The Annals of Statistics
1997, Vol. 25, No. 2, 756–770

CURVE ESTIMATION WHEN THE DESIGN DENSITY IS LOW

BY PETER HALL, J. S. MARRON, M. H. NEUMANN AND D. M. TITTERINGTON

Australian National University

In problems where a high-dimensional design is projected into a lower number of dimensions, the density of the new design is typically not bounded away from zero over its support, even if the original one was. Contexts where this problem arises include projection pursuit regression, estimation in single index models and application of the projection-slice method of Radon transform inversion. Theoretical work in these settings typically involves ignoring data toward the ends of the support of the projected design, but in practice that waste of information is not an attractive option. Motivated by these difficulties, we analyze the way in which local linear smoothing is affected by unboundedly sparse design and apply the conclusions of that study to develop empirical, adaptive bandwidth choice methods. Our results even add to knowledge in the familiar case of a design density that is bounded away from zero, where they provide adaptive bandwidth selectors that are optimal right to the ends of the design interval.

1. Introduction. This paper is motivated by a problem which arises in dimension reduction, when nonparametric curve estimation is applied to data obtained by projection. The problem occurs in projection pursuit regression, estimation in single index models, and the projection-slice method of Radon transform inversion, to name only three contexts. Briefly, if design points have a probability density which is bounded within a given multivariate region, then the density of the projection of those points onto a lower-dimensional Euclidean space usually decreases gradually to zero at the extremities of its support. This is true even if the original high-dimensional density was bounded away from zero on its support. Therefore, when using nonparametric methods to recover a target function from its projection, one is forced to either accommodate design densities that are arbitrarily low, or waste some of the information in the multivariate data set by staying away from the ends of the support of the projected design. Quite apart from the inefficiency of the latter solution, it is awkward because it means ignoring different parts of the data for different projections.

In this paper we present an account of nonparametric regression in settings where the design density decreases to zero and apply it to the development of empirical bandwidth choice methods which allow full use of the data. This problem is distinctly different from more familiar ones of adaptive bandwidth choice [e.g., Gasser, Kneip and Köhler (1991), Fan and Gijbels (1995)], where

Received September 1995; revised April 1996
AMS 1991 *subject classifications*. Primary 62G07; secondary 62G20.
Key words and phrases. Bandwidth choice, design density, kernel methods, local adaptivity, local linear smoothing, projection pursuit, single index model.

the theory on which the methodology is based assumes that the design density is bounded away from zero. If that condition is violated, then both the order of magnitude of the appropriate bandwidth and the rate of convergence of the function estimator alter. We propose bandwidth selectors that adjust to variability in both the target function and the design density, even when the latter is very low, and which are nearly optimal in a mean-squared error sense. By way of contrast, and even in the much simpler setting of a design density that is bounded away from zero, some commonly used bandwidth selectors address only variation of the density.

Our analysis demands new theoretical techniques for approximating the variance of a nonparametric regression estimator when its numerator and denominator are both close to zero. The variance does converge to zero at the extremities of the design support, provided the bandwidth is chosen appropriately, but its rate of decrease depends intimately on the unknown rate at which the design density decreases. Our approach to adaptive bandwidth choice involves implicitly estimating the density's rate of decrease, as is made precise in Theorem 2.1. We focus on local linear regression methods, because they are fast becoming the most popular kernel-type approach. However, analogues of our techniques may be developed for Nadaraya–Watson and Gasser–Müller kernel weights, among others. They are discussed in a longer version of this paper [Hall, Marron, Neumann and Titterington (1995)]. There it is shown that neither of these alternative methods performs as well as local linear smoothing. The former method generally suffers a larger order of bias, and the latter from a larger order of variance, when the design density converges to zero.

To set our work in context, suppose independent observations $\mathcal{X} = \{(Y_i, Z_i), 1 \le i \le n\}$ are made of a vector (Y, Z), where Y is a scalar and Z a p-vector. It is desired to estimate $g_\theta(x) = E(Y|\theta \cdot Z = x)$, where θ is a unit p-vector and x is a scalar. (Versions of our results may also be developed in the case where the Z_i's are projected into q dimensions, where $1 \le q < p$.) In the case of single index models [e.g., Brillinger (1983); Härdle, Hall and Ichimura (1993)], (Y, Z) might be generated as $Y = g(\theta_0 \cdot Z) + \varepsilon$, where g is an unknown univariate function, θ_0 is an unknown unit vector, and the error ε is independent of Z, with zero mean. Here the main parameter of interest is generally θ_0, and the univariate function g_θ is estimated for various θ's as a prelude to estimating θ_0. In another setting, if θ is chosen to optimize a measure of "interestingness," often based on entropy or orthogonal polynomials, then g_θ represents the first step in developing a projection pursuit approximation to $E(Y|Z)$ [e.g., Friedman and Stuetzle (1981)]. To understand the effect that projection has on the density of the $\theta \cdot Z_i$'s, suppose the Z_i's are distributed over a region $\mathcal{Q} \subseteq \mathbb{R}^p$, with a p-variate density that is bounded away from both zero and infinity there. Let ψ_θ denote the univariate density of $\theta \cdot Z$, with support $\mathcal{I}_\theta$. If $\mathcal{Q}$ is a rectangular prism and if the unit vector θ is not parallel to one or other of its faces, then ψ_θ decreases to zero at rate x^{p-1} (as $x \downarrow 0$) at either end of $\mathcal{I}_\theta$. The rate is $x^{(p-1)/2}$ if $\mathcal{Q}$ is an ellipsoid.

These sparse data problems also arise in errors-in-variables regression [e.g., Raj and Ullah (1981), Nicholls and Pagan (1985)]. There, data are available on a pair (U, V), related by the identity $V = AU + B$, and the variables (A, B) have a joint distribution whose form is the subject of investigation. The joint characteristic function $\phi_{A,B}$ is expressible via the conditional characteristic function $\phi_{R|\Theta}$ of $R = V/(1+U^2)^{1/2}$ given $\Theta = \arctan U$, through the equation $\phi_{A,B}(r\cos\theta, r\sin\theta) = \phi_{R|\Theta}(r|\theta)$ (an example of the so-called projection-slice theorem). An attractive method for estimating the joint density $f_{A,B}$ of (A, B) is as follows. First estimate $\phi_{R|\Theta}$ using local linear smoothing, encountering exactly the same sparse design problems as in the earlier discussion. Then invert first a Fourier transform and then a Radon transform to obtain an estimate of $f_{A,B}$. (The conditional density of R given Θ is the Radon transform of $f_{A,B}$.)

Curve fitting by local polynomials, of which the local linear method studied in this paper is an example, is well known for its excellent computational and theoretical features, discussed by (for example) Hastie and Loader (1993). It has a long and distinguished history, going back 125 years, which is recounted in an excellent survey paper by Cleveland and Loader (1996). It lies at the heart of widely used software such as LOESS; see Cleveland (1979, 1993), Cleveland and Devlin (1988), Cleveland and Grosse (1991) and Fan (1992). This enduring numerical attraction and the minimax optimality of local linear smoothing [Fan (1993)], have earned that method the accolade of the "golden standard" for nonparametric regression [Seifert and Gasser (1996)].

Section 2 will present our main theoretical results on performance of curve estimators when the design is sparse. The conclusions drawn there will be developed into empirical bandwidth choice methods in Section 3. Numerical work will be presented in Section 4. Technical arguments behind our main result in Section 2 will be sketched in Section 5. Further details of proofs are available in Hall, Marron, Neumann and Titterington (1995).

2. Formulas for variance and bias. We begin by introducing the model and defining the estimator. The data $\{Y_i,\ 1 \le i \le n\}$ are assumed to be generated as $Y_i = m(x_i) + \varepsilon_i$, where m is a smooth function, the design points x_i are conditioned values of independent and identically distributed random variables with density f and, conditional on the x_i's, the random variables ε_i are independent with zero mean and variance σ^2. Here, f represents the design density ψ_θ discussed earlier. We take the support, $\mathscr{I}$, of f to be the interval $(0, 1)$, and assume that f is bounded away from zero on $(\xi, 1-\xi)$ for each $\xi > 0$. Put

$$w_i(x) = v_i(x)\left\{\sum_{j=1}^{n} v_j(x) + \delta(x)\right\}^{-1},$$

where for a kernel function K we define $v_i(x) = K\{(x-x_i)/h\}\{s_2 - (x-x_i)s_1\}$ and $s_k = \sum_{j=1}^{n} K\{(x-x_j)/h\}(x-x_j)^k$ $(k = 1, 2)$. The ridge parameter $\delta \ge 0$ is chosen to be nonrandom. Fan (1993) suggests taking δ equal to n^{-2}, and that

choice would be appropriate for our purposes. In this notation, our estimator of m is

$$\widehat{m}(x) = \sum_{i=1}^{n} w_i(x)\, Y_i.$$

Our main result in this section is Theorem 2.1, which shows that, as sample size increases and bandwidth decreases at an appropriate rate, the variance of $\widehat{m}(x)$ decreases like $\{nhf(x \circ h)\}^{-1}$, where $x \circ h = x \vee h$ if $0 \le x \le \frac{1}{2}$, $x \wedge (1-h)$ if $\frac{1}{2} < x \le 1$; and that bias decreases like h^2. A sequence of remarks following the theorem will describe its main implications.

We assume the following conditions.

Of the design density f,

(2.1) f is continuous on $(0, 1)$, bounded away from zero on $(\xi, 1-\xi)$ for each $\xi > 0$ and satisfies $f(x) \sim c_1 x^{\alpha_1}$, $f(1-x) \sim c_2 x^{\alpha_2}$ as $x \downarrow 0$, where $c_1, c_2 > 0$ and $\alpha_1, \alpha_2 \ge 0$.

Of the target function m,

(2.2) m'' is bounded and uniformly continuous on $(0, 1)$.

Of the error distribution,

(2.3) $E(\varepsilon_i) = 0, \qquad E(\varepsilon_i^2) = \sigma^2 > 0$ for each i, $\qquad E(\varepsilon_i \varepsilon_j) = 0$ for each $i \ne j$.

Of the kernel function K,

(2.4) K is bounded, symmetric, Hölder continuous, nonnegative and supported on $(-1, 1)$.

Of the bandwidth function h,

(2.5) for some $\eta > 0$, $h(x) = O(n^{-\eta})$ and $h(x)^{-1}[\{x \vee h(x)\}^{\alpha_1} \wedge \{(1-x) \vee h(x)\}^{\alpha_2}]^{-1} = O(n^{1-\eta})$ uniformly in $0 < x < 1$.

Of the ridge function δ,

(2.6) $$\sup_{0<x<1} \left(n^2 h(x)^6 \left[\{x \vee h(x)\}^{\alpha_1} \wedge \{(1-x) \vee h(x)\}^{\alpha_2}\right]\right)^{-1} \delta(x) \to 0\,.$$

Next we introduce notation for dominant terms in formulas for variance and bias. Put

$$q_{ilk}(w) = \int u^k\, K(u)^l \left\{1 \wedge w - (1 \wedge w^{-1})\, u\right\}_+^{\alpha_i} du, \qquad w > 0,$$

(2.7) $$v_i = \left(q_{i12}^2\, q_{i20} - 2 q_{i11}\, q_{i12}\, q_{i21} + q_{i11}^2\, q_{i22}\right)\left(q_{i10}\, q_{i12} - q_{i11}^2\right)^{-2},$$

(2.8) $$b_i = \left(q_{i12}^2 - q_{i11}\, q_{i13}\right)\left(q_{i10}\, q_{i12} - q_{i11}^2\right)^{-1};$$

and with r denoting either v or b, let $r(x, h) = r_1(xh^{-1})$ or $r_2\{(1-x)\, h^{-1}\}$ according as $0 < x \le \frac{1}{2}$ or $\frac{1}{2} < x < 1$.

THEOREM 2.1. *Assume conditions* (2.1)–(2.6). *Then, for a sequence of design points* $x_1, x_2, \ldots$ *arising with probability* 1,

$$\operatorname{var}\{\widehat{m}(x)\} = \sigma^2[nh(x)\, f\{x \circ h(x)\}]^{-1}[v\{x, h(x)\} + o(1)], \tag{2.9}$$

$$E\{\widehat{m}(x)\} - m(x) = \tfrac{1}{2}\, h(x)^2[m''(x)\, b\{x, h(x)\} + o(1)], \tag{2.10}$$

where the $o(1)$ *terms are of that order uniformly in* $0 < x < 1$.

We intend Theorem 2.1 to be interpreted conditional on the design sequence, and our outline proof is for that setting. However, if the ridge parameter δ is chosen appropriately ($\delta = n^{-2}$ is adequate) then the theorem has a direct analogue for unconditional variance and bias.

REMARK 2.1 (Properties of v and b). Condition (2.4) implies that v and $|b|$ are bounded uniformly in $0 < x < 1$ and $h > 0$. Except in pathological cases, the set of values (x, h) in $(0, 1) \times (0, \infty)$ such that $b(x, h) = 0$ is of measure zero, and $v > 0$. The classical variance and bias formulas, $\operatorname{var}(\widehat{m}) \sim \sigma^2(nhf)^{-1}\,\kappa_1$ and $E(\widehat{m}) - m = \frac{1}{2}\, h^2\, m''\, \kappa_2 + o(h^2)$ where $\kappa_1 = (\int K^2)/(\int K)^2$ and $\kappa_2 = \{\int u^2\, K(u)\, du\}/(\int K)$, are implied by Theorem 2.1 in regions where f is bounded away from zero. Methods of Cheng, Fan and Marron (1996) may be used to show that, using the triangular kernel and an appropriate bandwidth, the mean-squared error for $\widehat{m}(0)$ formed from the variance and the bias given in Theorem 2.1 achieves the minimum possible value, in an asymptotic sense.

REMARK 2.2 (Order-of-magnitude approximations). Observe from (2.9) and (2.10) that balancing variance against squared bias at x produces, in order of magnitude terms, the identity $\{nh(x \vee h)^{\alpha_1}\}^{-1} = h^4$ for $0 < x < \xi$, any fixed $\xi > 0$. The analogous result in the upper tail is also valid. Therefore, the bandwidth $h_0 = h_0(x)$ that minimizes $E\{\widehat{m}(x) - m(x)\}^2$, and the minimum of the latter, satisfy

$$h_0(x) \simeq \begin{cases} n^{-1/(5+\alpha_1)}, & \text{if } 0 < x \le n^{-1/(5+\alpha_1)}, \\ (nx^{\alpha_1})^{-1/5}, & \text{if } n^{-1/(5+\alpha_1)} < x \le \tfrac{1}{2}, \\ \{n(1-x)^{\alpha_2}\}^{-1/5}, & \text{if } \tfrac{1}{2} < x \le 1 - n^{-1/(5+\alpha_2)}, \\ n^{-1/(5+\alpha_2)}, & \text{if } 1 - n^{-1/(5+\alpha_2)} < x < 1, \end{cases}$$

$$\inf_h\, E\{\widehat{m}(x) - m(x)\}^2 \simeq \begin{cases} n^{-4/(5+\alpha_1)}, & \text{if } 0 < x \le n^{-1/(5+\alpha_1)}, \\ (nx^{\alpha_1})^{-4/5}, & \text{if } n^{-1/(5+\alpha_1)} < x \le \tfrac{1}{2}, \\ \{n(1-x)^{\alpha_2}\}^{-4/5}, & \text{if } \tfrac{1}{2} < x \le 1 - n^{-1/(5+\alpha_2)}, \\ n^{-4/(5+\alpha_2)}, & \text{if } 1 - n^{-1/(5+\alpha_2)} < x < 1. \end{cases}$$

Noting the comments in Remark 2.1 concerning zeros of $b(x, h)$ we may show that, except in pathological cases, the approximations above are accurate in

the sense that in both, the $\simeq$ signs may be interpreted as $\asymp$ signs for all but a finite set of x's. That is, except for those x's the ratio of the left- and right-hand sides above is bounded away from zero and infinity as $n \to \infty$. Typical exceptional x's will be near points of inflection of m.

This concise asymptotic interpretation of the approximations is valid uniformly on any set $\mathscr{I}\backslash\mathscr{N}$, where $\mathscr{N}$ denotes any neighborhood, no matter how small, of the finite set of exceptional points noted in the previous paragraph. At those exceptional points, the optimal rate of convergence of $\inf_h E\{\widehat{m}(x) - m(x)\}^2$ is actually faster than that described by the approximation above, and likewise, that approximation provides an upper bound to the fastest rate of convergence throughout the neighborhood $\mathscr{N}$. Arguing thus, it may be shown that the locally optimized mean integrated squared error is given by

$$\begin{aligned} \text{LOMISE} &= \int_0^1 \inf_h E\{\widehat{m}(x) - m(x)\}^2 \, dx \\ &\asymp \begin{cases} n^{-4/5}, & \text{if } \alpha < 5/4, \\ n^{-4/5} \log n, & \text{if } \alpha = 5/4, \\ n^{-5/(5+\alpha)}, & \text{if } \alpha > 5/4 \end{cases} \\ &\asymp n^{-4/5} \int_{n^{-1/(5+\alpha)}}^{1/2} x^{-4\alpha/5} \, dx, \end{aligned} \tag{2.11}$$

where $\alpha = \max(\alpha_1, \alpha_2)$. Note particularly that the ratio of the quantities on the far left and far right sides of (2.11) is bounded away from zero and infinity as $n \to \infty$. When $\alpha < 5/4$, the asymptotic constants of proportionality in (2.11) have particularly simple expressions, and indeed

$$\text{LOMISE} \sim n^{-4/5} \int_0^1 (b_0 v_0^2 f^{-2} |m''|)^{2/5}.$$

Similar expressions are readily developed when $\alpha \geq 5/4$, although they are driven by behavior of f at the ends of $\mathscr{I}$ and are consequently more complex.

REMARK 2.3 (Projection of high-dimensional designs on to lower-dimensional structures). Here we return to the problem discussed in Section 1, that of estimating $m(x) = m_\theta(x) = E(Y|\theta \cdot Z = x)$ from random data on the vector (Y, Z), where Y is a scalar, Z is a p-vector and θ is a unit p-vector. Suppose the support of Z is a bounded, open, contiguous set $\mathscr{D} \subseteq \mathbb{R}^p$ and that Z has a density that is uniformly continuous and bounded away from zero on $\mathscr{D}$. If $\mathscr{D}$ is a rectangular prism then, provided θ is not parallel to any of the sides of $\mathscr{D}$, f satisfies (2.1) with $\alpha_1 = \alpha_2 = p - 1$. (The values are $\alpha_1 = \alpha_2 = 0$ when θ is parallel to an edge. Intermediate cases, where θ is parallel to a side of $\mathscr{D}$ but not to an edge, may be treated similarly.) If $\mathscr{D}$ is an ellipsoid then $\alpha_1 = \alpha_2 = \frac{1}{2}(p - 1)$. Noting these properties, we may deduce from (2.11) that LOMISE $= O(n^{-4/5})$ for all θ's if and only if $p \leq 2$ (when $\mathscr{D}$ is a rectangular prism) or $p \leq 3$ (when $\mathscr{D}$ is an ellipsoid).

3. Empirical bandwidth choice. We begin by applying Theorem 2.1 to develop a formula for the optimal bandwidth. Observe that $v(x, h) = V(x, h) f(x \circ h) + o(1)$ and $b(x, h) = B(x, h) + o(1)$, where

$$V = (f_{12}^2 f_{20} - 2 f_{11} f_{12} f_{21} + f_{11}^2 f_{22})(f_{10} f_{12} - f_{11}^2)^{-2},$$
$$B = (f_{12}^2 - f_{11} f_{13})(f_{10} f_{12} - f_{11}^2)^{-1},$$

and $f_{lk}(x, h) = \int u^k K(u)^l f(x - hu)\, du$. In that notation, $E\{\widehat{m}(x) - m(x)\}^2 \sim \sigma^2 \{nh(x)\}^{-1} V(x, h) + \frac{1}{4} h(x)^4 m''(x)^2 B(x, h)^2$. Minimizing this quantity takes no account of zeros caused by points of inflection of m or zeros of $B(x, h)$. Therefore, we suggest adjoining a small ridge parameter $t > 0$ to the quantity $m''(x)^2 B(x, h)^2$, prior to minimization. Hence, we develop an empirical approximation to the bandwidth h_a that minimizes

$$\sigma^2 (nh)^{-1} V(x, h) + \tfrac{1}{4} h^4 \{m''(x)^2 B(x, h)^2 + t\}. \tag{3.1}$$

Let $\tilde{V}$ and $\tilde{B}$ denote the versions of V and B, respectively, in which each $f_{lk}(x, h)$ is replaced by its unbiassed estimator

$$\tilde{f}_{lk}(x, h) = (nh)^{-1} \sum_{i=1}^{n} \{(x - X_i)/h\}^k K\{(x - X_i)/h\}^l.$$

There is a wide variety of ways of estimating σ^2 [see, e.g., Gasser, Sroka and Jennen-Steinmetz (1986); Buckley, Eagleson and Silverman (1988); Buckley and Eagleson (1989); Hall and Marron (1990); Hall, Kay and Titterington (1990); Carter and Eagleson (1992) and Seifert, Gasser and Wolf (1993)]. For the sake of definiteness, we employ the root-n consistent estimator,

$$\hat{\sigma}^2 = (2\nu)^{-1} \sum_{i:\, \xi < x_i < 1 - \xi} (Y_i - Y_{i+1})^2,$$

where $\xi \in (0, \frac{1}{2})$, ν denotes the number of summands in the series, and it is assumed that the design points x_i are indexed in order of increasing size. We take $\widetilde{m}''(x)$ to be the second derivative estimator constructed using a local sth degree polynomial, as suggested by Ruppert and Wand (1994), employing kernel $K_1 \geq 0$ and bandwidth $h_1 = h_1(x)$. Although Ruppert and Wand (1994) imply that we should use $s \geq 3$, we found that better numerical results were achieved with $s = 2$. Finally, we take $\tilde{h}_a = \tilde{h}_a(x)$ to be the minimizer of

$$\hat{\sigma}^2 (nh)^{-1} \tilde{V}(x, h) + \tfrac{1}{4} h^4 \{\widetilde{m}''(x)^2 \tilde{B}(x, h)^2 + t\}, \tag{3.2}$$

and let $\hat{h}_a = \tilde{h}_a$ if both $\tilde{h}_a$ and $\tilde{h}_a^{-1} \leq n$ and $\hat{h}_a = n^{-1/5}$ otherwise. (The latter restrictions serve only to ensure that the empirical bandwidth selector does not take grossly large or small values. The boundary of n^{-1} is somewhat arbitrary and could be replaced by many alternatives.)

We close this section by showing that the estimator $\hat{h}_a$ achieves first-order minimization of the quantity at (3.1). Let $\hat{h} = \hat{h}(x)$ denote a locally adaptive empirical bandwidth function, representing an estimator of a deterministic bandwidth $h = h(x)$. Write $\check{m}$ for the version of $\widehat{m}$ in which h is replaced

by $\hat{h}$, and recall that in Theorem 2.1 the estimator $\widehat{m}$ was constructed using bandwidth h. Theorem 3.1 below shows that first-order properties of deterministic, locally adaptive bandwidths are preserved by general methods for empirical choice, and that this result holds uniformly in x. From Theorem 3.2 we see that the particular method suggested above satisfies the conditions of Theorem 3.1.

THEOREM 3.1. *Assume the conditions of Theorem* 2.1, *that the errors* ε_i *are identically distributed and that all moments of the error distribution are finite. Suppose too that* m'' *is uniformly continuous*; *that for some* C, $\zeta > 0$,

$$|f(x) - f(y)| \le C\big(|x-y|/[(x \vee y) \wedge \{(1-x) \vee (1-y)\}]\big)^{\zeta} \times [(x \vee y)^{\alpha_1} \wedge \{(1-x) \vee (1-y)\}^{\alpha_2}] \tag{3.3}$$

for all x, $y \in (0,1)$ *and that for some* $\zeta > 0$ *and all* $\lambda > 0$,

$$P\Big\{ \sup_{0<x<1} |\hat{h}(x)\, h(x)^{-1} - 1| > n^{-\zeta} \Big\} = O(n^{-\lambda}), \tag{3.4}$$

$$P\Big[\sup_{0<x<1} \{\hat{h}(x) + \hat{h}(x)^{-1}\} \le n^{1/\zeta} \Big] = 1. \tag{3.5}$$

Then results (2.9) *and* (2.10) *hold for* $\check{m}$ *as well as* $\widehat{m}$.

Condition (3.3) is no more than the usual form of Hölder continuity assumption for a function that, as $x \downarrow 0$, decreases to zero like x^{α_1} in the left-hand tail and x^{α_2} in the right-hand tail. The proofs of Theorems 3.1 and 3.2 may be found in Hall, Marron, Neumann and Titterington (1995).

Our final result is in a form which applies to the left-hand half of the support interval $(0,1)$, and for that purpose we assume that K_1 is not a right-hand kernel, that is, that $K_1(u) \ne 0$ for some $u < 0$.

THEOREM 3.2. *Assume the conditions of Theorem* 2.1, *that* f *and* m'' *are Hölder continuous on* $(0,1)$, *that* K_1 *satisfies* $\int K_1 = 1$ *and is compactly supported and Hölder continuous on* $(-\infty, \infty)$, *that* $K_1(u) \ne 0$ *for some* $u < 0$, *that all moments of the distribution of the errors* ε_i *are finite and that for some* $\eta > 0$, $h_1(x) = O(n^{-\eta})$ *and* $h_1(x)^{-3}\{x \vee h_1(x)\}^{-\alpha_1} = O(n^{1-\eta})$, *where the "*$O$*" terms are of that order uniformly in* $0 < x < 1 - \xi$ *for some* $\xi \in (0,1)$. *Then, for a sequence of design points* $x_1, x_2, \ldots$ *arising with probability* 1, *and for some* $\zeta > 0$ *and all* $\lambda > 0$, *results* (3.4) *and* (3.5) *hold.*

The conditions of the theorem concerning h_1 are satisfied if we take $h_1(x) \equiv n^{-\gamma}$, where $0 < \gamma < (3 + \alpha_1)^{-1}$.

REMARK 3.1. (i) Note that formula (3.1), which is the basis for the data-driven bandwidth choice, is also adequate in cases of low density without the special assumption on the rate of decay of f. [This is in contrast to formula

(2.9), which in particular includes a term $f\{x \circ h(x)\}$ in the case $x \leq 1/2$, and whose validity depends in some way on the monotonic decay of $f(x)$ as $x \to 0$.] Actually (and quite surprisingly, because its derivation was based on asymptotic considerations), the term $\sigma^2(nh)^{-1}\tilde{V}(x, h)$ is *exactly* equal to the conditional variance of $\widehat{m}(x)$, $\text{var}\{\widehat{m}(x)\} = \sigma^2 \sum_i w_i^2(x)$. Hence, also in the case of low density in the interior, its estimate $\hat{\sigma}^2(nh)^{-1}\tilde{V}(x, h)$ is a very reliable estimate of $\text{var}\{\widehat{m}(x)\}$.

(ii) More work could certainly be invested in studying different ways of choosing the ridge constant. However, any such method is likely to be very much ad hoc and to involve arbitrary features.

(iii) One referee was concerned about the finite-sample variance problems associated with local polynomial estimators. These can be ameliorated if one takes, if necessary, a larger bandwidth. Since we make a data-driven local bandwidth choice, the problem of large variances should be in principle avoidable. This seems to be actually the case, because $\hat{\sigma}^2(nh)^{-1}\tilde{V}(x, h)$ is a very good estimate of $\text{var}\{\widehat{m}(x)\}$, as argued above. In this context note also that our criterion based on minimization of (3.2) automatically rejects bandwidths that lead to very high variances.

(iv) In view of the above discussion, it seems adequate to choose the bandwidth based on $\hat{\sigma}^2(nh)^{-1}\tilde{V}(x, h)$ plus some reasonable estimate of the squared bias. The bias part is perhaps "too asymptotic" for small values of n, since it includes only an estimate of $m''(x)$. On the other hand, a completely "nonasymptotic" estimate like $|\widehat{m}(x) - \sum_i w_i(x)\widehat{m}(x_i)|^2$ (with $\widehat{m}$ computed using a larger bandwidth), which would also not require the additional ridge parameter t, while being a possible alternative, is not so elegant.

4. Numerical example. A numerical study was made of the effects of implementing the ideas in this paper. In the illustration reported here the true curve was the parabola $m(x) = \frac{1}{2}x(1 - x)$, and the design density was chosen to be the piecewise quadratic, smooth, bell-shaped density defined by

$$f(x) = \begin{cases} 16x^2, & \text{if } 0 \leq x \leq 1/4, \\ 2 - 4(1 - 2x)^2, & \text{if } 1/4 \leq x \leq 3/4, \\ 16(1 - x)^2, & \text{if } 3/4 \leq x \leq 1. \end{cases}$$

Figure 1 displays results corresponding to sample size $n = 10000$ and noise standard deviation $\sigma = 0.05$. Displayed in each frame in the figure are the curve estimates from seven replications, along with their average (dashed curve) and the true parabolic curve (solid curve). The noise variance was estimated as described in Section 3. For the results in Figure 1(a), the bandwidth was chosen to be constant, at a level that was locally effective for x in the middle of the range. The instability of the resulting curve estimates near the ends of the interval is clear. In the other two displays, the bandwidth was chosen to vary with x, according to the methodology of Section 3. In Figure 1(b), the true value of $m''(x)$ was assumed known ($= -1$, for all x). Clearly, the results are much better, although there remains, inevitably, some disparity in

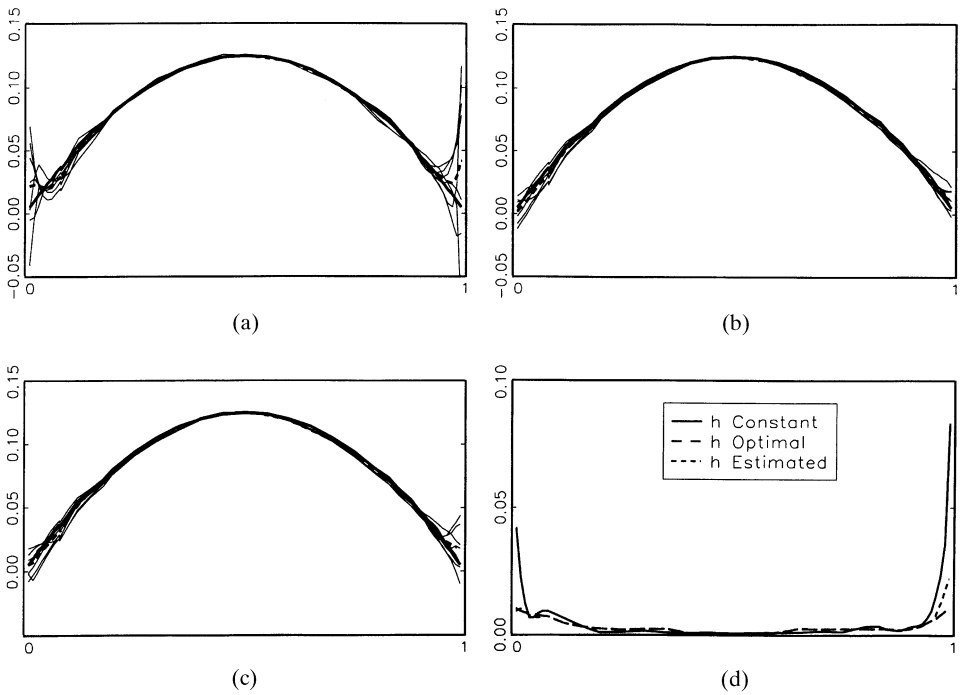

FIG. 1. *Target curve* (*heavy solid line*), *seven realizations of local linear estimates* (*thin solid line*), *and the average of the estimates* (*heavy dashed line*), *based on* $n = 10,000$ *and* $\sigma = 0.05$, *using* (a) *constant bandwidth,* (b) *location adaptive bandwidth based on true* m'', (c) *location adaptive bandwidth based on estimated* m'', (d) *root mean-squared errors associated with the three methods.*

performance between the middle and the ends of the interval. In Figure 1(c), $m''(x)$ was estimated by $\widetilde{m}''(x)$, as suggested in Section 3, and constructed using a local quadratic. Not surprisingly, the results are degraded relative to those in Figure 1(b), but they show an undeniable improvement over those in Figure 1(a). For Figures 1(b) and 1(c), the ridge parameter t was set to zero. Increasing t initially led to little change in the curve estimates, although the values chosen for the bandwidth in the middle of the interval changed somewhat. Further increase in t led to increasingly noticeable biases in the curve estimates. Figure 1(d) displays the estimated root mean-squared errors from the three methods.

Figure 2 displays corresponding results for the case of $n = 500$ and $\sigma = 0.02$. The case of $n = 1000$ and $\sigma = 0.05$, was also considered and led to qualitatively similar conclusions.

In summary, for this particular true curve, the experiments showed that the choice of smoothing parameter in the middle of the interval was not a particularly sensitive issue, but that considerable disadvantage accrued if the sparsity of the design near the ends was ignored by choosing too small a bandwidth.

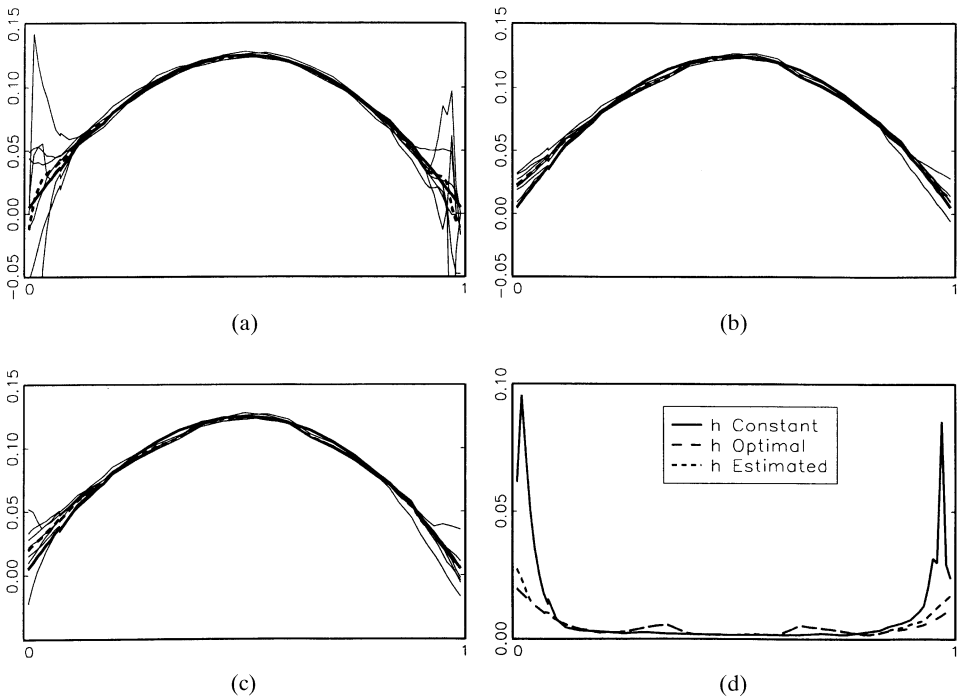

FIG. 2. *Target curve* (*heavy solid line*), *seven realizations of local linear estimates* (*thin solid line*), *and the average of the estimates* (*heavy dashed line*), *based on* $n = 500$ *and* $\sigma = 0.02$, *using* (a) *constant bandwidth,* (b) *location adaptive bandwidth based on true* m'', (c) *location adaptive bandwidth based on estimated* m'', (d) *root mean-squared errors associated with the three methods.*

5. Outline of proof of Theorem 2.1. Steps (i)–(iv) and (v) treat variance and bias contributions, respectively.

Step (i). *Preliminaries*. Observe that $\operatorname{var}(\widehat{m}) = \sigma^2(\sum v_i^2)/(\sum v_i + \delta)^2$. Put $Q_{lk}(x) = (nh)^{-1} \sum\{(x - x_i)/h\}^k K\{(x - x_i)/h\}^l$, for $l = 1, 2$ and $k = 0, 1, 2, 3$. In this notation, $\sum v_i = n^2 h^4(Q_{10} Q_{12} - Q_{11}^2)$ and $\sum v_i^2 = n^3 h^7(Q_{12}^2 Q_{20} - 2Q_{11} Q_{12} Q_{21} + Q_{11}^2 Q_{22})$. Therefore, with $\delta_1 = (n^2 h^4)^{-1} \delta$,

$$\begin{aligned} &nh \operatorname{var}(\widehat{m}) \\ &\quad = \sigma^2(Q_{12}^2 Q_{20} - 2Q_{11} Q_{12} Q_{21} + Q_{11}^2 Q_{22})(Q_{10} Q_{12} - Q_{11}^2 + \delta_1)^{-2}. \end{aligned} \tag{5.1}$$

Step (ii). *Nominal expected value of* Q_{lk}. If we regard $x_1, \ldots, x_n$ as independent random variables with density f, then $E\{Q_{lk}(x)\} = \int u^k K(u)^l f(x - hu)\, du$, whence it follows that as $h \to 0$,

$$\sup_{0 < x \le 1 - \xi} \left| E\{Q_{lk}(x)\} f(x \vee h)^{-1} - q_{lk}(xh^{-1}) \right| \to 0\,, \tag{5.2}$$

for each $\xi > 0$, where

$$q_{lk}(w) = \int u^k K(u)^l \{1 \wedge w - (1 \wedge w^{-1})\, u\}_+^{\alpha_1}\, du\,.$$

The analogous result, where the supremum in (5.2) is taken over $\xi \le x < 1$, is also true, by the same argument.

Step (iii). *Nominal error about mean of* Q_{lk}. Again treating $x_1, \dots, x_n$ as independent random variables, we may show that, uniformly in $0 < x < 1$,

$$nh \operatorname{var}\{Q_{lk}(x)\} \le C_1 \int_{-1}^{1} f(x - hu)\, du \le C_2 (x \vee h)^{\alpha_1},$$

where $C_1, C_2, \dots$ denote positive generic constants depending only on f, K, σ and any indicated arguments. Therefore, applying Rosenthal's inequality [see Hall and Heyde (1980), page 23], we obtain for each $r \ge 1$,

$$E|Q_{lk}(x) - E\{Q_{lk}(x)\}|^{2r} \le C_3(r)\big[\{(nh)^{-1}(x \vee h)^{\alpha_1}\}^r + (nh)^{-(2r-1)}(x \vee h)^{\alpha_1}\big].$$

It follows that if $h = h(x)$ is chosen so that $\inf_{0<x<1} h(x)\{x \vee h(x)\}^{\alpha_1} > C_4 n^{\eta_1 - 1}$ for some $\eta_1 > 0$, then

$$\sup_{0<x<1} \{x \vee h(x)\}^{-2r\alpha_1} E|Q_{lk}(x) - E\{Q_{lk}(x)\}|^{2r} = O(n^{-r\eta_1})$$

for all $k \ge 1$. If k is even, then $E\{Q_{lk}(x)\} \ge C_5(\xi)(x \vee h)^{\alpha_1}$ uniformly in $0 < x \le 1 - \xi$. Hence by Markov's inequality,

$$\sup_{0<x\le 1-\xi} P\big[|Q_{lk}(x) - E\{Q_{lk}(x)\}| > \eta_2 |E\{Q_{lk}(x)\}|\big] = O(n^{-\lambda})$$

for all even k and all η_2, $\lambda > 0$. The identity is valid for all $k \ge 1$ if we replace $\eta_2 |E\{Q_{lk}(x)\}|$ by $\eta_2 (x \vee h)^{\alpha_1}$. Therefore, if $\mathscr{A} = \mathscr{A}(n)$ represents any set consisting of $O(n^a)$ elements of $(0, 1 - \xi]$, for arbitrary but fixed $a > 0$, then for even m,

$$P\Big[\sup_{x\in\mathscr{A}} |Q_{lk}(x) - E\{Q_{lk}(x)\}|\,|E\{Q_{lk}(x)\}|^{-1} > \eta_2\Big] = O(n^{-\lambda})$$

for all $\eta_2, \lambda > 0$. The Hölder continuity of K may be used to extend the supremum to all $x \in (0, 1 - \xi]$:

$$P\Big[\sup_{0<x\le 1-\xi} |Q_{lk}(x) - E\{Q_{lk}(x)\}|\,|E\{Q_{lk}(x)\}|^{-1} > \eta_2\Big] = O(n^{-\lambda}).$$

It follows that there exists a class $\mathscr{X}_1$ of sequences $(x_1, x_2, \dots)$ such that $P\{(x_1, x_2, \dots) \in \mathscr{X}_1\} = 1$ if we regard $x_1, x_2, \dots$ as independent and identically distributed random variables with density f, and also,

$$\sup_{(x_1,x_2,\dots)\in\mathscr{X}_1} \sup_{0<x\le 1-\xi} |Q_{lk}(x) - E\{Q_{lk}(x)\}|\,|E\{Q_{lk}(x)\}|^{-1} \to 0. \tag{5.3}$$

The analogous result for the upper tail follows by the same argument. [Result (5.3) is valid for all even $k \ge 0$ and for all $k \ge 0$ if the factor $|E\{Q_{lk}(x)\}|^{-1}$ is replaced by $(x \vee h)^{-\alpha_1}$.]

Step (iv). *Asymptotic variance of* $\widehat{m}$. By (5.2) and (5.3),

$$\sup_{(x_1,x_2,\ldots)\in\mathscr{X}_1} \sup_{0<x\le 1-\varepsilon} \left|Q_{lk}(x)\, f(x\vee h)^{-1} - q_{lk}(xh^{-1})\right| \to 0\,.$$

Provided that $\sup_{0<x\le 1-\xi}\left[n^2 h(x)^4\{x\vee h(x)\}^{2\alpha_1}\right]^{-1}\delta(x)\to 0$, this result and (5.1) ensure that

$$\sup_{(x_1,x_2,\ldots)\in\mathscr{X}_1} \sup_{0<x\le 1-\xi} |nh(x)\, f(x\vee h)\, \mathrm{var}\{\widehat{m}(x)\} - \sigma^2 v\{x, h(x)\}| \to 0\,.$$

Formula (2.9) follows from this fact and its analogue for the right-hand tail.

Step (v). *Calculation of bias*. Note that

$$E(\widehat{m}) - m = \Big\{\sum v_i(m_i - m) - \delta m\Big\}\Big/\Big(\sum v_i + \delta\Big),$$

where $m_i = m(x_i)$. By Taylor expansion,

$$\sum v_i(m_i - m) = \tfrac{1}{2}\, n^2\, h^6\big\{m''\big(Q_{12}^2 - Q_{13}\, Q_{11}\big) + \big(R_1\, Q_{12} - R_2\, Q_{11}\big)\big\}\,, \tag{5.4}$$

where

$$R_j(x) = (nh)^{-1}\sum_{i=1}^{n}\{(x-x_i)/h\}^{j+1}[m''\{x+\theta_i(x_i-x)\} - m''(x)]\, K\{(x-x_i)/h\}$$

and θ_i (depending only on x, x_i and m) satisfies $0\le\theta_i\le 1$. The argument leading to (5.2) and (5.3) and uniform continuity of m'' may be employed to prove that there exists a set $\mathscr{X}_2$ of sequences $(x_1, x_2, \ldots)$ such that $P\{(x_1, x_2, \ldots)\in\mathscr{X}_2\} = 1$ (if $x_1, x_2, \ldots$ are regarded as independent and identically distributed random variables with density f) and

$$\sup_{(x_1,x_2,\ldots)\in\mathscr{X}_2} \sup_{0<x\le 1-\xi} |R_j(x)|(x\vee h)^{-\alpha_1} \to 0\,. \tag{5.5}$$

Results (5.2)–(5.5) imply that provided

$$\sup_{0<x\le 1-\varepsilon} [n^2\, h(x)^6\{x\vee h(x)\}^{2\alpha_1}]^{-1}\,\delta(x)\to 0,$$

we have

$$\sup_{(x_1,x_2,\ldots)\in\mathscr{X}_3} \sup_{0<x\le 1-\xi} h^{-2}\Big|E(\widehat{m}) - m - \tfrac{1}{2}\, h^2\, m''\big(q_{12} - q_{13}\, q_{11}\big) \times \big(q_{10}\, q_{12} - q_{11}^2\big)^{-1}\Big| \to 0 \tag{5.6}$$

as $n\to\infty$, where $\mathscr{X}_3 = \mathscr{X}_1\cap\mathscr{X}_2$, the functions h, m, m'' and $\widehat{m}$ are evaluated at x and the functions q_{lk} are evaluated at $xh(x)^{-1}$. Result (2.10) follows from (5.6) and its analogue in the upper tail.

Acknowledgments. The helpful comments of the Editor and two referees on a previous version of the paper were much appreciated.

REFERENCES

BRILLINGER, D. R. (1983). A generalized linear model with "Gaussian" regressor variables. In *A Festschrift for Erich Lehmann* (P. J. Bickel, K. A. Doksum and J. L. Hodges, eds.) 97–114. Wadsworth, Belmont, CA.

BUCKLEY, M. J. and EAGLESON, G. K. (1989). A graphical method for estimating the residual variance in nonparametric regression. *Biometrika* **76** 203–210.

BUCKLEY, M. J., EAGLESON, G. K. and SILVERMAN, B. W. (1988). The estimation of residual variance in nonparametric regression. *Biometrika* **75** 189–199.

CARTER, C. K. and EAGLESON, G. K. (1992). A comparison of variance estimators in nonparametric regression. *J. Roy. Statist. Soc. Ser. B* **54** 773–780.

CHENG, M. Y., FAN, J. and MARRON, J. S. (1996). Minimax efficiency of local polynomial fit estimators at boundaries. Mimeo Series No. 2098, Institute of Statistics, Univ. North Carolina.

CLEVELAND, W. S. (1979). Robust locally weighted regression and smoothing scatterplots. *J. Amer. Statist. Assoc.* **74** 829–836.

CLEVELAND, W. S. (1993). *Visualizing Data*. Hobart Press, Summit, NJ.

CLEVELAND, W. S. and DEVLIN, S. J. (1988). Locally weighted regression: an approach to regression analysis by local fitting. *J. Amer. Statist. Assoc.* **83** 596–610.

CLEVELAND, W. S. and GROSSE, E. H. (1991). Computational methods for local regression. *Statist. and Computing* **1** 47–62.

CLEVELAND, W. S. and LOADER, C. (1996). Smoothing by local regression: principles and methods. In *Statistical Theory and Computational Aspects of Smoothing* (W. Härdle and M. Schmlick, eds.) 10–49. Physica, Heidelberg.

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

FAN, J. (1993). Local linear smoothers and their minimax efficiency. *Ann. Statist.* **21** 196–216.

FAN, J. and GIJBELS, I. (1995). Data-driven bandwidth selection in local polynomial fitting: variable bandwidth and spatial adaptation. *J. Roy. Statist. Soc. Ser. B* **57** 371–394.

FRIEDMAN, J. H. and STUETZLE, W. (1981). Projection pursuit regression. *J. Amer. Statist. Assoc.* **76** 817–823.

GASSER, T., KNEIP, A. and KÖHLER, W. (1991). A flexible and fast method for automatic smoothing. *J. Amer. Statist. Assoc.* **86** 643–652.

GASSER, T., SROKA, L. and JENNEN-STEINMETZ, C. (1986). Residual variance and residual pattern in nonlinear regression. *Biometrika* **73** 625–633.

HALL, P. and HEYDE, C. C. (1980). *Martingale Limit Theory and Its Application*. Academic Press, New York.

HALL, P., KAY, J. W. and TITTERINGTON, D. M. (1990). Asymptotically optimal difference-based estimation of variance in nonparametric regression. *Biometrika* **77** 521–528.

HALL, P. and MARRON, J. S. (1990). On variance estimation in nonparametric regression. *Biometrika* **77** 415–419.

HALL, P., MARRON, J. S., NEUMANN, M. H. and TITTERINGTON, D. M. (1995). On local linear smoothing when the design density is low. Research Report SRR027-95, Centre for Mathematics and Its Applications, Australian National Univ., Canberra.

HÄRDLE, W., HALL, P. and ICHIMURA, H. (1993). Optimal smoothing in single-index models. *Ann. Statist.* **21** 157–178.

HASTIE, T. and LOADER, C. (1993). Local regression: automatic kernel carpentry. *Statist. Sci.* **8** 120–143.

NICHOLLS, D. F. and PAGAN, A. R. (1985). Varying coefficient regression. In *Handbook of Statistics* **5** (E. J. Hannan, P. R. Krishnaiah and M. M. Rao, eds.) 413–449. North-Holland, Amsterdam.

RAJ, B. and ULLAH, A. (1981). *Econometrics: A Varying Coefficients Approach*. Croom-Helm, London.

RUPPERT, D. and WAND, M. P. (1994). Multivariate locally weighted least squares regression. *Ann. Statist.* **22** 1346–1370.

SEIFERT, B., GASSER, T. and WOLF, A. (1993). Nonparametric estimation of residual variance revisited. *Biometrika* **80** 373–383.

SEIFERT, B. and GASSER, T. (1996). Finite sample analysis of local polynomials: analysis and solutions. In *Statistical Theory and Computational Aspects of Smoothing* (W. Härdle and M. Schmlick, eds.) 50–79. Physica, Heidelberg.

PETER HALL
CENTRE FOR MATHEMATICS
AND ITS APPLICATIONS
AUSTRALIAN NATIONAL UNIVERSITY
CANBERRA, ACT 0200
AUSTRALIA

J. S. MARRON
DEPARTMENT OF STATISTICS
UNIVERSITY OF NORTH CAROLINA
CHAPEL HILL, NORTH CAROLINA

M. H. NEUMANN
WEIERSTRASS-INSTITUT FÜR
ANGEWANDTE ANALYSIS UND STOCHASTIK
BERLIN
GERMANY

D. M. TITTERINGTON
DEPARTMENT OF STATISTICS
UNIVERSITY OF GLASGOW
GLASGOW G12 8QQ
SCOTLAND