 1(a), using the EBBS bandwidth shown in Figure 1(f). This bandwidth is based upon the variance function estimate in Figure 1(c).

The solid curve in Figure 1(g) is the same as the curve in Figure 1(c), except that the residuals in Figure 1(g) are from the curve in Figure 1(e), not the local linear fit with a bandwidth equal to 30 as in Figure 1(c). Notice that the two curves appear identical, showing that the effect of h_1 on $\hat{v}$ is minimal. The solid curve in Figure 1(g) and the curve in Figure 1(c) are *not* corrected by dividing by $(\mathbf{1} + S_2\Delta)$. The dashed curve in Figure 1(g) is the corrected estimate. The correction is not sizeable, but it does increase the estimated variance as expected.

The squared residuals in Figure 1(b) suggest that v might be bimodal. However, our local bandwidth selector chooses bandwidths large enough to smooth away the bimodality, suggesting that the apparent bimodality is merely a chance phenomenon and, in fact, v is monotonically increasing.

7.2. 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, U_i) where X_i is position and U_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. 20,000 values, $\{X_i : i = 1, \ldots, 20,000\}$, where taken uniformly distributed on $[0, .1]$, and at each X_i, U_i was 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, e.g., $v''(0)$, could be estimated accurately in the Pdf method. This quantity is of special interest since it is a boundary condition on turbulent dissipation. 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 6 with $n_{\text{bin}} = 200$ (100 observations/bin) and $p_1 = 2$ (piecewise quadratic binning). The residual mean squares are plotted in Figure 2(a) as a function of $\bar{x}$. Figure 2(b) is a plot of a local quadratic smooth of the data in Figure 7(a) (solid) and the v (dashed). The local bandwidth given in Figure 2(c) was generated by EBBS (Ruppert 1995) assuming that the variance function of the $\bar{v}$'s is proportional to the square of their mean function. As can be seen in Figure 2(a), this assumption is, in fact, true since the ε_i's are in a scale family.

In Figure 2(d) we have $\hat{v}''$ (solid) using local cubic smoothing as discussed in Section 2.4 and with the EBBS bandwidth shown in Figure 2(e). Also in Figure 2(d) is v'' (dashed).

The engineers concluded that estimation of $v''(0)$ is feasible, but that sample sizes of at least 20,000 are necessary.

Acknowledgement

We thank Stephen Pope and Tom Dreeben for supplying the turbulence data and for helpful discussions.

Appendix: Proofs of Theorems

Proof of Theorem 1

First note that

$$\widehat{v} = \frac{S_2 \text{diagonal}\{(S_1 - I)YY^T(S_1 - I)^T\}}{\mathbf{1} + S_2\Delta}.$$

For the bias we have

$$\begin{aligned} E(\widehat{v}|\mathcal{X}) &= \frac{S_2 \text{diagonal}\{(S_1 - I)(mm^T + V)(S_1 - I)^T\}}{\mathbf{1} + S_2\Delta} \\ &= \frac{S_2\{\text{diagonal}(b_1 b_1^T) + v + \text{diagonal}(S_1 V S_1^T - 2S_1 V)\}}{\mathbf{1} + S_2\Delta}. \end{aligned}$$

Direct algebra then leads to the stated result.

The result for $\mathcal{C}(\widehat{v}|\mathcal{X})$ depends heavily on:

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,*

$$\begin{aligned} \mathcal{C}\{(AY)^2\} = (A \odot A)(T - 3V^2)(A \odot A)^T + 2\{&\text{diag}(Am)AG(A \odot A)^T \\ &+ (A \odot A)GA^T \text{diag}(Am)\} + 2(AVA^T) \odot (AVA^T) + 4(AVA^T) \odot \{(Am)(Am)^T\}. \end{aligned}$$

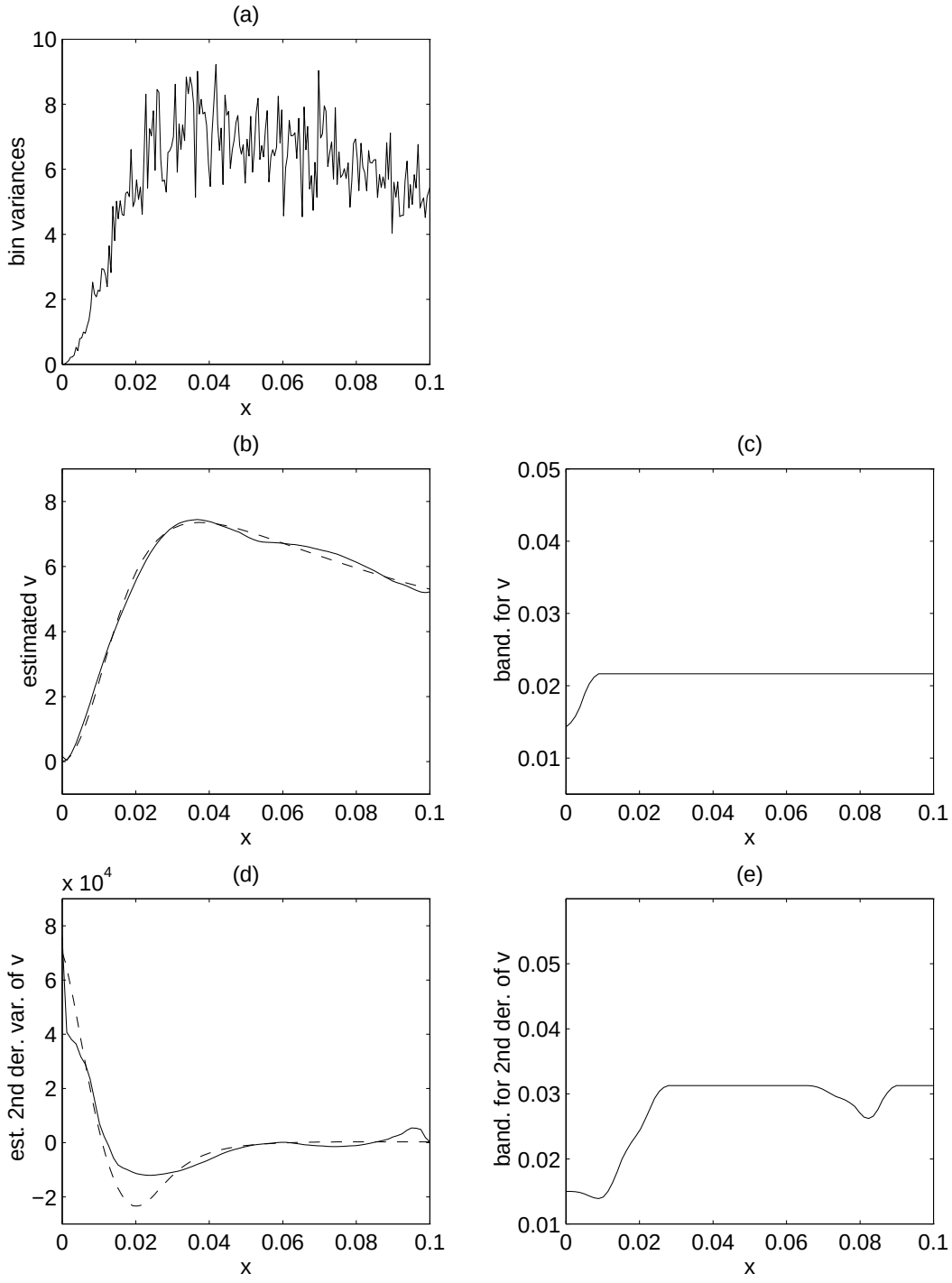

Figure 2: *Turbulence data. (a) Residual mean squares plotted against bin means of x. (b) Local quadratic smooth of data in (a) (solid) and v (dashed). (c) Local EBBS bandwidth used in (b). (d) Local cubic estimate of v'' using data from (a) (solid) and v'' (dashed). (e) Local EBBS bandwidth used in (d).*

PROOF. We will use the tensor notation and results of McCullagh (1987). Let a_{ij} denote the (i,j) entry of A. Moments of products of the entries of Y will be denoted by κ with appropriate superscripts. For example,

$$\kappa^{ij} = E(Y_iY_j) \qquad \text{and} \qquad \kappa^{iijk} = E(Y_i^2Y_jY_k).$$

Generalised cumulants will be denoted using partitioned superscript notation. For example,

$$\kappa^{i,j} = \text{cum}(Y_i, Y_j) = \text{cov}(Y_i, Y_j) \qquad \text{and} \qquad \kappa^{i,j,k\ell} = \text{cum}(Y_i, Y_j, Y_kY_\ell).$$

Then the (m,n) entry of $\mathcal{C}\{(AY)^2\}$ is easily shown to be

$$\mathcal{C}\{(AY)^2\}_{mn} = \sum_i \sum_j \sum_k \sum_\ell a_{mi}a_{mj}a_{nk}a_{n\ell}\kappa^{ij,k\ell}.$$

One of the fundamental identities for generalised cumulants, given on p.58 of McCullagh (1987), states that

$$\begin{aligned}\kappa^{ij,k\ell} &= \kappa^{i,j,k,\ell} + \kappa^i\kappa^{j,k,\ell} + \kappa^j\kappa^{i,k,\ell} + \kappa^k\kappa^{i,j,\ell} + \kappa^\ell\kappa^{i,j,k} + \kappa^{i,k}\kappa^{j,\ell} + \kappa^{i,\ell}\kappa^{j,k} \\ &\quad + \kappa^i\kappa^k\kappa^{j,\ell} + \kappa^i\kappa^\ell\kappa^{j,k} + \kappa^j\kappa^k\kappa^{i,\ell} + \kappa^j\kappa^\ell\kappa^{i,k}.\end{aligned}$$

This implies that, because of the mutual independence of the Y_i's,

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

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

∎

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

LEMMA 2. *If a is a constant vector having the same length as Y then*

$$\mathcal{C}(aY) = (aa^T) \odot \mathcal{C}(Y).$$

The result for $\mathcal{C}(\widehat{v}|\mathcal{X})$ follows immediately from Lemmas 1 and 2 and the well-known result: $\mathcal{C}(AY) = A\mathcal{C}(Y)A^T$.

∎

Proof of Theorem 2

For a rth differentiable function g we let $g^{(r)} = [g^{(r)}(X_1), \ldots, g^{(r)}(X_n)]^T$. 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. *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*

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

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

$$(3)\quad \text{diagonal}(S_{p,h}) = O_P\{(nh)^{-1}\},$$

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

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 (4) is $(S_2 - I)v$. Since 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 (5) is expanded out, the dominating terms are seen to be

$$S_2\{(T - 3V^2) + 2V^2\} S_2^T = S_2\, \text{diag}(\eta) S_2^T$$

where $\eta = [\eta(X_1), \ldots, \eta(X_n)]^T$. Application of (2) of Lemma 3 then leads to the desired result.

∎

REFERENCES

Bickel, P.J. (1992). On adaptive estimation. *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. *Annals of Statistics*, **10**, 1224-1233.

Dreeben, T.D., and Pope, S.B. (1995). "Pdf and Reynolds-stress modeling of near-wall turbulent flows." In *Tenth Symposium on Turbulent Shear Flows*, (J. Wyngaard, ed.). pp. 2.1–2.6.

Fan, J. and Gibjels, I. (1992). Variable bandwidth and local linear regression smoothers. *Annals of Statistics*, **20**, 2008–2038.

Fan, J. and Gijbels, I. (1995). Data-driven bandwidth selection in local polynomial fitting: variable bandwidth and spatial adaptation. *Journal of the Royal Statistical Society, Series B*, **57**, 371–394.

Hall, P. and Carroll, R.J. (1989). Variance function estimation in regression: the effect of estimating the mean. *Journal of the Royal Statistical Society, Series B*, **51**, 3–14.

Hastie, T.J. and Loader, C. (1993) Local regression: automatic kernel carpentry (with discussion). *Statist. Sci.*, **8**, 120–143.

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

Holst, U., Hössjer, O., Björklund, C., Ragnarsson, P., and Edner, H. (1995). Locally weighted least squares kernel regression and statistical evaluation of LIDAR measurements, under revision for *Environmetrics*.

Mathur, A. (1995). On estimation of residual variance function (abstract). In *Summaries of papers presented at the Joint Statistical Meetings, Orlando, Florida, August 13–17, 1995*, pp. 279.

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

Müller, H.G. and Stadtmüller, U. (1987). Estimation of heteroscedasticity in regression analysis. *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.

Ruppert, D. (1995). Empirical-bias bandwidths for local polynomial nonparametric regression and density estimation. In preparation.

Ruppert, D. and Wand, M.P. (1994). Multivariate locally weighted least squares regression. *Annals of Statistics*, **22**, 1346–1370.

Sigrist, M., editor. *Air monitoring by spectroscopic techniques (Chemical Analysis Series, Vol. 127).* Wiley, 1994.

Turlach, B.A. and Wand, M.P. (1995). Fast computation of auxiliary quantities in local polynomial smoothing. *The University of New South Wales, Australian Graduate School*

of Management Working Paper Series, 95–009.
(URL `http://www.agsm.unsw.edu.au/stats/Working.html`)

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