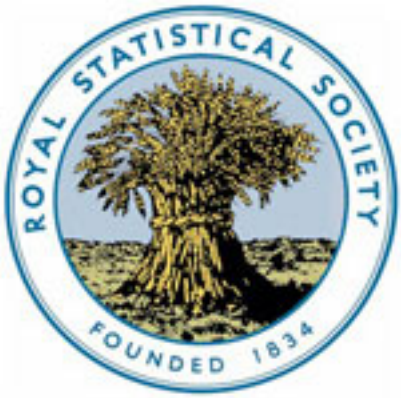


J. R. Statist. Soc. B (1988)
50, *No.* 3, *pp.* 413–436

Kernel Smoothing in Partial Linear Models

By PAUL SPECKMAN†

University of Missouri–Columbia, USA

[Received November 1986. Revised January 1988]

SUMMARY

Kernel smoothing is studied in partial linear models, i.e. semiparametric models of the form $y_i = \xi_i'\beta + f(t_i) + \varepsilon_i$ $(1 \leqslant i \leqslant n)$, where the ξ_i are fixed known p vectors, β is an unknown vector parameter and f is a smooth but unknown function. Two methods of estimating β and f are considered, one related to partial smoothing splines and the other motivated by partial residual analysis. Under suitable assumptions, the asymptotic bias and variance are obtained for both methods, and it is shown that estimating β by partial residuals results in improved bias with no asymptotic loss in variance. Application to analysis of covariance is made, and several examples are presented.

Keywords: ANALYSIS OF COVARIANCE; KERNEL SMOOTHING; PARTIAL LINEAR MODELS; SEMIPARAMETRIC MODELS

1. INTRODUCTION

Inference in semiparametric additive models has received considerable attention recently. In these models the mean response is assumed to be linearly related to one or more variables, but the relation to an additional variable or variables is not assumed to be easily parameterized. The specific model considered here is

$$y_i = \xi_i'\beta + f(t_i) + \varepsilon_i \quad (1 \leqslant i \leqslant n), \tag{1.1}$$

where the ξ_i are fixed known p vectors, β is an unknown vector of parameters and the t_i are in some bounded domain $D \subset \mathbb{R}^k$. The error terms ε_i will be assumed uncorrelated with mean zero and variance σ^2. We shall call $f(t)$ the smooth part of the model and assume that it represents a smooth unparameterized functional relationship. Because of its relation to the classical linear model, the semiparametric setting of equation (1.1) will be called a 'partial linear model'.

There have been several approaches to estimating β and f. One primary approach is the method of penalized least squares introduced by Engle *et al.* (1986), Green *et al.* (1985), Shiau *et al.* (1986) and Wahba (1984a, b) among others. Estimates are obtained by minimizing over β and g the quantity

$$\sum_{i=1}^{n} [y_i - \xi_i'\beta - g(t_i)]^2 + \lambda J(g), \tag{1.2}$$

where $J(g)$ is a functional chosen to penalize for roughness in the fitted g. If t is one dimensional with $D = [0, 1]$, for example, we can take

$$J(g) = \int_0^1 [g^{(m)}(t)]^2 \, dt,$$

which gives the cubic smoothing spline for $m = 2$. The constant λ is a 'tuning

† *Address for correspondence*: Department of Statistics, 222 Mathematical Sciences Building, University of Missouri–Columbia, Columbia, MO 65211, USA.

 0035–9246/88/50413 $2.00

parameter' chosen by the statistician for a suitable fit. Solving this minimization problem produces simultaneous estimates of $\boldsymbol{\beta}$ and f. Because equation (1.2) is the extension of the equation defining smoothing splines to the partial linear model, this estimate of f has been called a 'partial smoothing spline' by Wahba (1984b).

There have been other proposals for estimation in this context. Green *et al.* (1985) suggested a method for simultaneous estimation motivated by their penalized least squares approach in which rather arbitrary scatterplot smoothers can be used to estimate both $\boldsymbol{\beta}$ and f. They did not pursue this method, but it has been used, apparently independently, by Hastie and Tibshirani (1986) for inference in generalized additive models, and the idea forms one step in Breiman and Friedman's (1985) ACE algorithm. Other procedures for this problem include the projection method of Chen (1988) and a minimax approach due to Heckman (1986a).

Partial smoothing splines are attractive for several reasons. The principle of adding a penalty term to a sum of squares or more generally to a log-likelihood applies to a wide variety of linear and non-linear problems (see O'Sullivan (1986), for example). There is also a Bayesian interpretation to the method as in, for example, Shiller (1984), Green *et al.* (1985) or Eubank (1986). Most importantly, these researchers report that the method simply seems to work well. Less is known about the theory of partial smoothing splines, although Rice (1986) and Heckman (1986b) have made important contributions. In balanced cases of analysis of covariance, Heckman established asymptotic normality for the estimator of $\boldsymbol{\beta}$ and showed that its bias is asymptotically negligible. However, Rice showed that the bias of the estimator of $\boldsymbol{\beta}$ can asymptotically dominate the variance in unbalanced cases where ξ and t are correlated unless f is undersmoothed.

The startling negative result of Rice helped to motivate the study here. The results focus on kernel smoothing, defined in the next section, although the ideas apply as well to other linear smoothers. Kernel smoothing is well understood in nonparametric regression, but its theoretical properties have apparently not been studied in the partial linear model (1.1).

In Section 3 we present two methods of estimation for model (1.1). Both are based on methods for scatterplot smoothing in the simplified model

$$y_i = f(t_i) + \varepsilon_i \tag{1.3}$$

obtained from model (1.1) with $\boldsymbol{\beta} = \mathbf{0}$. Let $\mathbf{S}$ be the smoother matrix for ordinary spline smoothing in this case (corresponding to minimization of equation (1.2) with $\boldsymbol{\beta} = \mathbf{0}$), i.e. the $n \times n$ smoother matrix which transforms the vector of observations $\mathbf{y} = (y_1, \ldots, y_n)'$ into fitted values $\hat{\mathbf{y}} = \mathbf{S}\mathbf{y}$. Then we can show (see Section 3) that the estimators for $\boldsymbol{\beta}$ and $\mathbf{f} = (f(t_1), \ldots, f(t_n))'$ in equation (1.1) obtained from equation (1.2) are

$$\begin{aligned} \hat{\boldsymbol{\beta}} &= (\mathbf{X}'(\mathbf{I} - \mathbf{S})\mathbf{X})^{-1}\mathbf{X}'(\mathbf{I} - \mathbf{S})\mathbf{y} \\ \hat{\mathbf{f}} &= \mathbf{S}(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}), \end{aligned} \tag{1.4}$$

where $\mathbf{X} = (\xi_1, \ldots, \xi_n)'$ is the $n \times p$ design matrix for the parametric part of equation (1.1). More generally, Green *et al.* (1985) suggested replacing $\mathbf{S}$ by an arbitrary scatterplot smoother, and their approach is carried out here for kernel smoothers.

The second method treated here, motivated by consideration of partial regression plots, is to form partial residual vectors adjusting first for t by defining $\tilde{\mathbf{y}} = (\mathbf{I} - \mathbf{S})\mathbf{y}$

and $\tilde{\mathbf{X}} = (\mathbf{I} - \mathbf{S})\mathbf{X}$. In analogy with ordinary least squares (OLS), we can then define estimates

$$\hat{\boldsymbol{\beta}}_p = (\tilde{\mathbf{X}}'\tilde{\mathbf{X}})^{-1}\tilde{\mathbf{X}}'\tilde{\mathbf{y}}$$

and

$$\hat{\mathbf{f}}_p = \mathbf{S}(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}_p).$$

We adopt the notation $\hat{\boldsymbol{\beta}}_p$ here to denote the fact that the estimate is computed by regression on partial residuals. This approach was suggested independently by Denby (1984, 1986) and is closely related to the cross-validation method of Cuzick (1987). Further justification is given in Section 3.

The asymptotic results in Sections 4 and 5 detail the case for kernel smoothers. It is shown there that the first and second methods achieve comparable fits to f, but that $\hat{\boldsymbol{\beta}}$ from equation (1.4) has the bias problem noted by Rice (1986). However, the bias of $\hat{\boldsymbol{\beta}}_p$ is of lower order and asymptotically (at least) negligible with no need for undersmoothing f. This suggests that $\hat{\boldsymbol{\beta}}_p$ is more appropriate if the main interest is inference on $\boldsymbol{\beta}$. In view of the relationship between spline smoothing and kernel smoothing shown by Silverman (1985), the results here give further insight as well into partial smoothing splines. The ideas of Section 4 are not limited to kernel smoothers but apply in principle to arbitrary linear smoothers. For simplicity, however, only the case of kernel smoothing is discussed in detail.

Section 5 is devoted to a discussion on the automatic choice of bandwidth parameter for kernel smoothing in model (1.1). The results pertain to the method of generalized cross-validation (GCV) and show that the first-order properties of the GCV function are essentially unchanged by the addition of the parametric term $\xi_i'\boldsymbol{\beta}$ to the model. These results suggest that a data-driven choice for the bandwidth can be used in practice with the second estimator.

Section 6 contains three examples. The first analyses a data set given by Green *et al.* (1985) and uses kernel smoothing essentially to replicate their analysis. The second shows an outlier rejection feature of the method in an analysis of covariance, and the third demonstrates the bias of both methods on simulated data.

2. KERNEL SMOOTHING

We consider first the nonparametric regression problem of estimating f in equation (1.3). For simplicity, we assume that the domain of interest of f is $t \in [0, 1]$. The results to follow clearly hold for any bounded domain $[a, b]$ and also extend to the case where t is in higher dimensional Euclidean space.

One general formulation for linear smoothing is as follows. Suppose for each n there is a bivariate function $K_{nb}(s, t)$ indexed by a parameter b (the 'bandwidth') such that the estimate of $f(t)$ at fixed t is

$$\hat{f}_{nb}(t) = \sum_{i=1}^{n} K_{nb}(t, t_i)y_i. \tag{2.1}$$

The weight function (or delta sequence) $\{K_{nb}\}$ will be said to be of order v if for functions f with $f^{(v)}$ bounded and continuous on [0, 1] there exist bounded functions $h_1(t)$ and $h_2(t)$ such that

$$\begin{aligned} \text{bias}(\hat{f}_{nb}(t)) &= E[\hat{f}_{nb}(t)] - f(t) \\ &= b^v h_1(t) f^{(v)}(t) + o(b^v) \end{aligned} \tag{2.2a}$$

and

$$\operatorname{var}(\hat{f}_{nb}(t)) = \frac{h_2(t)\sigma^2}{nb}\,[1 + o(1)], \tag{2.2b}$$

where $o(1)$ denotes terms tending to zero uniformly on $[0, 1]$ as $b \to 0$ and $nb \to \infty$. Thus for a weight function K of order v, the mean-squared error (MSE) at a point is asymptotically

$$\text{MSE}(t) = \{b^{2v}h_1(t)^2[f^{(v)}(t)]^2 + h_2(t)\sigma^2/nb\}[1 + o(1)]. \tag{2.3}$$

One specific estimator that we have in mind here is the kernel estimator of Nadaraya (1964) and Watson (1964) defined by weights

$$K_{nb}(s, t_i) = w((s - t_i)/b) \Big/ \sum_{k=1}^{n} w((s - t_k)/b), \tag{2.4}$$

where the kernel (or window) $w(t)$ can be taken with finite support, say on $[-1, 1]$, with

$$\int_{-1}^{1} w(t)t^j \, dt = \begin{cases} 1 & (j = 0) \\ 0 & (1 \leqslant j \leqslant v - 1) \\ c_1 & (j = v) \end{cases}$$

for a non-zero constant c_1. Assuming for simplicity that observations are equally spaced with $t_k = k/n$ and ignoring potential difficulties at the boundaries, equation (2.2) holds with $h_1(t) = c_1/v!$ and

$$h_2 = \int_{-1}^{1} w(s)^2 \, ds.$$

There is an extensive literature on kernel smoothing with special attention given to specific forms of w, more efficient normalizations in the denominator of equation (2.4) and kernels modified at the boundaries. Some of the relevant references are Priestley and Chao (1972), Sacks and Ylvisaker (1978), Gasser and Müller (1979), Gasser *et al.* (1985) and Collomb (1981). The results here also apply to nearest neighbour kernels of the type considered by Stone (1977).

From equation (2.3), it follows that the best rate for the MSE at a point t is exactly $O(n^{-2v/(2v+1)})$ and is achieved by taking $b_n \sim n^{-1/(2v+1)}$. (The notation $a_n \sim b_n$ will be taken to mean that there exist constants $0 < m < M < \infty$ such that $m \leqslant a_n/b_n \leqslant M$ for all n.) This rate also holds for the integrated MSE and can be shown to be the best possible uniform rate in certain general nonparametric regression situations where f is assumed to be sufficiently smooth to admit v derivatives (see Stone (1982), Speckman (1985) or Nussbaum (1985)).

The smoothing spline estimator defined by equation (1.2) with $\boldsymbol{\beta} = \mathbf{0}$ is linear and satisfies equation (2.1). Unfortunately, condition (2.2a) does not hold unless a periodic smoother is taken and f has periodic boundary conditions (see Rice and Rosenblatt (1983)). Thus the results to follow will not apply directly to partial smoothing splines.

3. SMOOTHING IN PARTIALLY LINEAR MODELS

For motivation, suppose that f in model (1.1) can be parameterized as $(f(t_1), \ldots, f(t_n))' = \mathbf{T}\boldsymbol{\gamma}$, where $\mathbf{T}$ is an $n \times q$ matrix of full rank and $\boldsymbol{\gamma}$ is an additional parameter

vector. To assume that the $n \times (p+q)$ matrix $(\mathbf{X}, \mathbf{T})$ has full rank, we assume for simplicity that the unit vector $(1, \ldots, 1)'$ is in the span of $\mathbf{T}$ but not of $\mathbf{X}$. With matrix notation

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{T}\boldsymbol{\gamma} + \boldsymbol{\varepsilon},$$

the normal equations are

$$\begin{aligned} \mathbf{X}'\mathbf{X}\boldsymbol{\beta} &= \mathbf{X}'(\mathbf{y} - \mathbf{T}\boldsymbol{\gamma}), \\ \mathbf{T}\boldsymbol{\gamma} &= \mathbf{P}_{\mathbf{T}}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}), \end{aligned} \tag{3.1}$$

where $\mathbf{P}_{\mathbf{T}} = \mathbf{T}(\mathbf{T}'\mathbf{T})^{-1}\mathbf{T}'$ denotes projection on to the column span of $\mathbf{T}$. Green *et al.* (1985) proposed replacing $\mathbf{P}_{\mathbf{T}}$ in equation (3.1) by a (perhaps non-linear) smoother M and simultaneously solving

$$\begin{aligned} \boldsymbol{\beta} &= R(\mathbf{y} - \mathbf{f}) \\ \mathbf{f} &= M(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}), \end{aligned}$$

where R is a (possibly robust and non-linear) estimator of treatment effects. Taking M to be the matrix $\mathbf{K}$ from kernel smoothing and letting $R = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ produces the defining equations for the Green–Jennison–Seheult (GJS) (1985) estimators

$$\hat{\mathbf{f}}_{\mathrm{GJS}} = \mathbf{K}(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}_{\mathrm{GJS}}) \tag{3.2a}$$

and

$$\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}}_{\mathrm{GJS}} = \mathbf{X}'(\mathbf{y} - \mathbf{K}(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}_{\mathrm{GJS}})).$$

Hence

$$\hat{\boldsymbol{\beta}}_{\mathrm{GJS}} = (\mathbf{X}'(\mathbf{I} - \mathbf{K})\mathbf{X})^{-1}\mathbf{X}'(\mathbf{I} - \mathbf{K})\mathbf{y}. \tag{3.2b}$$

If M is a spline smoother (discretized or continuous) and $\mathbf{K}$ is the corresponding smoother matrix, equation (3.2b) is precisely the form of the estimator of $\boldsymbol{\beta}$ obtained by minimizing equation (1.2).

The second nonparametric method treated here can be motivated in several ways. By analogy with the partial correlation analysis of a subset of independent variables in OLS, let

$$\tilde{\mathbf{X}} = (\mathbf{I} - \mathbf{K})\mathbf{X}$$

and

$$\tilde{\mathbf{y}} = (\mathbf{I} - \mathbf{K})\mathbf{y}$$

denote the variables $\mathbf{X}$ and $\mathbf{y}$ after 'adjustment' for dependence on t. Assuming that $\tilde{\mathbf{X}}$ has full rank, we estimate $\boldsymbol{\beta}$ from partial residuals by

$$\hat{\boldsymbol{\beta}}_p = (\tilde{\mathbf{X}}'\tilde{\mathbf{X}})^{-1}\tilde{\mathbf{X}}'\tilde{\mathbf{y}}. \tag{3.3a}$$

If $\mathbf{K}$ is symmetric and idempotent, equations (3.2b) and (3.3a) will be identical (OLS) estimators, but in general they are distinct. Note also that the estimator of $\boldsymbol{\beta}$ in equation (3.3a) is formally the solution to the weighted least squares criterion

$$\underset{\boldsymbol{\beta}}{\text{minimize}} \; \|(\mathbf{I} - \mathbf{K})(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})\|^2,$$

where $\| \ \|$ denotes the Euclidean norm $\| \mathbf{v} \|^2 = \Sigma v_i^2$ for $\mathbf{v} = (v_1, \ldots, v_n)' \in \mathbb{R}^n$. To estimate f, we mimic equation (3.2a) and define

$$\hat{\mathbf{f}}_p = \mathbf{K}(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}_p). \tag{3.3b}$$

In both cases, the estimator of $f(t)$ will be

$$\hat{f}(t) = \sum_{i=1}^{n} K(t, t_i)(y_i - \xi_i'\hat{\boldsymbol{\beta}}), \tag{3.4}$$

where $\hat{\boldsymbol{\beta}}$ is defined by either equation (3.2b) or equation (3.3a).

A referee has pointed out that $\hat{\boldsymbol{\beta}}_p$ is the solution to equation (3.2b) if $\mathbf{K}$ is replaced by the smoother matrix $\mathbf{S} = \mathbf{K} + \mathbf{K}' - \mathbf{K}'\mathbf{K}$. (However, the resulting $\hat{f}$ from equation (3.3b) will not in general be $\hat{f}_p$.) If $\mathbf{K}$ is symmetric, $\mathbf{S}$ can be written as $\mathbf{S} = \mathbf{I} - (\mathbf{I} - \mathbf{K})^2$, the smoother derived from 'twicing' $\mathbf{K}$ (see Stuetzle and Mittal (1979)). As $\mathbf{S}$ ranges over a broad class of linear smoothers, we obtain both $\hat{\boldsymbol{\beta}}_p$ and $\boldsymbol{\beta}_{GJS}$ among other estimators. We could thus view the difference in the two methods as a difference in estimating f rather than $\boldsymbol{\beta}$.

The method of model (3.2) does have an advantage in models where there is more than one additive smooth term such as

$$y_i = f_1(t_{i1}) + f_2(t_{i2}) + \cdots + f_k(t_{ik}) + \varepsilon_i.$$

The ACE algorithm of Breiman and Friedman (1985) and the method of Hastie and Tibshirani (1986) estimate the f_i by iteratively solving extended versions of equation (3.2a). With certain assumptions, this method produces unique estimates of the f_i which are independent of the order of the f_i. In contrast, equation (3.3) is hierarchical in the sense that the adjustment is made for t first. Adjusting for $\mathbf{X}$ first would produce a different estimator.

The partial approach of equation (3.3) applies in principle to rather arbitrary linear or even non-linear smoothers. One advantage of $\hat{\boldsymbol{\beta}}_p$ is that it can be computed without iteration even if a non-linear smoother is used. Thus the method requires only a standard regression routine if a smoothing program is available to compute $\tilde{\mathbf{X}}$ and $\tilde{\mathbf{y}}$.

4. ASYMPTOTIC BEHAVIOUR

We consider an asymptotic setting to gain insight into the behaviour of these estimators. As with all asymptotics, the results here should perhaps be regarded only as suggestive, and small sample behaviour should be further investigated in individual cases. However, the asymptotics have been verified qualitatively in several sampling experiments including that reported in Section 6.

For simplicity, suppose that t is univariate as the extension to the multivariate case is straightforward. We henceforth let $\xi_i' = (x_{i1}, \ldots, x_{ip})$ in model (1.1) and write $\mathbf{X} = \{x_{ij}\}_{i=1,n;\, j=1,p}$. Following Rice (1986), the assumptions given here are motivated by the case in which the variables x_{ij} and t are related via the regression model $E(x_{ij} | t_i) = g_j(t_i)$, where the $g_j(t)$ $(1 \leqslant j \leqslant p)$ are smooth functions with v continuous derivatives and the $\{t_i\}$ are equally spaced or more generally have an asymptotic density $p(t)$. In particular, suppose

$$x_{ij} = g_j(t_i) + \eta_{ij} \quad (1 \leqslant i \leqslant n,\ 1 \leqslant j \leqslant p), \tag{4.1}$$

for η_{ij} independent mean zero random variables independent of the ε_i. An important special case is the design problem where x_{ij} is zero or unity. For example, in analysis of covariance when t is the covariate and $x_{ij} = 1$ denotes the presence of some treatment level randomly assigned to the ith subject, $g_j(t_i) = P[x_{ij} = 1 | t_i]$. If there is no relationship between treatment level and covariate t, g_j will be constant as in Heckman (1986b). However, a systematic trend resulting in unbalanced allocation to treatment could result in non-constant g_j.

Using vector notation, write

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{f} + \boldsymbol{\varepsilon} \tag{4.2}$$

where $\mathbf{y} = (y_1, \ldots, y_n)'$, $\mathbf{f} = (f(t_1), \ldots, f(t_n))'$ and $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n)'$, and let

$$\mathbf{X} = \mathbf{g} + \boldsymbol{\eta} \tag{4.3}$$

where we now write $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_p)$, $\mathbf{g} = (\mathbf{g}_1, \ldots, \mathbf{g}_p)$, and $\boldsymbol{\eta} = (\boldsymbol{\eta}_1, \ldots, \boldsymbol{\eta}_p)$ with $\mathbf{x}_j = (x_{1j}, \ldots, x_{nj})'$, $\mathbf{g}_j = (g_j(t_1), \ldots, g_j(t_n))'$ and $\boldsymbol{\eta}_j = (\eta_{1j}, \ldots, \eta_{nj})'$.

We assume the existence of a smoother matrix $\mathbf{K} = \{K_{nb}(t_i, t_j)\}$ (suppressing the dependence on n and b), perhaps given by equation (2.4), such that equation (2.2) holds for non-linear regression models with $f^{(\nu)}$ continuous. Throughout, tr(**A**) denotes the trace of a matrix $\mathbf{A} = \{A_{ij}\}_{i=1,\, n;\, j=1,\, n}$, namely

$$\operatorname{tr}(\mathbf{A}) = \sum_{i=1}^{n} A_{ii}.$$

The assumptions to follow are consequences of the regression model outlined above. Justification under equation (4.1) with kernel smoothing is provided here and in Appendix A. However, the author believes that the assumptions can realistically hold in other contexts as well when the regression relation between x_{ij} and t_i may be difficult to justify but where the η_{ij} can be taken to behave as if they were 'random'. A possible example is the analysis of the field trial in Section 6.

For convenience, all the conditions required for the main theorems are stated here. The dependence of **K** on n and b is suppressed, and it is understood that $n \to \infty$, $b \to 0$ and $nb \to \infty$.

Assumptions.

(a) $n^{-1}\boldsymbol{\eta}'\boldsymbol{\eta} \to \mathbf{V}$

where $\mathbf{V} = \{V_{ij}\}$ is positive definite.

(b) $\operatorname{tr}(\mathbf{K}'\mathbf{K}) = \sum_{i=1}^{n} \sum_{j=1}^{n} K_{ij}^2 = O(b^{-1})$.

(c) $\|\mathbf{K}\boldsymbol{\eta}_j\|^2 = O(b^{-1}) = \|\mathbf{K}'\boldsymbol{\eta}_j\|^2 \quad (1 \leqslant j \leqslant p)$.

(d) $\tilde{g}_j(t_i) = b^{\nu} h_1(t_i) g_j^{(\nu)}(t_i) + o(b^{\nu}) \quad (1 \leqslant i \leqslant n,\ 1 \leqslant j \leqslant p)$.

(e) $\|(\mathbf{I} - \mathbf{K})\mathbf{f}\|^2 = \|\tilde{\mathbf{f}}\|^2 = O(nb^{2\nu})$.

(f) $n^{-1}\boldsymbol{\eta}'\tilde{\mathbf{f}} = O(n^{-1/2}b^{\nu})$.

(g) There is a probability density $p(t)$ on $[0, 1]$ such that

$$n^{-1}\sum_{i=1}^{n} c(t_i) \to \int_0^1 c(t)\,p(t)\,dt$$

as $n \to \infty$ for any continuous function $c(t)$.

If the regression model relating $\mathbf{x}_j$ to t of equation (4.1) holds the $O(\cdot)$ terms are to be interpreted as bounds holding in probability. Specifically, suppose the rows of $\boldsymbol{\eta}$, $(\eta_{i1}, \ldots, \eta_{ip})$, $i = 1, \ldots, n$, are independent identically distributed random vectors with mean zero and finite variance–covariance matrix $\mathbf{V} = \{V_{ij}\}$. Then assumption (a) holds in probability. From equation (2.2b),

$$\sum_{i=1}^{n} \operatorname{var}(\hat{f}(t_i)) = \sigma^2 \operatorname{tr}(\mathbf{K}'\mathbf{K}) = O(b^{-1});$$

hence assumption (b) holds. Assumption (c) follows from the random model because $E\|\mathbf{K}\boldsymbol{\eta}_j\|^2 = E\|\mathbf{K}'\boldsymbol{\eta}_j\|^2 = V_{jj}\operatorname{tr}(\mathbf{K}'\mathbf{K})$. Equation (2.2a) applied to g_j gives assumption (d), and assumption (e) is also a direct consequence of equation (2.2a). Moreoever, under equation (4.1) $E(\boldsymbol{\eta}_j'\tilde{\mathbf{f}}) = 0$ and $\operatorname{var}(\boldsymbol{\eta}_j'\tilde{\mathbf{f}}) = V_{jj}\|\tilde{\mathbf{f}}\|^2$, so assumption (f) follows from assumption (e). Assumption (g) holds if, for example, the $\{t_i\}$ are equally spaced (with $p(t) = 1$), if the $\{t_i\}$ are $n$tiles of a probability density p, or if the $\{t_i\}$ are a random sample with density p.

Two additional assumptions discussed in Appendix A will be used for proving theorems 4–6:

(h) $\operatorname{tr}(\mathbf{K}) = O(b^{-1})$;

(i) $\max_i \sum_j |K_{ij}| = O(1)$;

$\max_j \sum_i |K_{ij}| = O(1)$,

where both bounds hold uniformly as $n \to \infty$, $b \to 0$ and $nb \to \infty$.

In the following theorems, expectation is always understood to be with respect to the distribution of ε and conditioned on the observed $\mathbf{X}$ and $\{t_i\}$. In particular, under model (4.1), the ε_i and the η_{ij} are assumed independent. For this case the $O(\cdot)$ terms should be interpreted as holding in probability with respect to the distribution of the η_{ij}.

Theorem 1. Under assumptions (a)–(g), if $n \to \infty$ and $b \to 0$, then

$$E(\hat{\boldsymbol{\beta}}_{\mathrm{GJS}}) - \boldsymbol{\beta} = b^{\nu}\mathbf{V}^{-1}\int_0^1 \mathbf{g}(t)\,f^{(\nu)}(t)\,h_1(t)\,p(t)\,dt + O(b^{\nu}n^{-1/2}). \tag{4.4a}$$

If in addition $nb^2 \to \infty$, then

$$\operatorname{var}(\hat{\boldsymbol{\beta}}_{\mathrm{GJS}}) = n^{-1}\sigma^2\mathbf{V}^{-1} + n^{-2}\sigma^2\mathbf{V}^{-1}\mathbf{g}'(\mathbf{I}-\mathbf{K})(\mathbf{I}-\mathbf{K})'\mathbf{g}\mathbf{V}^{-1} + o(n^{-1}). \tag{4.4b}$$

If $\mathbf{K}$ is symmetric and these conditions hold,

$$\operatorname{var}(\hat{\boldsymbol{\beta}}_{\mathrm{GJS}}) = \sigma^2 n^{-1}\mathbf{V}^{-1} + o(n^{-1}). \tag{4.4c}$$

Theorem 2. Under assumptions (a)–(g), if $n \to \infty$ and $b \to 0$, then

$$E(\hat{\boldsymbol{\beta}}_p) - \boldsymbol{\beta} = b^{2v}\mathbf{V}^{-1} \int_0^1 \mathbf{g}^{(v)}(t)\, f^{(v)}(t)\, h_1(t)^2\, p(t)\, dt + o(b^{2v}) + O(b^v(nb)^{-1/2}). \quad (4.5a)$$

If in addition $nb^2 \to \infty$ and $nb^{4v} \to 0$,

$$\operatorname{var}(\hat{\boldsymbol{\beta}}_p) = \sigma^2 n^{-1}\mathbf{V}^{-1} + o(n^{-1}). \quad (4.5b)$$

These theorems are proven in Appendix A.

Remarks. The conditions of theorems 1 and 2 hold for $b \sim n^{-1/(2v+1)}$, the usual 'optimal' rate for smoothing f in equation (1.3). With this rate for symmetric $\mathbf{K}$, equations (4.4c) and (4.5b) show that var $(\hat{\boldsymbol{\beta}}_{GJS})$ and $\operatorname{var}(\hat{\boldsymbol{\beta}}_p)$ are equivalent and both $O(1/n)$, the rate of convergence for parametric estimation. If $\mathbf{K}$ is not symmetric, the results of theorem 1 are not as precise, but it can be shown using the methods of Appendix A that $n^{-1}\mathbf{V}^{-1}\mathbf{g}'(\mathbf{I}-\mathbf{K})(\mathbf{I}-\mathbf{K})'\mathbf{g}\mathbf{V}^{-1} = O(1)$, so again $\operatorname{var}(\hat{\boldsymbol{\beta}}_{GJS}) = O(1/n)$ from equation (4.4b). When β is a scalar it is plausible that $\operatorname{var}(\hat{\boldsymbol{\beta}}_{GJS}) \leqslant \operatorname{var}(\hat{\boldsymbol{\beta}}_p)$. This is observed in example 3 of Section 6, but the author has not been able to verify this in the general case of asymmetric $\mathbf{K}$.

The important conclusion of these theorems regards the bias of both estimators. If g is non-vanishing (i.e. x and t are correlated), $\hat{\boldsymbol{\beta}}_p$ can have substantially smaller bias than $\hat{\boldsymbol{\beta}}_{GJS}$. This is the problem of 'coccurvity' referred to by Hastie and Tibshirani (1986). Since $\operatorname{bias}(\hat{\boldsymbol{\beta}}_{GJS}) \sim b^v$, $\operatorname{bias}(\hat{\boldsymbol{\beta}}_{GJS}) \gg n^{-1/2}$ if the usual optimal bandwidth $b \sim n^{-1/(2v+1)}$ is used, and the square of the bias dominates the variance of the estimator. In contrast, $\operatorname{bias}(\hat{\boldsymbol{\beta}}_p) = O(b^{2v}) + O(b^v n^{-1/2}) = O(n^{-2v/(2v+1)}) \ll n^{-1/2}$ if $b \sim n^{-1/(2v+1)}$, so the square bias of $\hat{\boldsymbol{\beta}}_p$ is asymptotically negligible compared with its variance.

The results of theorem 1 parallel the conclusions in Rice (1986) for partial smoothing splines. He found exactly the same phenomena, namely parametric convergence for variance but nonparametric convergence of bias. The rates of theorem 1 do not apply directly to smoothing splines because equation (2.2a) does not hold without extraneous boundary conditions on f (see Rice and Rosenblatt (1983)), but the main idea of the proof is applicable and details will be carried out elsewhere. As a referee noted, it appears that the difference in performance between $\hat{\boldsymbol{\beta}}_{GJS}$ and $\hat{\boldsymbol{\beta}}_p$ is due to the fact that $\mathbf{K}$ is not a projection. It is possible to smooth data by projecting on to a subspace (whose dimension is allowed to grow with n). Such smoothers typically do not satisfy equation (2.2a), so the proofs of theorems 1 and 2 again are not directly applicable. However, Eubank and Speckman (1988) have recently shown that the estimator of $\boldsymbol{\beta}$ obtained by smoothing with projections on a particular combination of polynomial and trigonometric terms yields the desirable properties of theorem 2.

These results do not imply that $\hat{\boldsymbol{\beta}}_{GJS}$ is seriously biased in all cases because the integral term in equation (4.4a) can be small or we can undersmooth by taking $b \sim n^{-1/2v}$, for example. As Rice (1986) observed, the danger is that $\hat{\boldsymbol{\beta}}_{GJS}$ could be seriously biased if a method such as cross-validation is used which attempts to choose a b value to minimize the average mean-square error. Results in Section 5 support these observations. Estimation by $\hat{\boldsymbol{\beta}}_p$, however, appears safer and the potential bias problem is diminished with little if any loss in variance. Moreover, the partial kernel method with a single bandwidth b gives good estimates of both $\boldsymbol{\beta}$ and f in model (1.1), so the method would be suitable when both terms are of interest.

Heckman (1986b) considered the case where $g_i(t) \equiv$ constant, i.e. the case of no relationship between x_{ij} and t_j, and found asymptotic zero bias when $\boldsymbol{\beta}$ is estimated by partial smoothing splines. A balanced analysis of covariance design is an example of such a situation. It is not difficult to show that both $\hat{\boldsymbol{\beta}}_{GJS}$ and $\hat{\boldsymbol{\beta}}_p$ are unbiased in exactly balanced designs although the details will not be given here. However, some insight for balanced designs can be obtained easily if $\mathbf{K}$ is symmetric. Since $\mathbf{I}-\mathbf{K} = (\mathbf{I}-\mathbf{K})'$, equation (3.2b) can be rewritten as

$$\hat{\boldsymbol{\beta}}_{GJS} = (\mathbf{X}'\tilde{\mathbf{X}})^{-1}\tilde{\mathbf{X}}'\mathbf{y},$$

and the methods of the proof of theorem 1 can be used to obtain

$$E(\hat{\boldsymbol{\beta}}_{GJS}) - \boldsymbol{\beta} = b^{\nu}\mathbf{V}^{-1}\int_0^1 \mathbf{g}^{(\nu)}(t)\, f(t)\, h_1(t)\, p(t)\, dt + O(b^{\nu}n^{-1/2}). \tag{4.6}$$

Thus, if there is no asymptotic relationship between the x_{ij} and t_i so that $g_j(t) \equiv$ constant for $(1 \leqslant j \leqslant p)$, it follows that $g_j^{(\nu)}(t) \equiv 0$ and the bias is negligible. In kernel regression, $\mathbf{K}$ is typically not symmetric. However, for smoothing splines (either the continuous form or the discretized version of Green *et al.* (1985)), the matrix $\mathbf{K}$ is symmetric so that the extension of theorem 1 to equation (4.6) again suggests the corresponding result for smoothing splines.

Kernel estimation of f itself behaves asymptotically like equation (2.2) even in the presence of the parametric term $\boldsymbol{\beta}$. To simplify the statement of the next theorem, let $\hat{f}_0(t)$ denote the estimator of $f(t)$ that would be obtained by kernel smoothing if $\boldsymbol{\beta}$ were known exactly, i.e.

$$\hat{f}_0(t) = \sum_{i=1}^{n} K(t, t_i)(y_i - \boldsymbol{\xi}_i'\boldsymbol{\beta}).$$

(Note the connection with equation (3.4)). We then have the following result also proven in Appendix A.

Theorem 3. Under assumptions (a)–(g), if $n \to \infty$, $b \to 0$, $nb^2 \to \infty$ and $nb^{4\nu} \to 0$,

$$\text{bias}(\hat{f}_{GJS}(t)) = \text{bias}(\hat{f}_0(t)) + \mathbf{g}(t)'\text{bias}(\hat{\boldsymbol{\beta}}_{GJS})[1 + o(1)] = O(b^{\nu}), \tag{4.7a}$$

where $\mathbf{g}(t)' = (g_1(t), \ldots, g_p(t))$, and

$$\text{bias}(\hat{f}_p(t)) = \text{bias}(\hat{f}_0(t))[1 + o(1)] = O(b^{\nu}). \tag{4.7b}$$

For both GJS and partial kernel smoothing,

$$\text{var}(\hat{f}(t)) = \text{var}(\hat{f}_0(t))[1 + o(1)] = O((nb)^{-1}). \tag{4.8}$$

Remark. $\mathbf{K}$ is not assumed symmetric here.

As in the case studied by Heckman (1986b), we also have asymptotic normality.

Theorem 4. Under assumption (*i*) and the conditions of theorem 3, if the components of $\mathbf{X}$ are founded or if model (4.1) holds with $E|\eta_{ij}|^{2+\delta} < C < \infty$ for some $\delta > 0$ and $C < \infty$ $(1 \leqslant i \leqslant n,\ 1 \leqslant j \leqslant p)$, then

$$n^{1/2}[\hat{\boldsymbol{\beta}}_p - E(\hat{\boldsymbol{\beta}}_p)] \xrightarrow{D} N(0, \sigma^2\mathbf{V}^{-1}).$$

5. CROSS-VALIDATION

Various methods have been suggested for objectively choosing the bandwidth parameter in kernel smoothing from the data alone including cross-validation and GCV. Ordinary cross-validation based on omitting one observation at a time has been studied in detail by Härdle and Marron (1985). GCV, originally introduced by Craven and Wahba (1979) and Golub *et al.* (1979) for smoothing splines and ill-posed problems, has been applied to kernel smoothing by Rice (1984a). Green (1985) discussed the application of cross-validation to partial linear models, especially in the context of field trials, and he gave appropriate computational formulae. Because of its simplicity, we use GCV here and present theoretical results to support its application.

GCV attempts to provide a data-based estimate of the b which minimizes the unobservable

$$R(b) = n^{-1} \| E(\mathbf{y}) - \hat{\mathbf{y}} \|^2.$$

For model (1.1), suppose $\hat{\mathbf{y}}$ is the estimator of $E(\mathbf{y})$ (depending on b) given by

$$\left.\begin{aligned} \hat{\mathbf{y}} &= \mathbf{X}\hat{\boldsymbol{\beta}} + \hat{\mathbf{f}} \\ &= \mathbf{X}\hat{\boldsymbol{\beta}} + \mathbf{K}(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}) \\ &= \mathbf{K}\mathbf{y} + \tilde{\mathbf{X}}\hat{\boldsymbol{\beta}}, \end{aligned}\right\} \tag{5.1}$$

where $\hat{\boldsymbol{\beta}}$ is given by either equation (3.2b) or equation (3.3a). The definition of GCV requires the 'hat matrix' for equations (5.1), namely the matrix $\mathbf{A}(b)$ such that $\hat{\mathbf{y}} = \mathbf{A}(b)\mathbf{y}$. The GCV function is then taken to be

$$\mathrm{GCV}(b) = \frac{\mathrm{RSS}(b)}{[1 - n^{-1} \operatorname{tr}(\mathbf{A}(b))]^2},$$

where RSS(b) denotes the average residual sum of squares

$$\mathrm{RSS}(b) = n^{-1} \| (\mathbf{I} - \mathbf{A}(b))\mathbf{y} \|^2.$$

If we let $\mathbf{P}_{\tilde{\mathbf{X}}} = \tilde{\mathbf{X}}(\tilde{\mathbf{X}}'\tilde{\mathbf{X}})^{-1}\tilde{\mathbf{X}}'$ and $\mathbf{Q}_{\mathbf{X}} = \tilde{\mathbf{X}}(\mathbf{X}'\tilde{\mathbf{X}})^{-1}\mathbf{X}'$, equations (3.2b) and (3.3a) imply respectively

$$\mathbf{A}_{\mathrm{GJS}}(b) = \mathbf{K} + \mathbf{Q}_{\mathbf{X}}(\mathbf{I} - \mathbf{K}) \tag{5.2a}$$

and

$$\mathbf{A}_p(b) = \mathbf{K} + \mathbf{P}_{\tilde{\mathbf{X}}}(\mathbf{I} - \mathbf{K}). \tag{5.2b}$$

The GCV method is to select the bandwidth b which minimizes GCV(b).

We begin with a result on the best possible rate for $E(R(b))$ by relating it to kernel smoothing with no parametric part. Let

$$R_0(b) = n^{-1} \| E(\mathbf{y}) - \mathbf{K}\mathbf{y} \|^2$$

be the mean-square error under model (1.3) (i.e. with $\boldsymbol{\beta} = \mathbf{0}$), and define similar quantities for the GJS smoothing and partial estimators in model (1.1) by

$$R_{\mathrm{GJS}}(b) = n^{-1} \| E(\mathbf{y}) - \mathbf{X}\hat{\boldsymbol{\beta}}_{\mathrm{GJS}} - \hat{\mathbf{f}}_{\mathrm{GJS}} \|^2$$

and

$$R_p(b) = n^{-1} \| E(\mathbf{y}) - \mathbf{X}\hat{\boldsymbol{\beta}}_p - \hat{\mathbf{f}}_p \|^2.$$

An approximation for $E(R_0(b))$ can be obtained by summing equation (2.3) over the observation points to obtain

$$E(R_0(b)) = [c_1 b^{2v} + c_2 (nb)^{-1}][1 + o(1)], \tag{5.3}$$

where c_1 is a constant depending on $f^{(v)}$ and c_2 is a constant depending on σ^2. The value of b minimizing equation (5.3) is easily seen to be proportional to $n^{-1/(2v+1)}$ yielding the exact 'optimal' rate $E(R_0(b)) \sim n^{-2v/(2v+1)}$. The following theorem implies that this rate is also optimal for $E(R_{\text{GJS}}(b))$ and $E(R_p(b))$.

Theorem 5. Under assumptions (a)–(i), if $n \to \infty$, $b \to 0$ and $nb \to \infty$, then

$$E(R_p(b)) = E(R_0(b))[1 + o(1)]$$

and

$$\begin{aligned} E(R_{\text{GJS}}(b)) &= E(R_0(b))[1 + o(1)] + n^{-1}[\| \tilde{\mathbf{X}} \operatorname{bias}(\hat{\boldsymbol{\beta}}_{\text{GJS}}) \|^2 - 2\tilde{\mathbf{f}}'\tilde{\mathbf{X}} \operatorname{bias}(\hat{\boldsymbol{\beta}}_{\text{GJS}})] \\ &= E(R_0(b))[1 + o(1)] + O(b^{2v}). \end{aligned}$$

Theorem 6. Under the same assumptions, for either GJS or partial estimation,

$$E(\text{GCV}(b)) = \sigma^2 + E(R(b))[1 + o(1)]. \tag{5.4}$$

The proofs of these theorems are in Appendix A. Theorem 6 is an application of the GCV theorem of Craven and Wahba (1979) and Golub *et al.* (1979) and implies that the minimizer of $E(\text{GCV}(b))$ is essentially equivalent to the minimizer of $E(R(b))$ for both GJS and partial kernel estimators. This argument is of course heuristic in that equation (5.4) only shows that GCV(b) is an estimator of $E(R(b))$ with a nearly constant bias. However, Rice (1984a) has given a rigorous proof for using GCV and several variants in kernel smoothing in model (1.3), so the same results should be expected to hold here.

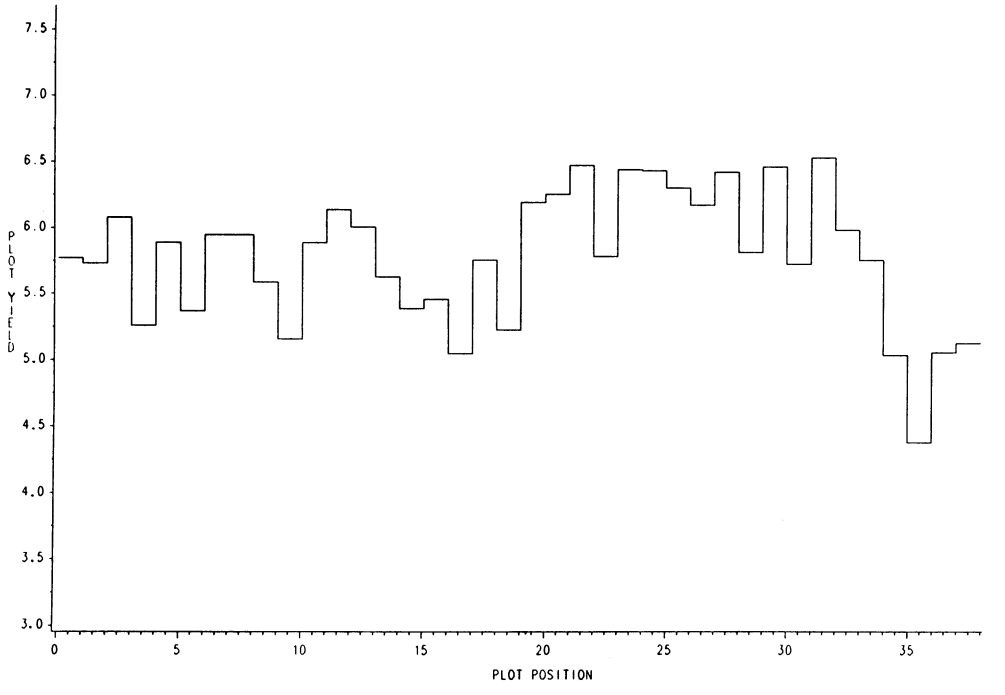
6. EXAMPLES

Example 1. The first data set is from a Rothamsted mildew control experiment analysed by Green *et al.* (1985). The data can be found in Draper and Guttman (1980). Fig. 1 contains a plot of the raw data and fitted values obtained by the partial kernel method. In the experiment, four mildew control spray applications were tested: none, early spring, late spring and repeated application. The experiment was arranged as a single column of 38 plots in nine blocks of four plots each with an extra plot on each end. Block position is taken as the covariate.

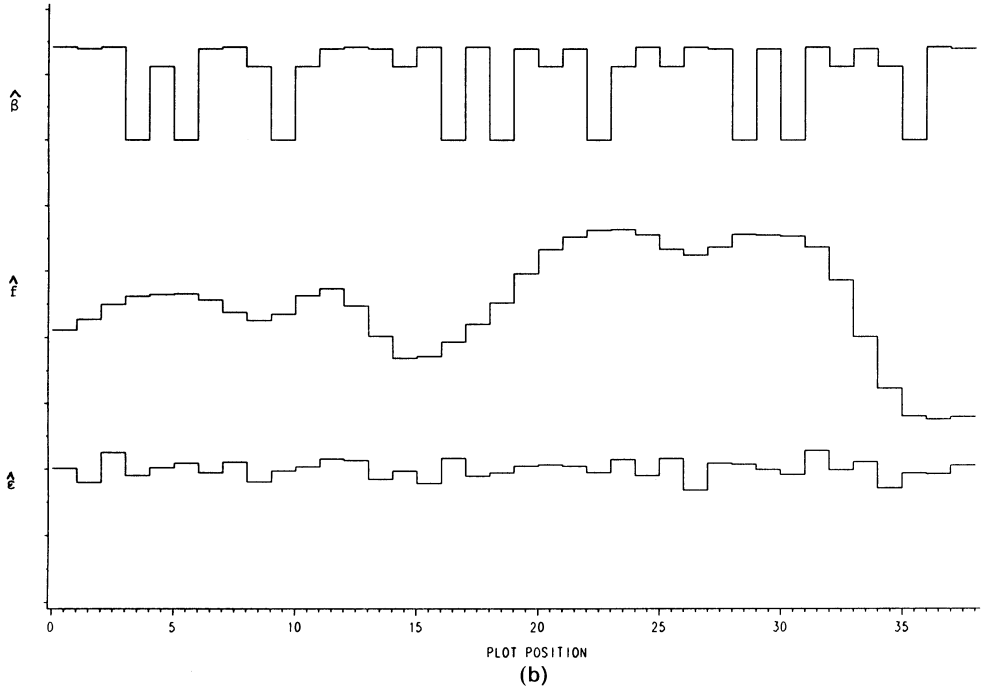
The fit in Fig. 1 was obtained using the quartic kernel (see Tapia and Thompson (1978), p. 60),

$$w(u) = \begin{cases} (1-u^2)^2, & |u| \leqslant 1, \\ 0, & |u| > 1. \end{cases} \tag{6.1}$$

For simplicity K was computed everywhere by equation (2.4) and no bias adjustment was made at the ends of the interval of estimation. This potentially could cause serious bias problems as noted by Gasser and Müller (1979) and Rice (1984b), but in the applications here the use of bias-adjusted smoothers did not appreciably affect the results, so unadjusted smoothers were used.



(a)



(b)

Fig. 1. (a) Plot yields (in tonnes per hectare) for the mildew control experiment; (b) partial kernel decomposition with $b = 2.17$ (vertical scale as in (a))

The bandwidth chosen by GCV for this example was $b = 2.17$, a fairly small value indicating the high variability in the plot position effect. The estimates obtained here using the partial regression estimator closely resemble those obtained by Green *et al.* (1985), and Fig. 1(b) is virtually identical with their decomposition. As they pointed out, blocking in the original groups of four is not appropriate.

The asymptotic normality from theorem 3 justifies an approximate F test. Let

$$\hat{\boldsymbol{\beta}}_p = \mathbf{U}'\mathbf{y},$$

where

$$\mathbf{U}' = (\tilde{\mathbf{X}}'\tilde{\mathbf{X}})^{-1}\tilde{\mathbf{X}}'(\mathbf{I} - \mathbf{K}),$$

and define

$$\text{SS(treatment)} = \hat{\boldsymbol{\beta}}_p'(\mathbf{U}'\mathbf{U})^{-1}\hat{\boldsymbol{\beta}}_p.$$

For an estimate of error, we let

$$\text{MSE}(b) = \frac{\text{RSS}(b)}{\text{tr}(\mathbf{I} - \mathbf{A}(b))'(\mathbf{I} - \mathbf{A}(b))}. \tag{6.2}$$

Using the techniques of Section 5, this can be shown to be an estimator of σ^2 with positive but asymptotically negligible bias, so the resulting 'F' statistic can be expected to be conservative. Monte Carlo trials not reported here support this conjecture. The denominator of equation (6.2) is 14.4, a term which can be taken as the 'degrees of freedom'. The values of MS(treatment) and MSE(b) are 0.682 and 0.0121 respectively yielding a ratio of 56.4, quite close to the 'F ratio' of 54.94 computed from Table 3 of Green *et al.* (1985) and obviously highly significant.

Example 2. The data for this example are from an experiment to determine whether a regular programme of mouthwash with a common brand of analgesic is effective in treating a type of gum disease (Platt, 1985). 30 volunteers with varying degrees of gum disease were randomized to two groups of 15 each, a control group using only a water rinse for mouthwash and an experimental treatment group. Base-line measurements were taken followed by weekly measurements over the three weeks of the study. Table 1 contains the base-line and week-three data on one of the variables, SBI, a measurement indicating gum shrinkage for which lower numbers indicate better health. The initial hypothesis was that the treatment group would show greater improvement than the control group, where improvement is measured by difference between the base-line and week-three SBI. However, the t test comparing the differences in the two groups actually suggested greater improvement for the controls ($t = -2.0803$ on 28 degrees of freedom).

Fig. 2 contains a plot of week-three SBI versus base-line SBI. It is clear that the control group received a disproportionate number of subjects with high base-line values, and a more appropriate analysis of the effect of the mouthwash treatment should account for the base-line value as a covariate. Thus the data were analysed with the partial kernel estimator using the kernel of equation (6.1) and $b = 0.38$ from GCV. The dependent variable was week-three SBI, the covariate was base-line SBI, and x_i was set to unity for the treatment group and zero for the control group. The estimated treatment effect was $\hat{\boldsymbol{\beta}} = 0.040$ with $F = 0.59$ on 1 and 24.7 degrees of freedom. This analysis shows no significant difference and adjusts for the obvious unlucky randomization. Fig. 2 shows the estimated regression functions, $\hat{g}(t)$ and

TABLE 1
Mouthwash data from example 2

Week-3 SBI	*Base-line SBI*	x_i	$\tilde{x}_i$
0.39	0.25	1	0.506
0.19	0.25	0	−0.494
0.30	0.30	1	0.485
0.15	0.33	0	−0.533
0.14	0.34	0	−0.539
0.15	0.38	1	0.439
0.17	0.40	0	−0.571
0.19	0.45	1	0.408
0.18	0.46	1	0.404
0.33	0.48	0	−0.603
0.29	0.54	0	−0.623
0.45	0.55	1	0.374
0.41	0.57	1	0.368
0.29	0.59	1	0.363
0.09	0.60	0	−0.640
0.65	0.63	1	0.354
0.45	0.63	1	0.354
0.15	0.63	1	0.354
0.44	0.65	1	0.353
0.51	0.65	0	−0.647
0.50	0.66	1	0.353
0.18	0.69	1	0.355
0.54	0.71	0	−0.640
0.47	0.71	0	−0.640
0.51	0.75	1	0.377
0.42	0.99	0	−0.201
0.42	0.99	0	−0.201
0.69	1.32	0	0.000
0.57	1.42	0	0.000
0.31	1.72	0	0.000

$\hat{g}(t) + \hat{\beta}$, superimposed on the scatterplot. Separate curves are drawn for the estimated treatment and control groups with constant difference $\hat{\beta}$.

The decomposition displayed in Fig. 2 appears realistic for data with base-line values t_i less than about 1.0, but there is clearly no way to tell whether the strict additive model (1.1) with constant variance and no interaction is even approximately valid for the outlying observations in the control group with large base-line values. Fortunately this potential problem with the model has little effect on the estimate $\hat{\beta}_p = \Sigma \tilde{x}_i \tilde{y}_i / \Sigma \tilde{x}_i^2$. As seen from Table 1, $\tilde{x}_i = 0$ for $t_i \geqslant 1.2$, and the two observations with $t_i = 0.99$ (this is not a misprint) have $\tilde{x}_i = -0.201$. The remaining data points are treated roughly equally. Thus observations which are isolated in the covariate are downweighted or eliminated entirely and do not contribute to the estimate of β, while data points which are not isolated receive relatively high weight. This outlier rejection property is intuitively reasonable because there is no way to tell whether the response for an observation with an isolated t_i is due to the parametric or nonparametric part of the model.

This example demonstrates one important aspect of the partial kernel analysis of covariance: the automatic removal of outliers. The usual linear analysis of covariance with the outliers removed produced essentially the same result. We note that the partial kernel (or other semiparametric) method may be of particular value in applications where the covariate has dimension greater than unity and where outliers are more difficult to identify.

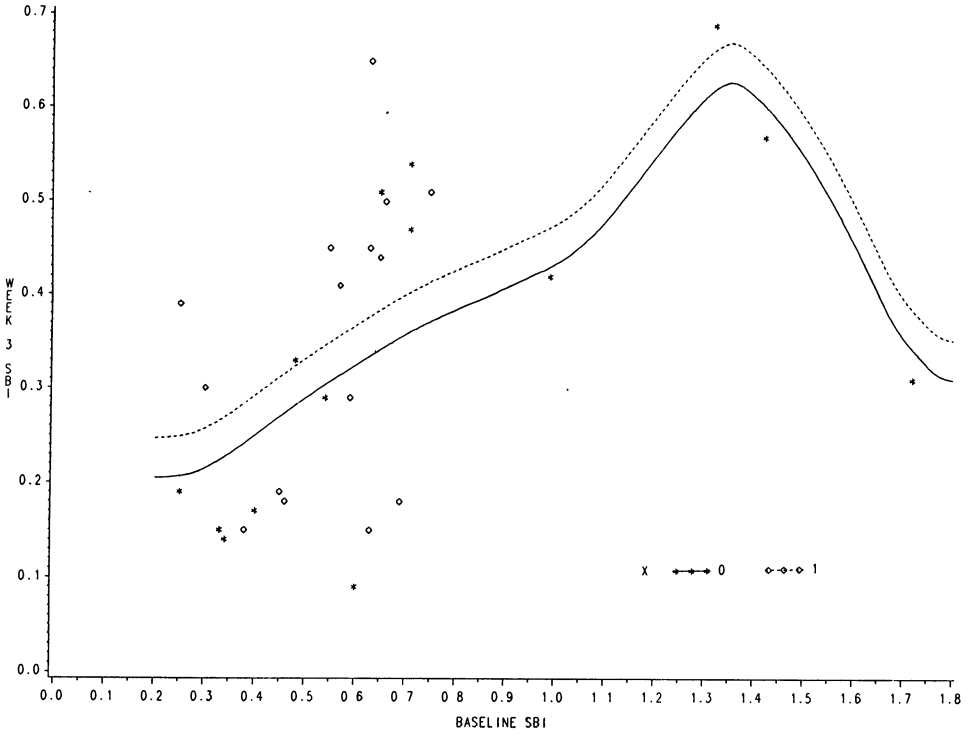


Fig. 2. Raw data and partial kernel estimates for example 2

Example 3. The final example is an analysis of covariance with simulated data and smooth part

$$f(t) = t/(1 + t^{2/3}).$$

A sample of 30 t_i values was drawn from a simulated $N(0.25, 1)$ distribution with observations $y_i = f(t_i) + 0.2N(0, 1)$. A second sample of t_i values was taken from an $N(-0.25, 1)$ distribution with $y_i = f(t_i) + 0.2 + 0.2N(0, 1)$. Fig. 3 shows the simulated data and the theoretical and estimated curves. In this case the estimates were obtained using the window of equation (6.1) and the bandwidth from GCV of 0.9.

The example was chosen to examine the difference in a small sample between the partial regression estimator and the GJS estimator. Table 2 contains the actual bias and standard deviation of $\hat{\beta}(/\sigma)$ for various values of b. The standard deviations for both estimators are comparable in all cases with that of $\hat{\boldsymbol{\beta}}_{GJS}$ slightly better than $\hat{\boldsymbol{\beta}}_p$. In both cases the bias is essentially negligible compared with the actual value of $\beta = 0.2$ with bandwidths near the area of 0.9 selected by GCV, and for very small bandwidths the GJS estimator performed better. However, as predicted by the asymptotics, the bias of $\hat{\boldsymbol{\beta}}_{GJS}$ deteriorates much more rapidly than that of $\hat{\boldsymbol{\beta}}_p$ for large b. In some applications where the variance is relatively large, the effect of bias may be negligible for either method, but here the partial method corrects for potentially important but unknowable bias with little loss in variance.

APPENDIX A: PROOFS OF THEOREMS

Proofs for $\hat{\boldsymbol{\beta}}_p$ and $\hat{f}_p$ are given in detail. The proofs for the GJS estimators are omitted

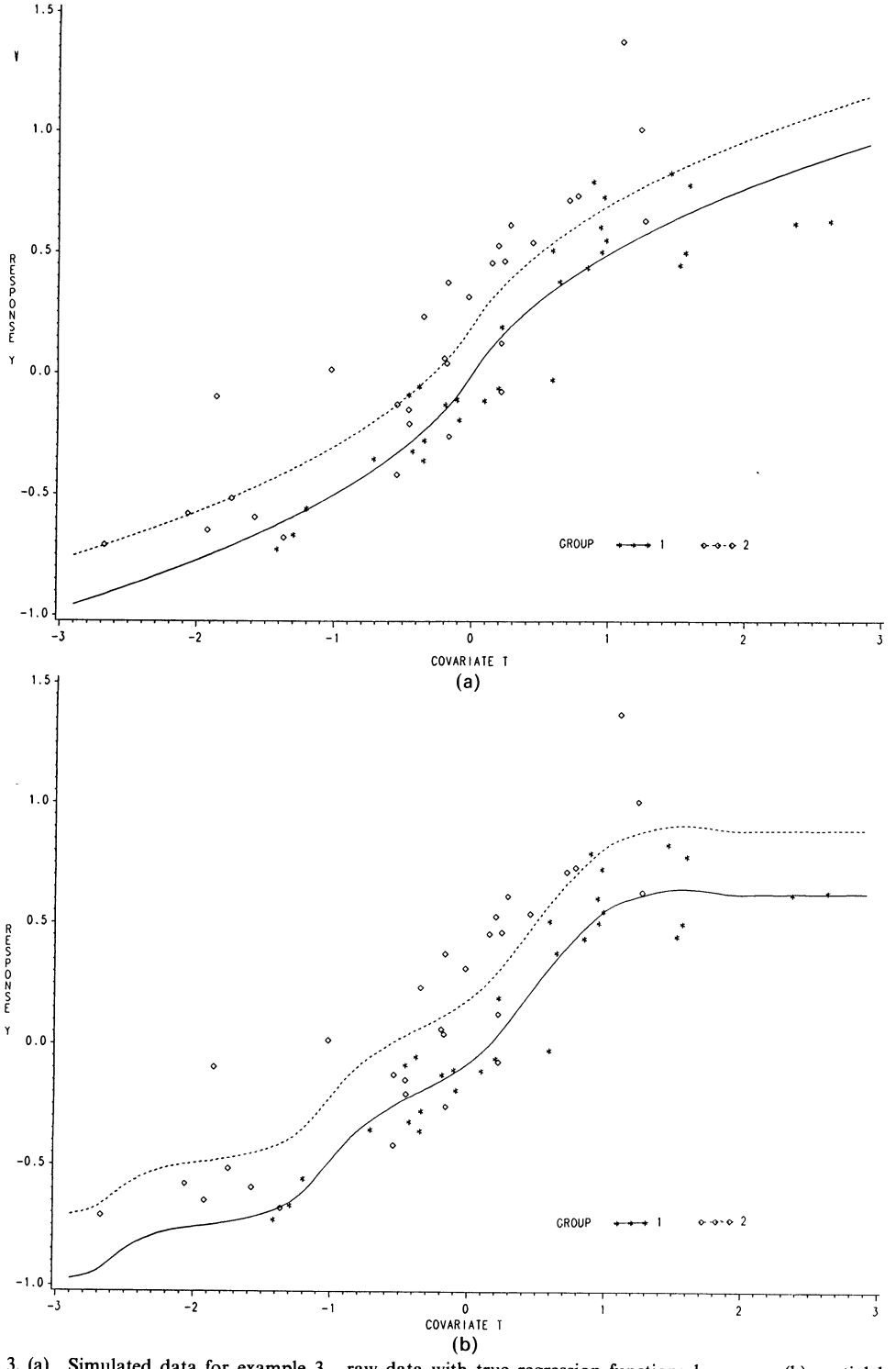


Fig. 3. (a) Simulated data for example 3—raw data with true regression functions by group; (b) partial kernel estimates for example 3—raw data with estimated regression functions by group

TABLE 2
Actual bias and standard deviation for both estimators using the simulated data of example 3

Bandwidth b	*Bias*		*Standard deviation/σ*	
	$\hat{\beta}_p$	$\hat{\beta}_{GJS}$	$\hat{\beta}_p$	$\hat{\beta}_{GJS}$
0.3	0.0048	0.0046	0.0992	0.0892
0.4	0.0076	0.0070	0.0898	0.0822
0.5	0.0050	0.0041	0.0850	0.0799
0.6	0.0024	−0.0001	0.0825	0.0795
0.7	0.0020	−0.0035	0.0810	0.0791
0.8	0.0010	−0.0084	0.0797	0.0783
0.9	−0.0016	−0.0157	0.0787	0.0776
1.0	−0.0033	−0.0227	0.0779	0.0771
1.1	−0.0038	−0.0284	0.0775	0.0765
1.2	−0.0041	−0.0337	0.0773	0.0759
1.3	−0.0054	−0.0397	0.0770	0.0753
1.4	−0.0080	−0.0466	0.0776	0.0746

when they are similar. To begin the calculations, expand $\hat{\boldsymbol{\beta}}_p$ via equations (3.3a) and (4.2) to obtain

$$\begin{aligned}\hat{\boldsymbol{\beta}}_p &= (\tilde{\mathbf{X}}'\tilde{\mathbf{X}})^{-1}\tilde{\mathbf{X}}'(\mathbf{I}-\mathbf{K})\mathbf{y}\\ &= (\tilde{\mathbf{X}}'\tilde{\mathbf{X}})^{-1}\tilde{\mathbf{X}}'[\tilde{\mathbf{X}}\boldsymbol{\beta}+\tilde{\mathbf{f}}+(\mathbf{I}-\mathbf{K})\boldsymbol{\varepsilon}]\\ &= \boldsymbol{\beta}+(\tilde{\mathbf{X}}'\tilde{\mathbf{X}})^{-1}\tilde{\mathbf{X}}'\tilde{\mathbf{f}}+(\tilde{\mathbf{X}}'\tilde{\mathbf{X}})^{-1}\tilde{\mathbf{X}}'(\mathbf{I}-\mathbf{K})\boldsymbol{\varepsilon}.\end{aligned}$$

Thus the bias of $\hat{\boldsymbol{\beta}}_p$ is

$$\text{bias}(\hat{\boldsymbol{\beta}}_p) = \mathrm{E}(\hat{\boldsymbol{\beta}}_p) - \boldsymbol{\beta} = (\tilde{\mathbf{X}}'\tilde{\mathbf{X}})^{-1}\tilde{\mathbf{X}}'\tilde{\mathbf{f}}, \tag{A.1}$$

and the variance–covariance matrix is

$$\text{var}(\hat{\boldsymbol{\beta}}_p) = \sigma^2(\tilde{\mathbf{X}}'\tilde{\mathbf{X}})^{-1}\tilde{\mathbf{X}}'(\mathbf{I}-\mathbf{K})(\mathbf{I}-\mathbf{K})'\tilde{\mathbf{X}}(\tilde{\mathbf{X}}'\tilde{\mathbf{X}})^{-1}. \tag{A.2}$$

Similarly,

$$\hat{\boldsymbol{\beta}}_{GJS} = \boldsymbol{\beta}+(\mathbf{X}'\tilde{\mathbf{X}})^{-1}\mathbf{X}'\tilde{\mathbf{f}}+(\mathbf{X}'\tilde{\mathbf{X}})^{-1}\mathbf{X}'(\mathbf{I}-\mathbf{K})\boldsymbol{\varepsilon},$$

with corresponding expressions for bias and variance.

We begin with a preliminary result.

Lemma 1. If $n\to\infty$, $b\to 0$ and $nb\to\infty$,

(a) $n^{-1}\mathbf{X}'\tilde{\mathbf{X}}\to\mathbf{V}$ and

(b) $n^{-1}\tilde{\mathbf{X}}'\tilde{\mathbf{X}}\to\mathbf{V}$.

Proof. The jth column of $\tilde{\mathbf{X}} = (\mathbf{I}-\mathbf{K})\mathbf{X}$ is

$$\tilde{\mathbf{x}}_j = \tilde{\mathbf{g}}_j + (\mathbf{I}-\mathbf{K})\boldsymbol{\eta}_j.$$

We consider $n^{-1}\tilde{\mathbf{x}}_i'\tilde{\mathbf{x}}_j$ to prove (b); the proof of (a) is similar. Since

$$n^{-1}(\tilde{\mathbf{x}}_i'\tilde{\mathbf{x}}_j) = n^{-1}(\tilde{\mathbf{g}}_i'\tilde{\mathbf{g}}_j+\tilde{\mathbf{g}}_i'\tilde{\boldsymbol{\eta}}_j+\tilde{\mathbf{g}}_j'\tilde{\boldsymbol{\eta}}_i+\boldsymbol{\eta}_i'\boldsymbol{\eta}_j-\boldsymbol{\eta}_i'\mathbf{K}\boldsymbol{\eta}_j-\boldsymbol{\eta}_i'\mathbf{K}'\boldsymbol{\eta}_j+\boldsymbol{\eta}_i'\mathbf{K}'\mathbf{K}\boldsymbol{\eta}_j),$$

it suffices to apply assumption (a) (Section 4) and then to show that all terms except

$n^{-1}\boldsymbol{\eta}_i'\boldsymbol{\eta}_j$ tend to zero. Assumptions (d) and (g) give

$$n^{-1}\tilde{\mathbf{g}}_i'\tilde{\mathbf{g}}_j = b^{2\nu}\int_0^1 g_i^{(\nu)}(t)\, g_j^{(\nu)}(t)\, h_1(t)^2\, p(t)\, dt + o(b^{2\nu}),$$

implying $n^{-1/2}\,\|\,\tilde{\mathbf{g}}_i\,\| = O(b^{\nu}) = o(1)$. Next, assumption (a) implies that $\|\,\boldsymbol{\eta}_j\,\| = O(n^{1/2})$, hence from assumption (c)

$$\begin{aligned}\|\,\tilde{\boldsymbol{\eta}}_j\,\| &= \|\,(\mathbf{I}-\mathbf{K})\boldsymbol{\eta}_j\,\| \leqslant \|\,\boldsymbol{\eta}_j\,\| + \|\,\mathbf{K}\boldsymbol{\eta}_j\,\| \\ &\leqslant O(n^{1/2}) + O(b^{-1/2}) \\ &= O(n^{1/2})\end{aligned}$$

since $nb \to \infty$. These two estimates show that the terms involving $\tilde{\mathbf{g}}_i$, $\tilde{\mathbf{g}}_j$, $\tilde{\boldsymbol{\eta}}_i$ and $\tilde{\boldsymbol{\eta}}_j$ are negligible. Again assumptions (a) and (c) imply $|\,n^{-1}\boldsymbol{\eta}_i'\mathbf{K}\boldsymbol{\eta}_j\,| = O(n^{-1/2}b^{-1/2})$ and $|\,n^{-1}\boldsymbol{\eta}_i'\mathbf{K}'\mathbf{K}\boldsymbol{\eta}_j\,| = O((nb)^{-1})$, and the proof is complete.

Proof of equation (4.5a). In view of lemma 1 and using equation (A.1), it suffices to consider

$$n^{-1}\tilde{\mathbf{x}}_i'\tilde{\mathbf{f}} = n^{-1}[\tilde{\mathbf{g}}_i'\tilde{\mathbf{f}} + \boldsymbol{\eta}_i'\tilde{\mathbf{f}} - (\mathbf{K}\boldsymbol{\eta}_i)'\tilde{\mathbf{f}}].$$

The first term is asymptotic to

$$b^{2\nu}\int_0^1 g_i^{(\nu)}(t) f^{(\nu)}(t)\, h_1^2(t)\, p(t)\, dt$$

by equation (2.2a) and assumptions (d) and (g). The second is $O(b^{\nu}n^{-1/2})$ by assumption (f) and the last term is dominated by $n^{-1}\,\|\,\mathbf{K}\boldsymbol{\eta}_i\,\|\,\|\,\tilde{\mathbf{f}}\,\| = n^{-1}O(b^{-1/2}b^{\nu}n^{1/2})$ by assumptions (c) and (e).

Proof of equation (4.4a). The proof is similar except that

$$\begin{aligned}n^{-1}\mathbf{x}_i'\tilde{\mathbf{f}} &= n^{-1}(\mathbf{g}_i'\tilde{\mathbf{f}} + \boldsymbol{\eta}_i'\tilde{\mathbf{f}}) \\ &= b^{\nu}\int_0^1 g_i(t) f^{(\nu)}(t)\, h_1(t)\, p(t)\, dt + O(b^{\nu}n^{-1/2}).\end{aligned}$$

The next proofs are simplified by the following notation. From assumptions (c) and (d), $n^{-1}\,\|\,\mathbf{g}_j - \mathbf{K}\mathbf{x}_j\,\|^2 = [O(b^{\nu}) + O((nb)^{-1/2})]^2 = O(b^{2\nu}) + O((nb)^{-1})$ uniformly $(1 \leqslant j \leqslant p,\ n \geqslant 1)$. (Note the connection with equation (2.3).) Thus there exists a bounding function

$$e(n) = M_1 b^{2\nu} + M_2(nb)^{-1} \tag{A.3}$$

with finite constants M_1 and M_2 such that $n^{-1}\,\|\,\mathbf{g}_j - \mathbf{K}\mathbf{x}_j\,\|^2 \leqslant e(n)$ uniformly in j and n. Note that $ne(n) \to \infty$ under the assumptions of theorems 1 and 2.

Proof of equation (4.5b). From equation (A.2),

$$\operatorname{var}(\hat{\boldsymbol{\beta}}_p) = \sigma^2(\tilde{\mathbf{X}}'\tilde{\mathbf{X}})^{-1} + \sigma^2(\tilde{\mathbf{X}}'\tilde{\mathbf{X}})^{-1}[-\tilde{\mathbf{X}}'\mathbf{K}\tilde{\mathbf{X}} - \tilde{\mathbf{X}}'\mathbf{K}'\tilde{\mathbf{X}} + \tilde{\mathbf{X}}'\mathbf{K}\mathbf{K}'\tilde{\mathbf{X}}](\tilde{\mathbf{X}}'\tilde{\mathbf{X}})^{-1}.$$

Using lemma 1, it will suffice to show that the terms in brackets are all componentwise $o(n)$. From assumption (b)

$$\|\,\mathbf{K}'\,\| \leqslant [\operatorname{tr}(\mathbf{K}'\mathbf{K})]^{1/2} = O(b^{-1/2}). \tag{A.4}$$

It then follows by assumption (c) and equation (A.3) that

$$\|\,\mathbf{K}'\tilde{\mathbf{x}}_i\,\| \leqslant \|\,\mathbf{K}'\boldsymbol{\eta}_i\,\| + \|\,\mathbf{K}'(\mathbf{g}_i - \mathbf{K}\mathbf{x}_i)\,\|$$

$$\leqslant \| \mathbf{K}'\boldsymbol{\eta}_i \| + \| \mathbf{K}' \| \, \| \mathbf{g}_i - \mathbf{K}\mathbf{x}_i \|$$
$$= O(b^{-1/2}) + O\{b^{-1/2}[ne(n)]^{1/2}\}$$
$$= O(ne(n))$$

as $ne(n) \to \infty$. By lemma 1, $\| \tilde{\mathbf{x}}_j \| = O(n^{1/2})$, hence

$$| \tilde{\mathbf{x}}_i' \mathbf{K} \tilde{\mathbf{x}}_j | \leqslant \| \mathbf{K}'\tilde{\mathbf{x}}_i \| \, \| \tilde{\mathbf{x}}_j \| = O(n^{3/2}e(n)),$$

and

$$| \tilde{\mathbf{x}}_i' \mathbf{K}\mathbf{K}' \tilde{\mathbf{x}}_j | \leqslant \| \mathbf{K}'\tilde{\mathbf{x}}_i \| \, \| \mathbf{K}'\tilde{\mathbf{x}}_j \| = O(n^2 e(n)^2).$$

Gathering estimates and invoking lemma 1 once more, we have

$$\operatorname{var}(\hat{\boldsymbol{\beta}}_p) - \sigma^2(\tilde{\mathbf{X}}'\tilde{\mathbf{X}})^{-1} = (n\mathbf{V})^{-1}[O(n^{3/2}e(n)) + O(n^2 e(n)^2)](n\mathbf{V})^{-1}$$
$$= n^{-1}[O(n^{1/2}e(n)) + O(ne(n)^2)].$$

The assumptions of theorem 1 applied to equation (A.3) imply that $e(n) \to 0$ and $ne(n)^2 \to 0$, so the last term is $o(n^{-1})$.

Proof of equations (4.4b) and (4.4c). Letting $\mathbf{B} = (\mathbf{I} - \mathbf{K})(\mathbf{I} - \mathbf{K})'$, we can use equation (4.3), lemma 1 and assumption (a) to write

$$\operatorname{var}(\hat{\boldsymbol{\beta}}_{\mathrm{GJS}}) = \sigma^2(\mathbf{X}'\tilde{\mathbf{X}})^{-1}\mathbf{X}'\mathbf{B}\mathbf{X}(\tilde{\mathbf{X}}'\mathbf{X})^{-1}$$
$$= n^{-2}\sigma^2\mathbf{V}^{-1}\{n\mathbf{V}[1 + o(1)] + \boldsymbol{\eta}'\mathbf{B}\mathbf{g} + \mathbf{g}'\mathbf{B}\boldsymbol{\eta} + \mathbf{g}'\mathbf{B}\mathbf{g}\}\mathbf{V}^{-1}[1 + o(1)].$$

Thus to prove equation (4.4b), it suffices to show that $\mathbf{g}'\mathbf{B}\boldsymbol{\eta} = o(n)$. But equation (A.4) and assumptions (a), (c) and (d) imply that

$$\| \mathbf{g}'\mathbf{B}\boldsymbol{\eta} \| \leqslant \| \mathbf{g}'(\mathbf{I} - \mathbf{K})'\boldsymbol{\eta} \| + \| \mathbf{g}'\mathbf{K}(\mathbf{I} - \mathbf{K})'\boldsymbol{\eta} \|$$
$$\leqslant \| \tilde{\mathbf{g}} \| \, \| \boldsymbol{\eta} \| + \| \mathbf{g} \| (\| \mathbf{K}\boldsymbol{\eta} \| + \| \mathbf{K} \| \, \| \mathbf{K}'\boldsymbol{\eta} \|)$$
$$= O(n^{1/2}b^{\nu})O(n^{1/2}) + O(n^{1/2})[O(b^{-1/2}) + O(b^{-1})] = o(n)$$

as $b \to 0$ and $nb^2 \to \infty$. The symmetric case (4.3c) follows immediately using assumption (d).

Proof of theorem 3. From equation (3.4) for both estimators,

$$\operatorname{bias}(\hat{f}(t)) = \operatorname{bias}(\hat{f}_0(t)) + \mathbf{k}(t)'\mathbf{X} \operatorname{bias}(\hat{\boldsymbol{\beta}}) \tag{A.5}$$

and

$$\operatorname{var}(\hat{f}(t)) = \operatorname{var}(\hat{f}_0(t)) + \mathbf{k}(t)'\mathbf{X} \operatorname{var}(\hat{\boldsymbol{\beta}})\mathbf{X}'\mathbf{k}(t) - 2\mathbf{k}(t)'\mathbf{X} \operatorname{cov}(\hat{\boldsymbol{\beta}}, \hat{f}_0(t)), \tag{A.6}$$

where

$$\mathbf{k}(t)' = (K(t, t_1), \ldots, K(t, t_n)).$$

Assumptions (c) and (d) imply that $\mathbf{k}(t)'\mathbf{X} = (\hat{g}_1(t), \ldots, \hat{g}_p(t)) \to (g_1(t), \ldots, g_p(t))$ as $n \to \infty$, so equation (4.7) follows from equations (4.4a) and (4.4b) applied to equation (A.5). For both estimators $\operatorname{var}(\hat{\boldsymbol{\beta}}) = O(n^{-1}) = o(\operatorname{var}(\hat{f}_0(t))$ from theorems 1 and 2, the remark following theorem 2 and equation (2.2b), so equation (4.8) follows from equation (A.6).

In the following we use the matrix L_p norm

$$\| \mathbf{A} \|_p = \max_{\| \mathbf{v} \|_p > 0} \| \mathbf{A}\mathbf{v} \|_p / \| \mathbf{v} \|_p,$$

where

$$\| \mathbf{v} \|_p = \left(\sum_{i=1}^{n} | v_i |^p \right)^{1/p}, \qquad 1 \leqslant p < \infty,$$

and

$$\| \mathbf{v} \|_\infty = \max_{1 \leqslant i \leqslant n} | v_i |.$$

It is well known that $\| \mathbf{A} \|_2$ is the modulus of the largest singular value of $\mathbf{A}$, that

$$\| \mathbf{A} \|_1 = \max_{1 \leqslant j \leqslant n} \sum_{i=1}^{n} | A_{ij} |$$

and that

$$\| \mathbf{A} \|_\infty = \max_{1 \leqslant i \leqslant n} \sum_{j=1}^{n} | A_{ij} |$$

(see Stewart (1973)).

Proof of theorem 4. The proof for $\hat{\boldsymbol{\beta}}_p$ is given. To establish asymptotic normality, first write $\hat{\boldsymbol{\beta}}_p = (\tilde{\mathbf{X}}_n' \tilde{\mathbf{X}}_n)^{-1} \tilde{\mathbf{X}}_n' (\mathbf{I}_n - \mathbf{K}_n) \mathbf{y}_n$, where we now explicitly display the dependence on n. In view of lemma 1, it suffices to show that $n^{-1/2} \mathbf{a}' \tilde{\mathbf{X}}_n' (\mathbf{I}_n - \mathbf{K}_n) \boldsymbol{\varepsilon}_n = n^{-1/2} \mathbf{c}_n' \boldsymbol{\varepsilon}_n \to N(0, \sigma^2 \mathbf{a}' \mathbf{V} \mathbf{a})$, where $\mathbf{c}_n' = \mathbf{a}' \tilde{\mathbf{X}}_n' (\mathbf{I}_n - \mathbf{K}_n) = (c_{1n}, \ldots, c_{nn})$ and $\mathbf{a}$ is an arbitrary fixed vector. This will follow from the Lindeberg condition (cf. Billingsley (1986), problem 27.6) by showing that

$$\max_{1 \leqslant i \leqslant n} c_{in}^2 \Big/ \sum_{i=1}^{n} c_{in}^2 \to 0. \tag{A.7}$$

From the proof of equation (4.5b),

$$n^{-1} \mathbf{c}_n' \mathbf{c}_n = n^{-1} \mathbf{a}' \tilde{\mathbf{X}}_n' (\mathbf{I}_n - \mathbf{K}_n)(\mathbf{I}_n - \mathbf{K}_n)' \tilde{\mathbf{X}}_n \mathbf{a} \to \mathbf{a}' \mathbf{V} \mathbf{a},$$

so equation (A.7) will hold if

$$\max_{1 \leqslant i \leqslant n} c_{in}^2 = o(n). \tag{A.8}$$

(Under model (4.1), convergence in equation (A.7) is in probability. The Lindeberg condition then applies conditioned on $\mathbf{X}_n$ with probability arbitrarily close to unity.)

Assumptions (i) imply that $\| \mathbf{K}_n \|_\infty$ and $\| \mathbf{K}_n' \|_\infty$ are uniformly bounded. Thus

$$\begin{aligned} \max_i | c_{in} | &= \| \mathbf{c}_n \|_\infty \\ &= \| (\mathbf{I}_n - \mathbf{K}_n')(\mathbf{I}_n - \mathbf{K}_n) \mathbf{X}_n \mathbf{a} \|_\infty \\ &\leqslant (1 + \| \mathbf{K}_n' \|_\infty)(1 + \| \mathbf{K}_n \|_\infty) \| \mathbf{X}_n \mathbf{a} \|_\infty. \end{aligned}$$

If the components of $\mathbf{X}_n$ are uniformly bounded, it follows immediately that $\| \mathbf{c}_n \|_\infty = O(1)$; hence equation (A.8) holds. For model (4.1) with $x_{ijn} = g_j(t_{in}) + \eta_{ijn}$ and g_j continuous, where now $\mathbf{X}_n = \{x_{ijn}\}_{i=1,n;\, j=1,p}$ and $\boldsymbol{\eta}_{jn}' = (\eta_{1jn}, \ldots, \eta_{njn})$, it suffices

to show that

$$\| \boldsymbol{\eta}_{jn} \|_\infty = o_p(n^{1/2}) \qquad (1 \leqslant j \leqslant p). \tag{A.9}$$

By assumption $E | \eta_{ijn} |^{2+\delta} \leqslant C < \infty$ for all i, j and n, so the Markov inequality yields

$$P(\max_{1 \leqslant i \leqslant n} | \eta_{ijn} | \geqslant m_n) \leqslant \sum_i P(| \eta_{ijn} | \geqslant m_n) \leqslant nCm_n^{-2-\delta}$$

for any constant m_n. Hence with $m_n = n^{1/2}/\log n$, $P(\| \boldsymbol{\eta}_{jn} \|_\infty > m_n) \to 0$ and equation (A.9) holds.

Remark. If $n \operatorname{var}(\hat{\boldsymbol{\beta}}_{GJS})$ converges in equation (4.4b), the same argument gives asymptotic normality for $\hat{\boldsymbol{\beta}}_{GJS}$.

The proofs of theorems 5 and 6 require a preliminary result.

Lemma 2. As $n \to \infty$, $b \to \infty$ and $nb \to \infty$,

$$\operatorname{tr}(\mathbf{A}(b)) = \operatorname{tr}(\mathbf{K}(b))[1 + o(1)]$$

and

$$\operatorname{tr}(\mathbf{A}(b)'\mathbf{A}(b)) = \operatorname{tr}(\mathbf{K}(b)'\mathbf{K}(b))[1 + o(1)],$$

where $\mathbf{A}(b)$ denotes either the GJS or the partial hat matrix given in equation (5.2).

Proof. For the partial regression estimator,

$$\operatorname{tr}(\mathbf{A}_p(b)) = \operatorname{tr}(\mathbf{K}(b)) + \operatorname{tr}(\mathbf{P}_{\tilde{\mathbf{X}}}[\mathbf{I} - \mathbf{K}(b)])$$

by equation (5.2b). We use the fact that for any matrix $\mathbf{B}$ of rank p,

$$| \operatorname{tr}(\mathbf{B}) | \leqslant p \| \mathbf{B} \|_2.$$

Because $\mathbf{P}_{\tilde{\mathbf{X}}}$ and $\mathbf{I}$ are projections, their norms are both unity. The product $\mathbf{P}_{\tilde{\mathbf{X}}}[\mathbf{I} - \mathbf{K}(b)]$ has rank p, hence

$$| \operatorname{tr}(\mathbf{P}_{\tilde{\mathbf{X}}}[\mathbf{I} - \mathbf{K}(b)]) | \leqslant p(1 + \| \mathbf{K}(b) \|_2).$$

But by an inequality of Reisz (see Hardy *et al.* (1952)),

$$\| \mathbf{K}(b) \|_2^2 \leqslant \| \mathbf{K}(b) \|_1 \| \mathbf{K}(b) \|_\infty = O(1)$$

under assumption (i). The first assertion of the lemma then follows for $\mathbf{A}_p(b)$ since $\operatorname{tr}(\mathbf{K}(b)) \to \infty$ as $b \to 0$, $n \to \infty$ and $nb \to \infty$. The second statement in the lemma is proved similarly for $\mathbf{A}_p(b)$ since $\operatorname{tr}(\mathbf{K}(b)'\mathbf{K}(b)) \to \infty$ by assumption (b).

The proof for $\mathbf{A}_{GJS}(b)$ is similar. It is only necessary to show that $\| \mathbf{Q}_{\mathbf{X}} \| = O(1)$, where $\mathbf{Q}_{\mathbf{X}}$ (given before equation (5.2)) replaces $\mathbf{P}_{\tilde{\mathbf{X}}}$. But

$$\begin{aligned} \| \mathbf{Q}_{\mathbf{X}} \|_2^4 &= \| \mathbf{Q}_{\mathbf{X}}' \mathbf{Q}_{\mathbf{X}} \|_2^2 \\ &\leqslant \operatorname{tr}(\mathbf{Q}_{\mathbf{X}}' \mathbf{Q}_{\mathbf{X}}) \\ &= \operatorname{tr}(\mathbf{X}(\tilde{\mathbf{X}}'\mathbf{X})^{-1}\tilde{\mathbf{X}}'\tilde{\mathbf{X}}(\mathbf{X}'\tilde{\mathbf{X}})^{-1}\mathbf{X}') \\ &\approx \operatorname{tr}(\mathbf{X}(\mathbf{X}'\tilde{\mathbf{X}})^{-1}\mathbf{X}') \\ &= \operatorname{tr}(\mathbf{X}'\mathbf{X}(\mathbf{X}'\tilde{\mathbf{X}})^{-1}) \end{aligned}$$

by lemma 1. But assumption (a) and the fact that $\mathbf{g}(t)$ is bounded on [0, 1] imply that $n^{-1}\mathbf{X}'\mathbf{X} = O(1)$; hence another application of lemma 1 gives the required bound.

Proof of theorem 5. For either GJS or partial smoothing, we have

$$\begin{aligned} E(\mathbf{y}) - \hat{\mathbf{y}} &= \mathbf{f} + \mathbf{X}\boldsymbol{\beta} - \mathbf{X}\hat{\boldsymbol{\beta}} - \mathbf{K}(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}) \\ &= \tilde{\mathbf{f}} + \tilde{\mathbf{X}}(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}) - \mathbf{K}\boldsymbol{\varepsilon} \\ &= \tilde{\mathbf{f}} - \tilde{\mathbf{X}}\,\mathrm{bias}(\hat{\boldsymbol{\beta}}) - \mathbf{A}\boldsymbol{\varepsilon} \end{aligned}$$

where $\mathbf{A}$ is given by equation (5.2). Thus

$$\begin{aligned} E(R(b)) &= n^{-1} \| \tilde{\mathbf{f}} - \tilde{\mathbf{X}}\,\mathrm{bias}(\boldsymbol{\beta}) \|^2 + n^{-1}\sigma^2 \,\mathrm{tr}(\mathbf{A}'\mathbf{A}) \\ &= n^{-1}[\| \tilde{\mathbf{f}} \|^2 - 2\tilde{\mathbf{f}}'\tilde{\mathbf{X}}\,\mathrm{bias}(\hat{\boldsymbol{\beta}}) + (\mathrm{bias}(\hat{\boldsymbol{\beta}}))'\tilde{\mathbf{X}}'\tilde{\mathbf{X}}\,\mathrm{bias}(\hat{\boldsymbol{\beta}})] \\ &\quad + n^{-1}\sigma^2 \mathrm{tr}(\mathbf{A}'\mathbf{A}). \end{aligned}$$

Because

$$\begin{aligned} E(R_0(b)) &= n^{-1} E \| \mathbf{f} - \mathbf{K}\mathbf{y} \|^2 \\ &= n^{-1} \| \tilde{\mathbf{f}} \|^2 + \sigma^2 n^{-1} \mathrm{tr}(\mathbf{K}'\mathbf{K}), \end{aligned}$$

lemma 1, assumption (e) and either equation (4.4a) or equation (4.5a) along with lemma 2 establish the theorem.

Proof of theorem 6. If $\mu_1(b) = n^{-1}\mathrm{tr}(\mathbf{A}(b))$ and $\mu_2(b) = n^{-1}\mathrm{tr}(\mathbf{A}(b)'\mathbf{A}(b))$, the GCV theorem as stated in Golub *et al.* (1979) gives equation (5.4), where $o(1)$ denotes a term tending to zero as $n \to \infty$, provided that

$$\begin{aligned} \mu_1(b) &\to 0 \\ \mu_1^2(b)/\mu_2(b) &\to 0. \end{aligned} \tag{A.10}$$

When $\boldsymbol{\beta} = \mathbf{0}$ in model (1.1) and $\mathbf{A} = \mathbf{K}$, equation (A.10) holds by assumptions (b) and (h). This result extends to GJS and partial kernel smoothing by applying lemma 2.

Justification of assumptions (h) and (i). The idea justifying assumption (h) is illustrated under equation (2.4) where

$$K(t_i, t_i) = w(0)/\sum_j w((t_j - t_i)/b).$$

Under assumption (g), we can approximate the denominator by an integral to obtain $K(t_i, t_i) = O((nb)^{-1})$, so assumption (h) follows. Assumption (i) is plausible because the matrix $\mathbf{K}$ should behave in some sense like a projection. The proof will vary from case to case but the idea can again be illustrated under equation (2.4). If w is non-negative, the row sums of K are identically unity, so the first bound is trivial. In the general case, assumption (g) implies that the row sums can be approximated by integrals and the dependence on b can be seen to vanish as $n \to \infty$. A similar argument works for the second assertion of assumption (i).

ACKNOWLEDGEMENTS

Many of the ideas in this paper originated in discussions with Professor Randy Eubank. He pointed out an error in an earlier version, and his advice and comments on successive drafts are gratefully acknowledged. The author is also indebted to the referees for their suggestions and to Dr Peter Green, Professor Nancy Heckman and Professor John Rice for providing copies of their manuscripts before publication. Rice's paper in particular provided motivation and insight for the present work.

REFERENCES

Billingsley, P. (1986) *Probability and Measure*, 2nd edn, problem 27.6. New York: Wiley.

Breiman, L. and Friedman, J. (1985) Estimating optimal transformations for multiple regression and correlation. *J. Amer. Statist. Ass.*, **80**, 580–598.

Chen, H. (1988) Convergence rates for parametric components in a partly linear model. *Ann. Statist.*, **16**, 136–146.

Collomb, G. (1981) Revue de la regression nonparamétrique. *Inst. Statist. Rev.*, **49**, 75–93.

Craven, P. and Wahba, G. (1979) Smoothing noisy data with spline functions. *Numer. Math.*, **31**, 377–403.

Cuzick, J. (1987) Semiparametric additive regression. Unpublished.

Denby, L. (1984) Smooth regression functions. *PhD Thesis.* Department of Statistics, University of Michigan, Ann Arbor.

———(1986) Smooth regression functions. *Statistical Research Report 26.* Murray Hill: AT&T Bell Laboratories.

Draper, N. and Guttman, I. (1980) Incorporating overlap effects from neighboring units into response surface models. *Appl. Statist.*, **29**, 128–134.

Engle, R., Granger, C., Rice, J. and Weiss, A. (1986) Nonparametric estimates of the relation between weather and electricity sales. *J. Amer. Statist. Ass.*, **81**, 310–320.

Eubank, R. L. (1986) A note on smoothness priors and nonlinear regression. *J. Amer. Statist. Ass.*, **81**, 514–517.

Eubank, R. L. and Speckman, P. (1988) Curve fitting by polynomial-trigonometric regression. Unpublished.

Gasser, Th. and Müller, H. G. (1979) Kernel estimation of regression functions. In *Smoothing Techniques for Curve Estimation* (eds Th. Gasser and M. Rosenblatt), pp. 23–68. Berlin: Springer.

Gasser, T., Müller, H.-G. and Mammitzsch, V. (1985) Kernels for nonparametric curve estimation. *J. R. Statist. Soc.* B, **47**, 238–252.

Golub, G., Heath, M. and Wahba, G. (1979) Generalized cross validation as a method for choosing a good ridge parameter. *Technometrics*, **21**, 215–223.

Green, P. (1985) Linear models for field trials, smoothing and cross-validation. *Biometrika*, **72**, 527–537.

Green, P., Jennison, C. and Seheult, A. (1985) Analysis of field experiments by least squares smoothing. *J. R. Statist. Soc.* B, **47**, 299–315.

Härdle, W. and Marron, J. S. (1985) Optimal bandwidth selection in nonparametric regression function estimation. *Ann. Statist.*, **12**, 1465–1481.

Hardy, G. H., Littlewood, J. E. and Polya, G. (1952) *Inequalities*, theorem 296. Cambridge: Cambridge University Press.

Hastie, T. and Tibshirani, R. (1986) Generalized additive models. *Statist. Sci.*, **1**, 297–310.

Heckman, N. (1986a) Minimax estimates in a semiparametric model. Unpublished.

———(1986b) Spline smoothing in partly linear models. *J. R. Statist. Soc.* B, **48**, 244–248.

Nadaraya, E. A. (1964) On estimating regression. *Theor. Probabil. Appl.*, **9**, 141–142.

Nussbaum, M. (1985) Spline smoothing in regression models and asymptotic efficiency in L^2. *Ann. Statist.*, **13**, 984–997.

O'Sullivan, F. (1986) Ill-posed inverse problems. *Statist. Sci.*, **1**, 502–527.

Platt, R. (1985) Personal communication.

Priestley, M. B. and Chao, M. T. (1972) Nonparametric function fitting. *J. R. Statist. Soc.* B, **34**, 385–392.

Rice, J. (1984a) Bandwidth choice for nonparametric regression. *Ann. Statist.*, **12**, 1215–1230.

———(1984b) Boundary modification for kernel regression. *Commun. Statist. Theor. Meth.*, **13**, 893–900.

———(1986) Convergence rates for partially splined models. *Statist. Probabil. Lett.*, **4**, 203–208.

Rice, J. and Rosenblatt, M. (1983) Smoothing splines: regression, differentiation, and deconvolution. *Ann. Statist.*, **11**, 141–156.

Sacks, J. and Ylvisaker, D. (1978) Linear estimation for approximately linear models. *Ann. Statist.*, **6**, 1122–1137.

Shiau, J., Wahba, G. and Johnson, D. R. (1986) Partial spline models for the inclusion of tropopause and frontal boundary information in otherwise smooth two and three dimensional objective analysis. *J. Atmos. Ocean. Technol.*, **3**, 714–725.

Shiller, R. J. (1984) Smoothness priors and nonlinear regression. *J. Amer. Statist. Ass.*, **79**, 609–615.

Silverman, B. W. (1985) Some aspects of the spline smoothing approach to non-parametric regression curve fitting. *J. R. Statist. Soc.* B, **47**, 1–52.

Speckman, P. (1985) Spline smoothing and optimal rates of convergence in nonparametric regression models. *Ann. Statist.*, **13**, 970–983.

Stewart, G. (1973) *Introduction to Matrix Computations.* New York: Academic Press.

Stone, C. (1977) Consistent nonparametric regression. *Ann. Statist.*, **5**, 595–645.

———(1982) Optimal global rates of convergence for nonparametric regression. *Ann. Statist.*, **10**, 1040–1053.

Stuetzle, W. and Mittal, Y. (1979) Some comments on the asymptotic behavior of robust smoothers. In *Smoothing Techniques for Curve Estimation* (eds Th. Gasser and M. Rosenblatt), pp. 191–195. Berlin: Springer.

Tapia, R. and Thompson, J. (1978) *Nonparametric Probability Density Estimation.* Baltimore. Johns Hopkins.

Wahba, G. (1984a) Cross validated spline methods for the estimation of multivariate functions from data on functionals. In *Statistics: an Appraisal, Proc. 50th Anniversary Conf.* (eds H. A. David and H. T. David). Ames: Iowa State University Press.

———(1984b) Partial spline models for the semiparametric estimation of functions of several variables. In *Analyses for Time Series, Japan–US Joint Sem.*, pp. 319–329. Tokyo: Institute of Statistical Mathematics.

Watson, G. S. (1964) Smooth regression analysis. *Sankhya* A, **26**, 359–372.