Biometrika (1995), **82**, 1, *pp.* 93–100
Printed in Great Britain

A kernel method of estimating structured nonparametric regression based on marginal integration

BY OLIVER LINTON
Cowles Foundation for Research in Economics, Yale University, Connecticut 06520, U.S.A.
AND JENS PERCH NIELSEN
PFA Pension, Copenhagen, Denmark

SUMMARY

We define a simple kernel procedure based on marginal integration that estimates the relevant univariate quantity in both additive and multiplicative nonparametric regression.

Some key words: Additive model; Backfitting; Kernel estimation; Model selection; Nonparametric regression.

1. INTRODUCTION

Nonparametric regression is frequently used as a preliminary diagnostic tool. It is a convenient method of summarising the relationship between a dependent and a univariate independent variable. However, when the explanatory variables are multidimensional, these methods are less satisfactory. In particular, the rate of convergence of standard estimators is poorer, while simple plots are not available to aid model selection.

There are a number of simplifying structures that have been used to avoid these problems. These include the regression tree structure of Gordon & Olshen (1980), the projection pursuit model of Friedman & Stuetzle (1981), semiparametric models such as considered by Engle et al. (1986), and the additive structure of Buja, Hastie & Tibshirani (1989): see Härdle (1990, pp. 257–87) for further discussion. We restrict our attention to fully nonparametric additive and multiplicative structures. Let $g(x, z)$ be a bivariate regression function; an additive sub-model is

$$g(x, z) = g_1(x) + g_2(z). \tag{1·1}$$

Stone (1985) shows that both g_1 and g_2 can be estimated with the one-dimensional convergence rate of $n^{2/5}$, where n is the sample size. In practice, the Hastie & Tibshirani (1990) estimation procedures are widely used. These involve multiple iterations, where the additive structure is used in each step, to obtain estimates of g_1 and g_2. A major disadvantage of this method is that its statistical properties are not well understood.

The main purpose of this paper is to introduce an alternative kernel procedure for estimating g_1 and g_2 that has several advantages over the conventional method. Our estimator is explicitly defined and its statistical properties are easily derived: it has convergence rate $n^{2/5}$. We provide consistent confidence intervals and an automatic bandwidth selection method. In addition, the multiplicative sub-model

$$g(x, z) = h_1(x)h_2(z) \tag{1·2}$$

is also plausible in many practical situations. If (1·2) is correct, it is not clear what the

Hastie & Tibshirani procedures are estimating, and an additional procedure such as suggested by Breiman (1991) must be used to extract the univariate effects. By contrast our method also estimates h_1 and h_2, when (1·2) is true. This is particularly important when choosing amongst these nonnested specifications; with our approach it is unnecessary to match the degrees of freedom from distinct additive and multiplicative fits. We construct residuals that can be used for selecting between the additive and multiplicative structures.

2. A GENERAL PROCEDURE

Let Q be a deterministic weighting function with $\int dQ(z) = 1$. We allow for both discrete and continuous Q, and integrals should be interpreted in the Stieltjes sense. Let q be the density of Q with respect to either Lebesgue or a counting measure. Now consider the following contrast:

$$\alpha_Q(x) = \int g(x, z)\, dQ(z). \tag{2·1}$$

In the additive case, $\alpha_Q(x) = g_1(x) + c_1$, where $c_1 = \int g_2(z)\, dQ(z)$, while in the multiplicative situation, $\alpha_Q(x) = h_1(x)c_2$, where $c_2 = \int h_2(z)\, dQ(z)$. Therefore $\alpha_Q(x)$ is, up to identifiability, the univariate component of interest in both additive and multiplicative structures.

Let $\{y_i, x_i, z_i\}_{i=1}^n$ be our sample. We estimate the regression of y on x and z, $g(x, z)$, by the local linear smoother of Fan (1992) with product kernels; thus $\hat{g}(x, z) = \sum w_j(x, z)y_j$ is the first element of

$$(D^{\mathrm{T}}KD)^{-1}D^{\mathrm{T}}Ky,$$

where $y = (y_1, y_2, \ldots, y_n)^{\mathrm{T}}$ and $D = (d_1, d_2, \ldots, d_n)^{\mathrm{T}}$ with $d_j = (1, x_j - x, z_j - z)^{\mathrm{T}}$, while K is a diagonal n by n matrix with typical diagonal entry $k_{b_1}(x_j - x)k_{b_2}(z_j - z)$. Here, b_1 and b_2 are scalar bandwidths, and $k_b(.) = b^{-1}k(b^{-1}.)$ for any b, where k is a univariate differentiable probability density function, symmetric about zero. Wand & Jones (1993) discuss the merits of using different bandwidths for each direction. In sum, we estimate $\alpha_Q(x)$ by the sample version of (2·1):

$$\hat{\alpha}_Q(x) = \int \hat{g}(x, z)\, dQ(z) = \sum_{j=1}^{n} w_{Qj}(x)y_j,$$

where $w_{Qj}(x) = \int w_j(x, z)\, dQ(z)$.

3. PROPERTIES OF THE ESTIMATOR

Here, we treat only the bivariate case, but in § 4·3 below we discuss the extension to general dimensions. Suppose that (y_i, x_i, z_i) are independent and identically distributed, with

$$y_i = g(x_i, z_i) + \varepsilon_i \quad (i = 1, 2, \ldots, n),$$

where $E(\varepsilon_i | x_i, z_i) = 0$ and $\mathrm{var}\,(\varepsilon_i | x_i, z_i) = \sigma^2$, while (x_i, z_i) has joint density $f(x, z)$ and marginals $f_x(x)$ and $f_z(z)$ with marginal cumulative distribution functions $F_x(x)$ and $F_z(z)$. Furthermore, we assume that f has a compact support $A_x \times A_z = A \subseteq \mathbb{R}^2$, and that the support of Q is contained within A_z.

We conduct our analysis conditional on $\{x_i, z_i\}_{i=1}^n$, and all expectations below are taken

with respect to the conditional distribution. Ruppert & Wand (1994) obtain the first two moments of $\hat{g}(x, z)$:

$$E\{\hat{g}(x, z)\} - g(x, z) = O(b_1^2) + O(b_2^2), \quad \text{var}\{\hat{g}(x, z)\} = O(n^{-1}b_1^{-1}b_2^{-1}).$$

These orders of magnitude hold also for boundary points. Employing their proof technique, we obtain by straightforward calculation the following result.

THEOREM. *Assume that g possesses two continuous partial derivatives in each direction, while f is continuously differentiable. Suppose also that $b_1, b_2 \to 0$ and $nb_1b_2^2 \to \infty$. Then, conditional on $\{x_i, z_i\}_{i=1}^n$,*

$$(nb_1)^{\frac{1}{2}}[\hat{\alpha}_Q(x) - E\{\hat{\alpha}_Q(x)\}] \to N\{0, s^2(x)\}, \tag{3·1}$$

in distribution, where $s^2(x) = v(k)\sigma^2 \int f^{-1}(x, z)q^2(z)\,dz$, with $v(k) = \int k(t)^2\,dt$, while

$$E\{\hat{\alpha}_Q(x)\} - \alpha_Q(x) = \{b_1^2\beta_1(x) + b_2^2\beta_2(x)\}\{1 + o(1)\}, \tag{3·2}$$

where

$$\beta_1(x) = \mu(k)\int \frac{\partial^2 g}{\partial x^2}(x, z)\,dQ(z)/2, \quad \beta_2(x) = \mu(k)\int \frac{\partial^2 g}{\partial z^2}(x, z)\,dQ(z)/2,$$

with $\mu(k) = \int t^2 k(t)\,dt$.

Proof. The deterministic part (3·2) follows by integrating the bias of the local linear estimator. The stochastic part (3·1) is dealt with in two stages. First, it is shown that the local linear estimator itself can be approximated by the internal estimator

$$\hat{g}_I(x, z) = n^{-1}\sum_{i=1}^{n} k_{b_1}(x - x_i)k_{b_2}(z - z_i)y_i/f(x_i, z_i),$$

with integrated, with respect to Q, error of $O_p(n^{-1}b_1^{-1}b_2^{-1}) + o_p(b_1b_2)$ in our case, and error $O_p(n^{-1}) + o_p(b_1^2b_2^2)$ under further smoothness (Jones, Davies & Park, 1994). This allows us to work with the stochastic part of $\hat{g}_I(x, z)$. Integrating with respect to Q, we get

$$\int [\hat{g}_I(x, z) - E\{\hat{g}_I(x, z)\}]\,dQ(z) = n^{-1}\sum_i k_{b_1}(x - X_i)\left\{\int k_{b_2}(z - z_i)\,dQ(z)\right\}\varepsilon_i/f(x_i, z_i).$$

Then, since $\int k_{b_2}(z - z_i)\,dQ(z) = q(z_i) + o(1)$, the result follows. □

When $b_1, b_2 = O(n^{-1/5})$, $\hat{\alpha}_Q(x)$ converges at rate $n^{2/5}$. Some simplification to the asymptotic bias results if one under-smooths in the direction of z, that is by taking $b_2 = o(n^{-1/5})$, in which case (3·2) depends only on derivatives with respect to x; that is the bias is $b_1^2\mu(k)\alpha_Q''(x)/2$. When g is additive, (3·2) is $b_1^2\mu(k)g_1''(x)/2$, while in the multiplicative case it is $b_1^2\mu(k)c_2h_1''(x)/2$. The arguments of Fan (1992) suggest that under-smoothing is superior according to a minimax criterion.

We now turn to the choice of Q. We first describe the minimum variance weighting function Q_{MV}, with density q_{MV}, that minimises $S = \int s^2(x)f_X(x)\,dx$. For any $q(z)$,

$$\int\left\{\int f^{-1}(x, z)q^2(z)\,dz\right\}f_X(x)\,dx = \int q^2(z)\tau(z)\,dz \geqslant 1 \Big/ \int \tau^{-1}(z)\,dz,$$

by the Fubini theorem and the Schwarz inequality, where $\tau(z) = \int f^{-1}(x, z)f_X(x)\,dx$.

Therefore,

$$q_{\mathrm{MV}}(z) = \tau^{-1}(z) \Big/ \int \tau^{-1}(z)\, dz,$$

in which case $S = v(k)\sigma^2 / \int \tau^{-1}(z)\, dz$. An integrated-mean-squared-error-optimal Q can also be obtained, but in general this will depend on both f and g, although in the special case that g is additive and $b_2 = o(n^{-1/5})$, the bias does not depend on Q; in this case, q_{MV} is integrated-mean-square-error-optimal. In practice, we recommend using the empirical distribution function $F_{zn}(z)$ that converges weakly to $F_z(z)$. Our approximations, with $F_z(z)$ in place of Q in (3·1) and (3·2), remain valid in this case. When x and z are independent, $Q_{\mathrm{MV}}(z) = F_z(z)$, which provides some justification for this choice. An alternative is the uniform weighting function $(\int_{A(z)} dz)^{-1} \chi(z \in A_z)$, where $\chi(B)$ is the indicator function of the event B.

We now compare with the backfitting algorithm of Hastie & Tibshirani (1990). Let W_1 and W_2 be the $n \times n$ smoother matrices for estimating the marginal regression functions at each sample point. The backfitting procedure converges to the $n \times 1$ vector linear smoothers

$$\tilde{g}_1^{\infty} = \{I - (I - W_1 W_2)^{-1}(I - W_1)\}\, y, \quad \tilde{g}_2^{\infty} = \{I - (I - W_2 W_1)^{-1}(I - W_2)\}\, y,$$

provided the matrix norm of $W_1 W_2$ is less than one in absolute value (Hastie & Tibshirani, 1990, p. 119). Here, I is the $n \times n$ identity matrix. Unfortunately, these expressions appear quite intractable, and, although consistent confidence intervals can be constructed by exploiting linearity, to our knowledge no expressions comparable to (3·1) and (3·2) exist for the bias and variance of $\tilde{g}_1^{\infty}$ and $\tilde{g}_2^{\infty}$. Without such expressions it is difficult to make precise statements about bandwidth choice or model selection.

4. EXTENSIONS

4·1. *Goodness of fit*

Since $\hat{\alpha}_Q(x)$ converges to $g_1(x) + c_1$ or $h_1(x)c_2$ depending on which structure is true, it is useful to have some way of discriminating between these two models. Let

$$\alpha_{Q_1}(x) = \int g(x, z)\, dQ_1(z), \quad \alpha_{Q_2}(z) = \int g(x, z)\, dQ_2(x),$$

and let $\hat{\alpha}_{Q_1}(x)$ and $\hat{\alpha}_{Q_2}(z)$ denote the estimated quantities. Further define

$$\hat{c} = \int \hat{g}(x, z)\, dQ_1(z)\, dQ_2(x).$$

Then under the additive structure

$$\hat{c} \to \alpha_{Q_1} + \alpha_{Q_2} - (g_1 + g_2)$$

in probability, while under the multiplicative structure

$$\hat{c} \to \frac{\alpha_{Q_1} \alpha_{Q_2}}{h_1 h_2}.$$

Therefore, the residuals

$$\hat{\varepsilon}_{ai} = y_i - \{\hat{\alpha}_{Q_1}(x_i) + \hat{\alpha}_{Q_2}(z_i) - \hat{c}\}, \quad \hat{\varepsilon}_{mi} = y_i - \hat{c}^{-1} \hat{\alpha}_{Q_1}(x_i) \hat{\alpha}_{Q_2}(z_i)$$

can be used to evaluate the fit and to select the appropriate model. One possible criterion is to select the additive model if

$$\sum_{i=1}^{n} \hat{\varepsilon}_{ai}^2 < \sum_{i=1}^{n} \hat{\varepsilon}_{mi}^2, \tag{4·1}$$

and the multiplicative structure if the reverse inequality holds.

Of course, one could also compare residuals from the Hastie & Tibshirani additive fit with residuals from Breiman's method of multiplicative fitting. But it is necessary to match the effective smoothing in each of these two procedures. This is less of a problem for our method, since our single procedure estimates both additive and multiplicative structures.

4·2. *Additive and interaction terms*

Suppose that

$$g(x, z) = g_1(x) + g_2(z) + h_1(x)h_2(z),$$

where for identification purposes it is assumed that

$$\int h_1(x) f_x(x)\, dx = \int h_2(z) f_z(z)\, dz = 0,$$

but $\int g_1(x) f_x(x)\, dx \neq 0$ and $\int g_2(z) f_z(z)\, dz \neq 0$. In this case, our procedure with $q_1 = f_z$ and $q_2 = f_x$ estimates the additive components g_1 and g_2 up to location. The additive residuals can then be used to detect interactions. If $e_i = y_i - g_1(x_i) - g_2(z_i)$ were observed, then h_1 and h_2 could be consistently estimated by applying our procedure with e_i replacing y_i, and taking Q_1 and Q_2 to be some other measures for which $\int h_1\, dQ_2$ and $\int h_2\, dQ_1$ are both nonzero. This procedure is still consistent when e_i is replaced by $\hat{\varepsilon}_{ai}$.

4·3. *High dimensions*

Now suppose that z is $(P-1)$-dimensional. Our procedure for extracting the univariate effect of x is exactly the same except that we must use a P-dimensional preliminary kernel estimator and a $(P-1)$-dimensional weighting function Q. In this case, $\alpha_Q(x)$ consistently estimates either $g_1(x)$ or $h_1(x)$, and does so without imposing any structure on the $(P-1)$-dimensional functions $g_2(z)$ or $h_2(z)$. By repeated application of our procedures one can obtain all univariate effects.

Unfortunately, our attempts to extend the theorem to higher dimensions have met with difficulties. We are unable to show that

$$\int \{\hat{g}(x, z) - \hat{g}_I(x, z)\}\, dQ(z)$$

is any smaller than $O_p(n^{-1}b^{-P})$, when a common bandwidth b is used for the initial smooth, under only second derivative assumptions. This approximation error can dominate unless b is chosen very large, thereby excluding the optimal bandwidth $b = O(n^{-1/5})$. This happens when $P \geqslant 3$. In this sense our procedure suffers from the curse of dimensionality. We believe that the problem is primarily associated with estimation of the nuisance function f, rather than with the fact that we work initially with a high-dimensional smoother: if we could work directly with $\hat{g}_I(x, z)$, then it is clear that there would be no such restraint on bandwidth. In any case, there is an alternative remedy which is to use

bias reduction. Suppose that an rth order kernel or local rth order polynomial estimator with a bandwidth $b_r = O(n^{-1/(2r+1)})$ is used for the pilot estimate of g. Because of the wider bandwidth, the approximation error in using $\hat{g}_I(x, z)$ is now small enough: provided $P < r+1$, the optimal convergence rate $n^{-r/(2r+1)}$ is achieved by $\hat{\alpha}_Q(x)$, when g has r continuous derivatives.

5. EXAMPLES AND SIMULATIONS

We now illustrate our procedure on a real example. The data are a random sample of 534 individuals from the 1985 Current Population Survey conducted by the U.S. Department of Commerce. Details of this data set are given by Berndt (1991, Ch. 5). We examine the relationship between logarithm of wages, y, and the covariates education in years, x, and experience in years, z.

Let $\hat{G}$ be the $n \times n$ matrix of estimates, with typical element $\hat{g}(x_i, z_j)$, calculated using the Gaussian kernel $k(t) = (2\pi)^{-\frac{1}{2}} \exp(-t^2/2)$ and a common bandwidth $b = b_1 = b_2$ defined below. Define the $n \times 1$ vectors $\hat{\alpha}_{Q_1}$ and $\hat{\alpha}_{Q_2}$ containing the estimated univariate components evaluated at each sample point

$$\hat{\alpha}_{Q_1} = \hat{G} q_1, \quad \hat{\alpha}_{Q_2} = \hat{G}^{\mathrm{T}} q_2. \tag{5·1}$$

We use the empirical distribution functions for weighting, that is $q_1, q_2 = (1, 1, \ldots, 1)^{\mathrm{T}}/n$. When estimating the x effect, we took $b = 3{\cdot}14$, while $b = 3{\cdot}93$ was used for the z effect; these were selected by the following rule of thumb:

$$\hat{b} = \left\{ \frac{\hat{\sigma}^2 \nu(k)(\max - \min)}{\mu(k)^2 (\hat{\theta}_1 + \hat{\theta}_2)^2} \right\}^{1/5} n^{-1/5}, \tag{5·2}$$

where 'max' and 'min' denote the sample maximum and minimum of the design variable of interest, $\hat{\theta}_1$ and $\hat{\theta}_2$ were the coefficients of $x^2/2$ and $z^2/2$ obtained from a least squares regression of y on a constant, x, z, $x^2/2$, xz and $z^2/2$, while $\hat{\sigma}^2$ was obtained from the residuals of this regression. This rule is asymptotically optimal with respect to density-weighted integrated mean squared error, when x and z are independent and g is a quadratic function. Figure 1(a) shows $\hat{\alpha}_{Q_1}$ and Fig. 1(b) gives $\hat{\alpha}_{Q_2}$; both figures also contain 95%

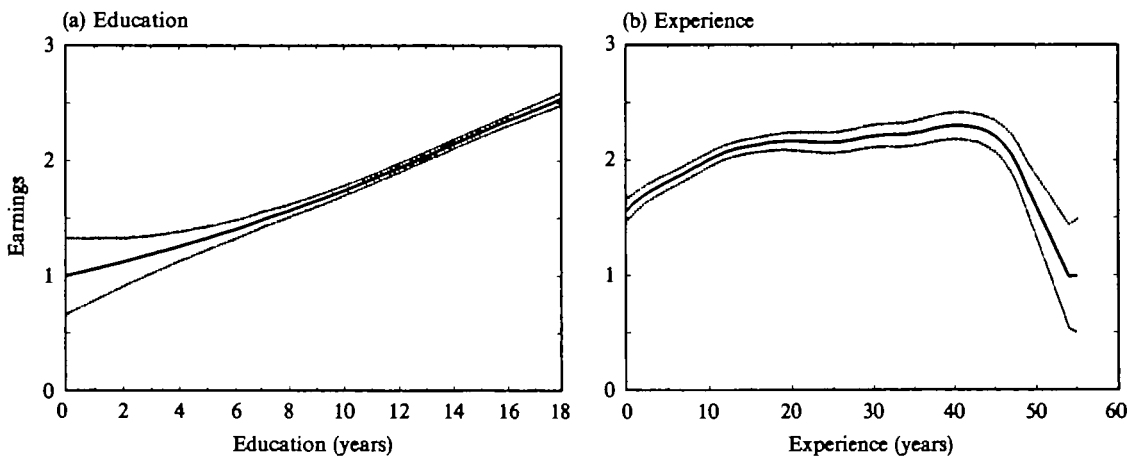

Fig. 1. Univariate effect of (a) education and (b) experience on the logarithm of wages, with 95% symmetric confidence intervals (dotted lines). A product of Gaussian kernels is used with common bandwidth $b = 3{\cdot}14$ for (a), and $b = 3{\cdot}93$ for (b).

pointwise confidence intervals $C_{0\cdot95} = \hat{\alpha}_{Q_1}(x_i) \pm 1\cdot96\hat{s}_i$, where

$$\hat{s}_i^2 = v(k)\hat{\sigma}^2 n^{-1} \sum_{j=1}^{n} \hat{f}^{-1}(x_i, z_j)\hat{f}_z(z_j)$$

in which $\hat{f}_z$ and $\hat{f}$ are the corresponding kernel estimates of f_z and f, while $\hat{\sigma}^2 = n^{-1}\sum\{y_i - \hat{g}(x_i, z_i)\}^2$. These intervals are consistent in the sense that pr $[E\{\hat{\alpha}_{Q_1}(x_i)\} \in C_{0\cdot95}] \to 0\cdot95$. Furthermore, when either $\beta_1, \beta_2 = 0$ or $nb^5 \to 0$, we can replace $E\{\hat{\alpha}_{Q_1}(x_i)\}$ by $\alpha_{Q_1}(x_i)$.

The effect of education on earnings is monotonic and essentially linear, while that of experience has a U-shaped pattern with rapidly increasing returns to the first 10 years of experience followed by slow but steady increase through 40 years followed by a decline in later years. This is consistent with other studies, although some have found a similar dip in the returns to education (Mukarjee & Stern, 1994). The sum of squared residuals was 112·7 for the additive model and 114·5 for the multiplicative, to be compared with 106·6 for the unrestricted regression and total sum of squared deviations from mean of 151·8. An interaction effect was fitted to the additive residuals using the uniform weighting function, but this added little explanatory power. We therefore interpret the effects as being primarily additive.

We investigated by simulation methods our procedure (4·1) for selecting between additive and multiplicative structures. The following designs were used:

Model 1: $y_i = x_i + z_i + \varepsilon_i$,
Model 2: $y_i = x_i z_i + \varepsilon_i$,
Model 3: $y_i = x_i^2 + z_i^2 + \varepsilon_i$,
Model 4: $y_i = x_i^2 z_i^2 + \varepsilon_i$,

where (x_i, z_i) were uniformly distributed on the unit square, while ε_i were independent and identically distributed $N(0, 0\cdot25^2)$. A total of 500 samples were generated for each experiment; we investigate $n = 50$ and $n = 100$. We used the estimation method described above in (5·1) and (5·2). Results are given in Table 1. The method seems to work moderately well even for these small sample sizes.

Table 1. *Percentage of samples correctly classified by* (4·1)

	Model			
n	1	2	3	4
50	79	69	87	63
100	86	86	95	80

6. Conclusion

Buja et al. (1989) discuss algorithms suitable for computing their estimator and the conditions under which these converge to a unique quantity. By contrast, our procedure is trivial to implement. What is not given by Buja et al. (1989) or by Breiman (1991) is a satisfactory description of the statistical properties of their method; this is due to the complicated dependence of their procedure on the marginal smoothing matrices. By contrast, our estimator, due to its simple form, is easier to analyse statistically: we obtained its limiting distribution and provided a result on optimal bandwidth choice. Finally, our approach is well suited for selecting between additive and multiplicative models because

both structures are consistently estimated by the same procedure. In view of the simple form of our estimator, it should be possible to develop analytical results for a variety of model selection and bandwidth selection procedures based on it.

Some problems remain in establishing convergence at rate $n^{2/5}$ for second order methods in dimensions higher than three. However, when sufficient bias reduction is used, the optimal one-dimensional convergence rate can be established for any dimension. This is reminiscent of the semiparametric literature, where often bias reduction is necessary to ensure $n^{\frac{1}{2}}$-consistency of Euclidean parameter estimates in the presence of high-dimensional nonparametric nuisance functions (Robinson, 1988). These issues remain to be investigated in future work.

ACKNOWLEDGEMENT

We would like to thank R. J. Carroll, D. R. Cox, W. Härdle, M. C. Jones and two referees for helpful comments.

REFERENCES

BERNDT, E. (1991). *The Practice of Econometrics.* Reading, MA: Addison-Wesley.

BREIMAN, L. (1991). The Π method for estimating multivariate functions from noisy data (with discussion). *Technometrics* **33**, 125–60.

BUJA, A., HASTIE, T. & TIBSHIRANI, R. (1989). Linear smoothers and additive models (with discussion). *Ann. Statist.* **17**, 453–555.

ENGLE, R. F., GRANGER, C. W. J., RICE, J. & WEISS, A. (1986). Semiparametric estimates of the relation between weather and electricity sales. *J. Am. Statist. Assoc.* **81**, 310–20.

FAN, J. (1992). Design-adaptive nonparametric regression. *J. Am. Statist. Assoc.* **87**, 998–1004.

FRIEDMAN, J. H. & STUETZLE, W. (1981). Projection pursuit regression. *J. Am. Statist. Assoc.* **76**, 817–23.

GORDON, L. & OLSHEN, R. A. (1980). Consistent nonparametric regression from recursive partitioning schemes. *J. Mult. Anal.* **10**, 611–27.

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

HASTIE, T. J. & TIBSHIRANI, R. J. (1990). *Generalized Additive Models.* London: Chapman and Hall.

JONES, M. C., DAVIES, S. J. & PARK, B. U. (1994). Versions of kernel-type regression estimators. *J. Am. Statist. Assoc.* **89**, 825–32.

MUKARJEE, H. & STERN, S. (1994). Feasible nonparametric estimation of multiargument monotone functions. *J. Am. Statist. Assoc.* **80**, 77–80.

ROBINSON, P. M. (1988). Root-*n*-consistent semiparametric regression. *Econometrica* **56**, 931–54.

RUPPERT, D. & WAND, M. P. (1994). Multivariate weighted least squares regression. *Ann. Statist.* To appear.

STONE, C. J. (1985). Additive regression and other nonparametric models. *Ann. Statist.* **13**, 689–705.

WAND, M. P. & JONES, M. C. (1993). Comparison of smoothing parameterizations in bivariate kernel density estimation. *J. Am. Statist. Assoc.* **88**, 520–9.

[*Received April* 1993. *Revised June* 1994]