The Annals of Statistics
1989, Vol. 17, No. 2, 573–588

ON PROJECTION PURSUIT REGRESSION

BY PETER HALL

Australian National University

We construct a tractable mathematical model for kernel-based projection pursuit regression approximation. The model permits computation of explicit formulae for bias and variance of estimators. It is shown that the bias of an orientation estimate dominates error about the mean—indeed, the latter is asymptotically negligible in comparison with bias. However, bias and error about the mean are of the same order in the case of projection pursuit curve estimates. Implications of our formulae for bias and variance are discussed.

1. Introduction. Difficulties which traditional estimators have coping with high-dimensional problems may be described in terms of "data sparseness." If a given amount of data is distributed in space, then the distance between adjacent data points increases with increasing dimension. Friedman and Stuetzle [5], Section 1 and Huber [10], Section 1, give numerical examples of this behaviour. Standard techniques of function estimation respond to data sparseness by giving more emphasis than they should to "transitory" features, such as randomly occurring clusters of data points. This behaviour is due largely to the fact that variance of traditional estimators increases rapidly with increasing dimension. Even though "optimal" rates of consistency (Stone [14, 15]) are achieved by balancing variance against squared bias, the optimum in high dimensions occurs with a high level of variance. For example, in a p-dimensional problem, assuming r bounded derivatives, the optimal convergence rate in mean squared error terms is $n^{-2r/(2r+p)}$, which is very poor if, say, $r = 2$ and $p = 4$. In principle, unwanted transitory features could be suppressed by deliberately constructing a suboptimal estimator, with smaller variance and greater bias. An example would be a kernel estimator with a relatively large window. Unfortunately, this modification flattens out *all* features, wanted or not, so that much of the baby is thrown out with the bathwater. In contrast, projection pursuit places emphasis on lower-dimensional features, which it estimates accurately with relatively low variance. It does not suffer from the "flattening" debility of low-variance, multivariate kernel estimators.

Our aim in the present article is to construct a tractable mathematical model describing projection pursuit regression, and to analyse this model so as to shed light on the way in which an "estimated" projective approximation tracks a "theoretical" projective approximation. The main results are as follows. We provide explicit formulae for bias and error about the mean in orientation estimates and curve estimates—see Theorems 4.2 and 4.4 in Section 4. These results show that the estimate of orientation has most of its error in the form of

Received May 1987; revised June 1988.

AMS 1980 *subject classifications*. Primary 62H99; secondary 62H05.

Key words and phrases. Bias, kernel method, nonparametric regression, orientation, projection pursuit.

bias—error about the mean is asymptotically negligible in comparison with bias. We also prove that the common form of kernel-based projection pursuit regression does estimate projections with convergence rates identical to those encountered in one-dimensional problems, although a greater degree of smoothness (in fact, an extra derivative) must be assumed to achieve this end. The extra derivative appears in the formula for bias of the projective approximation, and is needed to remedy difficulties which the estimator has orienting the projection in the right direction. This problem could be alleviated by first constructing an undersmoothed regression estimate, using that estimate to find the right orientation, and then reconstructing the regression estimate with the correct amount of smoothing. We shall discuss this two-stage procedure at the end of Section 4. Such two-stage estimators do improve the convergence rate of orientation estimators, but do not alter the convergence rate of curve estimates. They exacerbate the numerical problem of multiple minima of the orientation function.

We believe this is the first time that concise formulae have been given for variance and bias of kernel-based projection pursuit estimators. There is nothing unexpected about the variance formula, it being entirely analogous to its counterpart for univariate kernel estimators. However, the bias formula is considerably more complex than that for classical kernel estimators, due to bias in the estimate of orientation. It contains a contribution from the extra derivative discussed above.

The article is structured as follows. Our mathematical model for projection pursuit is described and justified in Section 2. Essential calculus for projective approximation is developed in Section 3. Section 4 states our main results, from which follow the conclusions discussed above. Proofs of theorems from Section 4 are given in Section 5.

Several generalizations of our arguments are possible. For example, there is no need to take the smoothing parameter (window size) to be nonrandom, as we do. Our results remain true for random windows which, when divided by our nonrandom window, have limit infimum bounded away from 0 and limit supremum bounded, with probability 1. And of course, one may use different windows for different projective approximations. The only change necessary to proofs is that the "continuity argument" [step (i) in Section 5] must be applied more often.

In common with most investigations of projection pursuit, we confine our attention to univariate projections. However our techniques may be employed to study q-dimensional projections, for any q less than the dimension p of the target function.

The concept and philosophy of projection pursuit in general, and of projection pursuit regression in particular, have been reviewed, consolidated and extended by Huber [10]. The idea of projecting multidimensional data onto a lower-dimensional subspace so as to obtain accurate estimates of lower-dimensional features goes back to Kruskal [11, 12], Switzer [18], Switzer and Wright [19] and Friedman and Tukey [6]. See also Stone's [17] work on dimension reduction. Some theory for projection pursuit has been developed by Diaconis and

Freedman [1], Donoho and Johnstone [3] and Fill and Johnstone [4]. See also Diaconis and Shashahani [2].

For the sake of brevity and clarity we shall confine our attention to the first stage of the projection pursuit algorithm, estimating the first projective approximation. Later projective approximations may be estimated similarly, with properties similar to those of the first.

By way of notation, we shall assume that the set of explanatory variables $\{x_k,\ 1 \le k \le n\}$ is a random sample from a p-variate distribution with density f. At each x_k a univariate observation Y_k is made, with the property that $E(Y_k|x_k = x) = G(x)$, where G is the target function. Let Ω be the set of all p-dimensional unit vectors, and let $\theta, \theta_1, \theta_2, \ldots$ be elements of Ω. If H is a p-variate function such as f or G, then the directional derivative of H in direction θ will be denoted by

$$H_{(\theta)}(x) \equiv \lim_{u \to 0} \{H(x + u\theta) - H(x)\}/u,$$

assuming that this limit exists. Higher-order derivatives will be written as $H_{(\theta_1, \theta_2)} = (H_{(\theta_1)})_{(\theta_2)}$ and so forth. The ith component of a vector $x \in \mathbb{R}^p$ will be written as $x^{(i)}$, so that the usual dot or scalar product is $x \cdot y \equiv \Sigma x^{(i)} y^{(i)}$. In this notation, the norm of x is $\|x\| = (x \cdot x)^{1/2}$. The symbol $\mathscr{A}$ will denote a subset of $\mathbb{R}^p$, usually convex, and $I(\cdot \in \mathscr{A})$ will be the indicator function of $\mathscr{A}$: $I(x \in \mathscr{A}) = 1$ if $x \in \mathscr{A}$, 0 otherwise. We shall reserve the symbol u for a real number.

2. Projective approximation. Let G be a function from $\mathbb{R}^p$ to $\mathbb{R}$, let f be a probability density on $\mathbb{R}^p$ and let X be a p-variate random variable with density f. For scalar u, put

$$g_\theta(u) \equiv E\{G(X)|\theta \cdot X = u\}, \qquad \theta \in \Omega. \tag{2.1}$$

The *first projective approximation* to G within a region $\mathscr{A} \subseteq \mathbb{R}^p$, and relative to f, is that function $G_1(x) \equiv g_{\theta_1}(\theta_1 \cdot x)$, where θ_1 minimizes

$$S(\theta) \equiv E\big[\{G(x) - g_\theta(\theta \cdot X)\}^2 I(X \in \mathscr{A})\big]. \tag{2.2}$$

We assume that the minimum is attained uniquely, except for the sign change $\theta \mapsto -\theta$.

The first projective approximation may be estimated from data as follows. Let $\{(Y_k, x_k),\ 1 \le k \le n\}$ be a sequence of pairs of observations such that $E(Y_k|x_k = x) = G(x)$ for each k, and $\{x_1, \ldots, x_n\}$ is a random sample from the p-variate distribution with density f. Write f_θ for the density of $\theta \cdot x$, which might be called the "marginal density of X in direction θ." A kernel estimate of f_θ, excluding the kth sample value x_k, is

$$\hat{f}_{\theta\langle k \rangle}(u) \equiv \{(n-1)h\}^{-1} \sum_{j \ne k} K\{(u - \theta \cdot x_j)h^{-1}\},$$

where $h > 0$ is window size and K is the kernel function. A kernel estimate of

g_θ, excluding x_k, is

$$(2.3)\qquad \hat{g}_{\theta\langle k\rangle}(u) \equiv \left[\{(n-1)h\}^{-1}\sum_{j\neq k} Y_j K\{(u-\theta\cdot x_j)h^{-1}\}\right]\Big/\hat{f}_{\theta\langle k\rangle}(u).$$

We estimate θ_1 as that value $\hat{\theta}_1 \in \Omega$ which minimizes

$$(2.4)\qquad \hat{S}(\theta) \equiv n^{-1}\sum_{k=1}^{n}\{Y_k - \hat{g}_{\theta\langle k\rangle}(\theta\cdot x_k)\}^2 I(x_k \in \mathscr{A}).$$

An estimate of the first projective approximation to G within $\mathscr{A}$, is $\hat{G}_{1\langle k\rangle}(x) \equiv \hat{g}_{\hat{\theta}_1\langle k\rangle}(\hat{\theta}_1\cdot x)$. We shall assume that K is continuous, so that minima are *achieved*.

Our reason for restricting attention to estimation over the set $\mathscr{A}$ is that the estimator which we use to estimate G_1, and which is also employed by practitioners, has a density estimator in its denominator. That denominator takes values close to 0 near the boundary of the support of f. It is common practice in the theory of regression estimation to study estimators on sets bounded away from those troublesome regions (e.g., [13], page 239ff.), and we very much regret that we found such assumptions necessary in our work too. The variance of kernel regression estimators can be excessively large towards the edges of the support [see formula (2.10) below], and there is every likelihood that projection pursuit estimators will also perform poorly in that region.

Our only purpose in excluding the kth observation x_k from the estimators above is to remove extraneous bias terms when estimating $\hat{\theta}_1$. Once $\hat{\theta}_1$ has been determined, we can put x_k back. Define

$$\hat{f}_\theta(u) \equiv (nh)^{-1}\sum_{k=1}^{n} K\{(u-\theta\cdot x_k)h^{-1}\},$$

$$(2.5)\qquad \hat{g}_\theta(u) \equiv \left[(nh)^{-1}\sum_{k=1}^{n} Y_k K\{(u-\theta\cdot x_k)h^{-1}\}\right]\Big/\hat{f}_\theta(u),$$

$$\hat{G}_1(u) \equiv \left[(nh)^{-1}\sum_{j=1}^{n} Y_j K\{(u-\hat{\theta}_1\cdot x_j)h^{-1}\}\right]\Big/\hat{f}_{\hat{\theta}_1}(u),$$

for real numbers u. We call $\hat{G}_1(u)$ "the" estimate of the first projective approximation to G within the region $\mathscr{A}$ and relative to f.

As we remark in Section 4, it is relatively easy to show that $\hat{\theta}_1, \hat{G}_1$ are consistent for θ_1, G_1, respectively. Our aim in this article is to describe the *rate* of consistency.

We conclude this section by giving details of the kernel K and window h. Recall that our kernel estimators are all one-dimensional, and so our choice of kernel will be dictated by smoothness conditions on one-dimensional projections of f and G. Restrictions of this nature are virtually the same as conditions imposed on directional derivatives. Therefore we assume:

(2.6) the first r directional derivatives of $f(x)$ and $G(x)$ exist and are continuous uniformly in $x \in \mathbb{R}^p$ and in all directions.

Define

$$\mathscr{A}^{\varepsilon} \equiv \{x \in \mathbb{R}^p\colon \text{for some } y \in \mathscr{A},\ \|x - y\| \le \varepsilon\}. \tag{2.7}$$

So that we do not have to impose integrability conditions and to keep the integrand in estimators such as $\hat{g}_j$ bounded away from 0, we further suppose:

(2.8) f vanishes outside a compact set, and is bounded away from 0 on $\mathscr{A}^{\varepsilon}$ for some $\varepsilon > 0$.

To ensure that the set $\{\theta \cdot x\colon x \in \mathscr{A}\}$ is a proper interval, for each $\theta \in \Omega$, we assume that $\mathscr{A}$ is a nonempty, open, p-dimensional convex set.

For fixed θ, estimators such as $\hat{f}_{\theta\langle k\rangle}$, $\hat{g}_{\theta\langle k\rangle}$, $\hat{f}_{\theta}$ and $\hat{g}_{\theta}$ are classical one-dimensional kernel estimators, based on the univariate sample $\{\theta \cdot x_k, 1 \le k \le n\}$, and being of the type reviewed by Prakasa Rao [13], Chapter 4, Section 4.2. Under conditions (2.6) and (2.8), fast rates of convergence may be obtained using an rth-order kernel, just as in the case of kernel density estimators [13, page 42ff.]. Therefore we stipulate that K fulfill the condition:

(2.9) $\int_{-\infty}^{\infty} u^j K(u)\,du = 1$ for $j = 0$, 0 for $1 \le j \le r - 1$; and K is Hölder continuous and compactly supported.

Hölder continuity means that for some $s > 0$ and $C > 0$, and all real u, v, $|K(u) - K(v)| \le C|u - v|^s$. Compact support and Hölder continuity are needed for the "continuity argument" in Section 5. The integrability condition in (2.9) characterizes an rth-order kernel.

We now specify window size. Consider the model

$$Y_k = G(x_k) + \varepsilon_k, \qquad 1 \le k \le n,$$

where the ε_k's are independent and identically distributed with zero mean and finite variance σ^2, and are stochastically independent of the x_k's. Assume conditions (2.6), (2.8) and (2.9), and that $h = h(n) \to 0$ and $nh \to \infty$. Then for fixed $\theta \in \Omega$, and assuming $f_\theta(u) > 0$,

$$\begin{aligned}\hat{g}_\theta(u) = g_\theta(u) &+ (nh)^{-1/2}\Big[E\big\{(G(X) - g_\theta(u))^2 | \theta \cdot X = u\big\} + \sigma^2\Big]^{1/2} \\ &\times \{f_\theta(u)\}^{-1/2}\Big(\int K^2\Big)^{1/2} Z(u) \\ &+ h^r \operatorname{const}(u, \theta) + o(h^r),\end{aligned} \tag{2.10}$$

where $Z(u)$ is asymptotically normal $N(0, 1)$. The rate of convergence of $\hat{g}_\theta(u)$ to $g_\theta(u)$ is maximized at $O_p(n^{-r/(2r+1)})$ by taking $h \sim \text{const.}\, n^{-1/(2r+1)}$, and this will be our choice of window in Sections 3, 4 and 5. The case $r = 2$ is by far the most common, and that is treated in detail for univariate kernel estimators by Prakasa Rao [13], page 239ff.

3. Calculus for projective approximation.

Let $\theta, \theta_0 \in \Omega$, with θ_0 fixed and θ converging to θ_0. To introduce Taylor expansions for quantities such as $S(\theta)$, let θ_{00} be either one of the two unit vectors in the same plane as both θ and θ_0, and perpendicular to θ_0. Provided θ is in the same half of the θ_0, θ_{00} plane as

θ_0, which must very quickly become the case since θ is converging to θ_0, we may write

$$(3.1) \qquad \theta = (1-\eta^2)^{1/2}\theta_0 + \eta\theta_{00},$$

where $-1 \le \eta \le 1$. This representation is unique up to the transformation $(\eta, \theta_{00}) \mapsto (-\eta, -\theta_{00})$, and $\eta = \theta \cdot \theta_{00} \to 0$ as $\theta \to \theta_0$.

Under mild regularity conditions, $S(\theta)$ admits a Taylor expansion of the form

$$(3.2) \qquad S(\theta) = S(\theta_0) + \eta S_1(\theta_0, \theta_{00}) + \tfrac{1}{2}\eta^2 S_2(\theta_0, \theta_{00}) + o(\eta^2)$$

as $\theta \to \theta_0$, for suitable functions S_1 and S_2. The following theorem makes this explicit.

THEOREM 3.1. *Assume that the first two directional derivatives of f and G exist and are bounded and continuous uniformly in $x \in \mathbb{R}^p$ and in all directions; that $\mathscr{A}$ is a nonempty, open, p-dimensional convex set whose boundary has two continuous derivatives; and that f satisfies* (2.8). *Let θ_0, θ_{00} be perpendicular unit vectors, and define $\theta = \theta(\theta_0, \theta_{00})$ by* (3.1). *There exist uniformly continuous functions S_1 and S_2 of θ_0 and θ_{00}, not depending on η, such that* (3.2) *holds uniformly in θ_0, θ_{00} as $\eta \to 0$.*

Such results may be developed from expansions of Radon transforms, as we now show. Let $\mathscr{T}$ be a p-dimensional sphere of radius t centered at the origin, and choose t sufficiently large for $\mathscr{T}$ to contain the support of f. [Recall from (2.8) that we assume f to have compact support.] Given $\theta \in \Omega$ and $u \in \mathbb{R}$, define $\Gamma_\theta = \Gamma_\theta(u)$ to be the $(p-1)$-dimensional "surface" formed from the set of points $\{x \in \mathscr{T}\colon \theta \cdot x = u\}$. Let $d\gamma_\theta(x)$ be an element of $(p-1)$-dimensional content, situated at $x \in \Gamma_\theta$ and aligned so that its normal is parallel to θ. Define the Radon transform (e.g., Helgason [9], page 2)

$$(3.3) \qquad A(u,\theta) \equiv \int_{\Gamma_\theta} a(x)\, d\gamma_\theta(x).$$

THEOREM 3.2. *Assume that the first two directional derivatives of a exist and are continuous uniformly in $x \in \mathscr{T}$ and in all directions. Let θ_0, θ_{00} be perpendicular unit vectors, and define $\theta = \theta(\theta_0, \theta_{00})$ by* (3.1). *Then there exist uniformly bounded, continuous functions A_1, A_2 such that*

$$(3.4) \qquad \begin{aligned} &\sup\left|A(u,\theta) - \left\{A(u,\theta_0) + \eta A_1(u,\theta_0,\theta_{00}) + \tfrac{1}{2}\eta^2 A_2(u,\theta_0,\theta_{00})\right\}\right| \\ &\quad = o(\eta^2) \end{aligned}$$

as $\eta \to 0$, where the supremum is over $u \ge 0$ and $\theta_0, \theta_{00} \in \Omega$ such that $\theta_0 \perp \theta_{00}$.

Smoothness of Radon transforms is discussed in Chapter 1 of Helgason [9].

Let A, B denote versions of A in cases $a \equiv fG$, $a \equiv f$, respectively. Then versions of A_1 are

$$A_1(u, \theta_0, \theta_{00}) \equiv \int_{\Gamma_{\theta_0}} \{(\theta_0 \cdot x)(fG)_{(\theta_{00})}(x) - (\theta_{00} \cdot x)(fG)_{(\theta_0)}(x)\}\, d\gamma_{\theta_0}(x), \tag{3.5}$$

$$B_1(u, \theta_0, \theta_{00}) \equiv \int_{\Gamma_{\theta_0}} \{(\theta_0 \cdot x) f_{(\theta_{00})}(x) - (\theta_{00} \cdot x) f_{(\theta_0)}(x)\}\, d\gamma_{\theta_0}(x), \tag{3.6}$$

respectively. Put

$$g_1(x, \theta_0, \theta_{00}) \equiv (\theta_{00} \cdot x) g'_{\theta_0}(\theta_0 \cdot x) + (A_1/B) - (AB_1/B^2),$$

where A_1 stands for $A_1(u, \theta_0, \theta_{00})$ with $u = \theta_0 \cdot x$ and so forth. Noting that $g_\theta(u) = A(u, \theta)/B(u, \theta)$, and noting also definition (2.2) of $S(\theta)$, we readily deduce that S_1 in (2.2) is given by

$$S_1(\theta_0, \theta_{00}) = 2\int_{\mathscr{A}} \{g_{\theta_0}(\theta_0 \cdot x) - G(x)\} g_1(x, \theta_0, \theta_{00}) f(x)\, dx.$$

In similar but more complex fashion we may obtain a formula for S_2.

Now we turn to estimated projective approximation, in which we minimize an estimate $\hat{S}(\theta)$ [defined at (2.4)] of

$$S(\theta) \equiv \int_{\mathscr{A}} \{G(x) - g_\theta(\theta \cdot x)\}^2 f(x)\, dx.$$

Our estimates of g_θ are functions $\hat{g}_{\theta\langle k\rangle}$ [defined at (2.3)], each of which is a ratio of two random variables. The ratio of the means is

$$g_\theta(u|h) \equiv A(u, \theta|h)/B(u, \theta|h), \tag{3.7}$$

where

$$A(u, \theta|h) \equiv h^{-1} \int_{\mathbb{R}^p} K\{(u - \theta \cdot x)h^{-1}\} f(x) G(x)\, dx,$$
$$B(u, \theta|h) \equiv h^{-1} \int_{\mathbb{R}^p} K\{(u - \theta \cdot x)h^{-1}\} f(x)\, dx. \tag{3.8}$$

As we shall show, $\hat{S}(\theta)$ is given accurately to first and second order by

$$S(\theta|h) \equiv \int_{\mathscr{A}} \{G(x) - g_\theta(\theta \cdot x|h)\}^2 f(x)\, dx, \tag{3.9}$$

up to terms which do not depend on θ. See Theorem 4.1 in the next section. Therefore (a) there is no pressing need to incorporate an "error about the mean" term to describe the difference between $\hat{S}(\theta)$ and $S(\theta)$, and (b) the key to second-order behaviour of $\hat{\theta}$ lies in a Taylor expansion of $S(\theta|h)$. We develop this expansion next.

Let $\theta, \theta_0 \in \Omega$, with θ_0 fixed and θ converging to θ_0. Parametrize θ in terms of θ_0 and $\theta_{00} \perp \theta_0$, as at (3.1). Taylor-expand $S(\theta|h)$ as we did $S(\theta)$ at (3.2), obtaining

$$S(\theta|h) = S(\theta_0|h) + \eta S_1(\theta_0, \theta_{00}|h) + \tfrac{1}{2}\eta^2 S_2(\theta_0, \theta_{00}|h) + o(\eta^2)$$

as $\eta \to 0$. Next, expand S_1 and S_2, obtaining

$$S_1(\theta_0, \theta_{00}|h) = S_1(\theta_0, \theta_{00}) + h^r S_{11}(\theta_0, \theta_{00}) + o(h^r),$$
$$S_2(\theta_0, \theta_{00}|h) = S_2(\theta_0, \theta_{00}) + o(1),$$

where S_1, S_2 are exactly as in (3.2). Therefore,

$$\begin{aligned} S(\theta|h) = S(\theta_0|h) &+ \eta S_1(\theta_0, \theta_{00}) + \eta h^r S_{11}(\theta_0, \theta_{00}) \\ &+ \tfrac{1}{2}\eta^2 S_2(\theta_0, \theta_{00}) + o(\eta^2 + h^{2r}). \end{aligned} \tag{3.10}$$

If θ_0 gives a (local) minimum of S [not of $S(\cdot|h)$] then $S_1(\theta_0, \theta_{00}) = 0$ and $S_2(\theta_0, \theta_{00}) > 0$ for all $\theta_{00} \perp \theta_0$. The value of η which gives a turning point of the sum of the third and fourth terms on the right-hand side of (3.10) is of course

$$\eta_0 \equiv -h^r S_{11}(\theta_0, \theta_{00})/S_2(\theta_0, \theta_{00}).$$

If η is asymptotic to η_0, and if θ_0 gives a regular minimum of S, then (3.10) reduces to

$$S(\theta|h) = S(\theta_0|h) - \tfrac{1}{2}h^{2r} S_{11}(\theta_0, \theta_{00})^2 S_2(\theta_0, \theta_{00})^{-1} + o(h^{2r}).$$

The second term (a negative number) on the right-hand side of this expansion is minimized by choosing θ_{00} so as to maximize $S_{11}(\theta_0, \theta_{00})^2/S_2(\theta_0, \theta_{00})$. Thus, we are led to the following theorem. (The regularity conditions will be discussed shortly.)

THEOREM 3.3. *Let $r \geq 2$. Assume that the first $r + 1$ directional derivatives of f and G exist and are continuous uniformly in $x \in \mathbb{R}^p$ and in all directions; that $\mathscr{A}$ is a nonempty, open, p-dimensional convex set whose boundary has two continuous derivatives; that f satisfies* (2.8); *and that K satisfies* (2.9). *Let θ_0 give a local minimum of $S(\theta)$ [defined at* (2.2)], *and suppose the minimum is achieved in a regular fashion, in the sense that $S_2(\theta_0, \theta) > 0$ for all $\theta \perp \theta_0$. Let θ_{00} uniquely maximize $S_{11}(\theta_0, \theta)^2/S_2(\theta_0, \theta)$ over $\theta \perp \theta_0$. Then the value $\theta_0(h)$ of θ which minimizes $S(\theta|h)$ [defined at* (3.9)] *over $\theta \in \Omega$, satisfies*

$$\theta_0(h) = \theta_0 - h^r \theta_{00}\{S_{11}(\theta_0, \theta_{00})/S_2(\theta_0, \theta_{00})\} + o(h^r) \tag{3.11}$$

as $h \to 0$.

Theorem 3.3 may be proved via tedious but straightforward calculus. The function $S_{11}(\theta_0, \theta_{00})$ is extremely complex, but it is easy to see that S_{11} explicitly involves $(r + 1)$st derivatives of f and G. Indeed, $S_{11}(\theta_0, \theta_{00}|h)$ involves first derivatives of f and G, and r derivatives of the integrand of this quantity (so, $r + 1$ derivatives in all) are needed to get the term of order h^r in (3.11).

4. Properties of projection pursuit estimators. Let $\hat{S}(\theta)$ and $S(\theta|h)$ be the quantities defined at (2.4) and (3.9), respectively. Given $\theta_0 \in \Omega$, and $\varepsilon > 0$, put

$$\Theta_\varepsilon \equiv \left\{\theta \in \Omega \colon \|\theta - \theta_0\| \le (nh)^{-1/2} n^{\varepsilon}\right\}.$$

Recall from Section 2.2 that we have agreed to take $h \sim \mathrm{const}\, n^{-1/(2r+1)}$ as $n \to \infty$. In this circumstance, $(nh)^{-1/2} \sim$ const. h^r as $n \to \infty$. When interpreting results in the present section it will be helpful to remember that the one-dimensional counterparts of our estimators, whose properties projection pursuit regression is emulating, converge to their limits at rate $(nh)^{-1/2}$ (equivalently, h^r). Quantities which are $o\{(nh)^{-1/2}\}$ or $o(h^r)$, are negligible in comparison.

We assume that the Y_k's are related to the x_k's via the model

$$Y_k = G(x_k) + \varepsilon_k, \qquad 1 \le k \le n, \tag{4.1}$$

where the $\{\varepsilon_k\}$ and $\{x_k\}$ samples are stochastically independent, the x_k's are independent and identically distributed (i.i.d.) with density f, and the ε_k's are i.i.d. with zero mean, variance $\sigma^2 > 0$ and all moments finite. More general models may be treated using similar methods—for example, the distribution of ε_k given x_k may be permitted to depend on x_k.

It is straightforward to prove that $\hat{S}(\theta)$ is uniformly consistent for $S(\theta)$, which means that if $\theta_1, \hat{\theta}$ minimize $S(\theta)$, $\hat{S}(\theta)$, respectively, then $\hat{\theta} \to \theta_1$. From this it follows easily that $\hat{G}_1 \to G_1$. Our task is to describe the *rate* of consistency, and for that we first show that for any given $\theta_0 \in \Omega$, $\hat{S}(\theta)$ is close to $S(\theta|h)$ uniformly in θ values near to θ_0. Notice that for this result we need only r derivatives of f and G; $(r+1)$st derivatives are not required until later in our argument.

THEOREM 4.1. *Let $r \ge 2$, and let θ_0 be any element of Ω. Assume conditions* (2.6) *and* (2.8) *on f and G, condition* (2.9) *on K, and that $\mathscr{A}$ is a nonempty, open, p-dimensional convex set whose boundary has two continuous derivatives. Then there exists a random variable T_n not depending on θ (but depending on θ_0), such that for any $\varepsilon < 1/\{2(2r+1)\}$,*

$$nh \sup_{\theta \in \Theta_\varepsilon} \left|\hat{S}(\theta) - S(\theta|h) - T_n\right| \to 0 \tag{4.2}$$

almost surely.

Now assume an extra continuous derivative of f and G, bringing the total to $r+1$. Choose θ_0 $(= \theta_1)$ to be the value of θ minimizing $S(\theta)$. Take θ_{00} to be that unit vector perpendicular to θ_0, which maximizes $S_{11}(\theta_0, \theta_{00})^2 / S_2(\theta_0, \theta_{00})$. The following result is immediate from (4.2) and the argument just prior to Theorem 3.3.

THEOREM 4.2. *Assume all the conditions and adopt the notation of Theorem 3.3. Let $\hat{\theta}$ be a value of θ which minimizes $\hat{S}(\theta)$ over θ-values satisfying $\|\theta - \theta_0\| \le (nh)^{-1/2} n^{\varepsilon}$, for any fixed $\varepsilon < 1/\{2(2r+1)\}$. Then $\hat{\theta}$ admits the* *expansion given for* $\theta_0(h)$ *in* (3.11),

$$\hat{\theta}_0 = \theta_0 - h^r \theta_{00}\{S_{11}(\theta_0, \theta_{00})/S_2(\theta_0, \theta_{00})\} + o(h^r)$$

almost surely as $n \to \infty$. *Also, if* $\mathscr{E}_n$ *is the event that* $\hat{\theta}$ *is a turning point of* $\hat{S}(\theta)$, *then* $P(\liminf \mathscr{E}_n) = 1$.

[Since we assume that the kernel K is Hölder continuous, then $\hat{S}(\theta)$ is a continuous function of θ, and so $\hat{\theta}$ is well-defined.]

Notice that by Theorem 4.2, $\|\hat{\theta} - \theta_0\| = O(h^r) = O\{(nh)^{-1/2}\}$ almost surely, and so $\hat{\theta}$ is well inside the cone Θ_ε. We shall show in Theorems 4.3 and 4.4 below that for this definition of $\hat{\theta}$, our estimator $\hat{g}_{\hat{\theta}}(\hat{\theta} \cdot x)$ converges to the "target" projection $g_{\theta_0}(\theta_0 \cdot x)$ at rate $(nh)^{-1/2}$. [Recall that $\hat{g}_\theta(u)$ was defined at (2.5).] It is easy to see (by considering the bias term) that if we were to choose our orientation estimator outside the cone, then the rate of convergence would be no better than $O\{(nh)^{-1/2}n^\varepsilon\}$. Therefore our restriction to the cone Θ_ε does not exclude any minima of interest, and is made without loss of generality. However, our argument gives us no information about possible difficulties which multiple minima might cause. That matter is perhaps best explored by simulation.

Let $\theta_0(h)$ denote the value of θ which minimizes $S(\theta|h)$. We know from Theorems 3.3 and 4.2 that $\hat{\theta} - \theta_0(h) = o\{(nh)^{-1/2}\}$ almost surely, and so it stands to reason that the projective approximation based on $\hat{\theta}$ should have performance very similar to that of its counterpart based on $\theta_0(h)$. The next theorem makes this clear.

THEOREM 4.3. *Assume the conditions of Theorem* 3.3, *and define* $\hat{\theta}$ *as in Theorem* 4.2. *Then*

$$(nh)^{1/2} \sup_{x \in \mathscr{A}} \left|\hat{g}_{\hat{\theta}}(\hat{\theta} \cdot x) - \hat{g}_{\theta_0(h)}\{\theta_0(h) \cdot x\}\right| \to 0$$

almost surely.

Therefore the kernel estimator with orientation estimated at $\hat{\theta}$, is asymptotically equivalent to the kernel estimator with nonrandom orientation $\theta_0(h)$. To describe asymptotic properties of the latter estimator, let A, B, A_1 and B_1 be as defined just before and in (3.5) and (3.6), and put

$$\beta_1 \equiv -S_{11}(\theta_0, \theta_{00})S_2(\theta_0, \theta_{00})^{-1}\{A_1(\theta_0 \cdot x, \theta_0, \theta_{00})B(\theta_0 \cdot x, \theta_0)^{-1} - A(\theta_0 \cdot x, \theta_{00})B_1(\theta_0 \cdot x, \theta_0, \theta_{00})B(\theta_0 \cdot x, \theta_0)^{-2}\},$$

$$\beta_2 \equiv -(\theta_{00} \cdot x)S_{11}(\theta_0, \theta_{00})S_2(\theta_0, \theta_{00})^{-1}g'_{\theta_0}(\theta_0 \cdot x)$$

and $\beta_3 \equiv \beta(\theta_0 \cdot x)$, where $\theta_0(= \theta_1)$ minimizes $S(\theta)$, and

$$\beta(u) \equiv (-1)^r(r!)^{-1}\left\{\int_{-\infty}^{\infty} v^r K(v)\, dv\right\} f_{\theta_0}(u)^{-1} \times \left[\int_{\Gamma_{\theta_0}(u)} (D_{\theta_0})^r (fG)(y)\, d\gamma_{\theta_0}(y) - g_{\theta_0}(u) \int_{\Gamma_{\theta_0}(u)} (D_{\theta_0})^r f(y)\, d\gamma_{\theta_0}(y)\right].$$

(The operator D_θ denotes directional differentiation in direction θ.) Recall that the "errors" ε_k in model (4.1) have variance σ^2. Define

$$\tau^2(x) \equiv \left[E\left\{\left(G(X) - g_{\theta_0}(\theta_0 \cdot x)\right)^2 \middle| \theta_0 \cdot X = \theta_0 \cdot x\right\} + \sigma^2\right] f_{\theta_0}(\theta_0 \cdot x)^{-1} \times \int_{-\infty}^{\infty} K^2(v)\, dv.$$

THEOREM 4.4. *Assume the conditions of Theorem* 3.3, *and define* $\hat{\theta}$ *as in Theorem* 4.2. *Then for each* $x \in \mathscr{A}$,

$$(nh)^{1/2}\left\{\hat{g}_{\hat{\theta}_0}\left(\hat{\theta}_0 \cdot x\right) - g_{\theta_0}(\theta_0 \cdot x)\right\} = \tau(x)Z + (nh)^{1/2}h^r(\beta_1 + \beta_2 + \beta_3) + o_p(1)$$

as $n \to \infty$, *where* Z *is asymptotically normal* $N(0,1)$.

Terms β_1 and β_2 derive from the bias of the orientation estimate, $\hat{\theta}$. Thus, they contain derivatives of f and G of order $r + 1$. On the other hand, β_3 is the usual expression for bias of a univariate nonparametric regression estimator; see Prakasa Rao [13], page 239ff. Therefore β_3 contains only derivatives of order r or less.

The procedures described in Theorems 4.1–4.4 all use nonrandom bandwidths. Our arguments could be extended to encompass random, data-driven bandwidths; the only change required to our proofs would be more frequent use of the "continuity argument," described early in the proof of Theorem 4.1 in Section 5.

The procedure studied above is a good approximation to techniques actually used in practice [5, 8]. One variant of it is a two-stage algorithm [7], which may be described as follows. Recall that throughout our work we have taken $h \sim$ const $n^{-1/(2r+1)}$, since this size of window is optimal in univariate problems. As we have shown, this gives orientation estimates with error $O(n^{-r/(2r+1)})$, which is typical of nonparametric problems. However, it is possible for orientation estimates to achieve convergence rates of $O_p(n^{-1/2})$, under nonparametric assumptions. For example, using a nonnegative kernel ($r = 2$) and a window size between $n^{-1/3}$ and $n^{-1/4}$ (instead of $n^{-1/5}$), it is possible to estimate orientation with error $O_p(n^{-1/2})$ under the assumption that second derivatives of f and G satisfy a Lipschitz condition of order $\frac{1}{2} + \varepsilon$ for some $\varepsilon > 0$. Of course, this window is suboptimal as far as estimation of ridge functions goes, and so we should use a second stage to estimate ridges. Go back and reconstruct the estimator $\hat{g}_{\hat{\theta}}(\hat{\theta} \cdot x)$, using the *new* orientation estimate $\hat{\theta}$ but the *old* window size h ($\sim$ const. $n^{-1/(2r+1)}$). In view of the exceptional accuracy of our new $\hat{\theta}$, the new estimator $\hat{g}_{\hat{\theta}}(\hat{\theta} \cdot x)$ does not include a bias contribution from the error between $\hat{\theta}$ and θ_0 ($= \theta_1$). In fact, Theorem 4.4 holds in the form

$$(nh)^{1/2}\left\{\hat{g}_{\hat{\theta}}(\hat{\theta} \cdot x) - g_{\theta_0}(\theta_0 \cdot x)\right\} = \tau(x)Z + (nh)^{1/2}h^r\beta_3 + o_p(1).$$

Using undersmoothed kernel estimates of ridge functions to estimate orientation with $n^{-1/2}$ accuracy is analogous to estimating a distribution function with $n^{-1/2}$ accuracy by integrating an undersmoothed density estimate. It is well-known that the latter is possible. For values of $r \geq 3$, the two-stage method does yield one-dimensional convergence rates of the projective approximation, under the assumption of only r derivatives.

A reasonable practical objection to the two-stage algorithm is that the first stage magnifies numerical difficulties in finding the global minimum of $\hat{S}(\theta)$. The functional $\hat{S}$ usually has a multitude of local minima, and these become more pronounced and even more numerous when the amount of smoothing is reduced. Note too that the two-stage procedure does not actually improve the rate of convergence of the estimated projective approximation.

5. Outline of proofs for Section 4. The following notation will be used throughout. Summation over $1 \le k \le n$ such that $x_k \in \mathscr{A}$ will be denoted by Σ'_k; X will be a p-variate random variable with density f; $f_\theta(u|h)$ will be the expected value of the estimator $\hat{f}_\theta(u)$ [identical to the function $B(u, \theta|h)$ defined at (3.8)]; $g_\theta(u|h)$ will be the function defined at (3.7); and Θ_ε will be the set $\{\theta \in \Omega\colon \|\theta - \theta_0\| \le (nh)^{-1/2}n^\varepsilon\}$, for arbitrary but fixed $\theta_0 \in \Theta$.

Proof of Theorem 4.1. The key to this proof is the "continuity argument"; see Stone [16] for an example of its use elsewhere. The argument runs as follows. Suppose that a certain stochastic process $Z_n(y)$, $y \in S \subseteq \mathbb{R}^q$ and $q \ge 1$, may be shown to have the properties

$$\text{for all } \varepsilon, \lambda > 0, \quad \sup_{y \in S} P\{|Z_n(y)| > \varepsilon\} = O(n^{-\lambda}),$$

$$\text{for each } \lambda_1 > 0 \text{ there exist } C, \lambda_2 > 0 \text{ such that}$$

$$E\Bigl\{\sup_{y_1,\, y_2 \in \mathscr{S} \text{ s.t. } \|y_1 - y_2\| \le n^{-\lambda_2}} |Z_n(y_1) - Z_n(y_2)|\Bigr\} \le Cn^{-\lambda_1}.$$

Then if $\mathscr{S}$ is a bounded set,

$$\text{for all } \varepsilon, \lambda > 0, \quad P\Bigl\{\sup_{y \in \mathscr{S}} |Z_n(y)| > \varepsilon\Bigr\} = O(n^{-\lambda}). \tag{5.1}$$

In view of the Borel–Cantelli lemma, result (5.1) implies

$$\sup_{y \in \mathscr{S}} |Z_n(y)| \to 0$$

almost surely.

To illustrate use of this argument, we employ it to prove the following lemma. Define

$$\tilde{g}_{\theta\langle k\rangle}(u) \equiv \Bigl[\{(n-1)h\}^{-1} \sum_{l \notin k} G(x_l) K\{(u - \theta \cdot x_l)h^{-1}\}\Bigr] \Big/ \hat{f}_{\theta\langle k\rangle}(u).$$

Let sup* denote supremum over $\theta \in \Omega$, $x \in \mathscr{A}$ and $1 \le k \le n$.

Lemma 5.1. *Under the stated conditions, and for each $\xi > 0$ and $\lambda > 0$,*

$$P\bigl\{\sup^* |\hat{f}_{\theta\langle k\rangle}(\theta \cdot x) - f_\theta(\theta \cdot x)| > (nh)^{-1/2} n^\xi\bigr\} = O(n^{-\lambda}),$$

$$P\bigl\{\sup^* |\hat{g}_{\theta\langle k\rangle}(\theta \cdot x) - g_\theta(\theta \cdot x)| > (nh)^{-1/2} n^\xi\bigr\} = O(n^{-\lambda}).$$

These results continue to hold if $\hat{g}_{\theta\langle k\rangle}(\cdot)$, $f_\theta(\cdot)$ *and* $g_\theta(\cdot)$ *are replaced by* $\tilde{g}_{\theta\langle k\rangle}(\cdot)$, $f_\theta(\cdot\,|h)$ *and* $g_\theta(\cdot\,|h)$.

PROOF. Using the continuity argument, it suffices to show that

$$\sup{}^*P\Big\{|\hat{l}_{\theta\langle k\rangle}(\theta\cdot x) - l_\theta(\theta\cdot x|h)| > (nh)^{-1/2}n^{\xi}\Big\} = O(n^{-\lambda}),$$

where $\hat{l}_{\theta\langle k\rangle}$ denotes $\hat{f}_{\theta\langle k\rangle}$, $\hat{e}_{\theta\langle k\rangle}$ or $\hat{d}_{\theta\langle k\rangle}$; $l_\theta(\cdot\,|h)$ denotes $f_\theta(\cdot\,|h)$, $e_\theta(\cdot\,|h)$ or 0 (respectively); and

$$\hat{e}_{\theta\langle k\rangle}(u) \equiv \{(n-1)h\}^{-1}\sum_{l\neq k} G(x_l)K\{(u-\theta\cdot x_l)h^{-1}\},$$

$$\hat{d}_{\theta\langle k\rangle}(u) \equiv \{(n-1)h\}^{-1}\sum_{l\neq k} \varepsilon_l K\{(u-\theta\cdot x_l)h^{-1}\}$$

and $e_\theta(u|h) \equiv f_\theta(u|h)g_\theta(u|h)$. The proofs proceed very easily by Bernstein's inequality in the case of f and e, and by Rosenthal's inequality [8, page 23] in the case of d. □

Next we introduce a decomposition of $\hat{S}(\theta)$:

$$\hat{S}(\theta) = \hat{S}_{[1]}(\theta) - 2\hat{S}_{[2]}(\theta) + n^{-1}\sum_k{}'\{\varepsilon_k^2 + 2\varepsilon_k G(x_k)\},$$

where

$$\hat{S}_{[1]}(\theta) \equiv n^{-1}\sum_k{}'\{G(x_k) - \hat{g}_{\theta\langle k\rangle}(\theta\cdot x_k)\}^2,$$

$$\hat{S}_{[2]}(\theta) \equiv n^{-1}\sum_k{}'\varepsilon_k\hat{g}_{\theta\langle k\rangle}(\theta\cdot x_k).$$

Expand $\hat{S}_{[1]}$ as

$$\hat{S}_{[1]}(\theta) = \hat{S}_{[3]}(\theta) - 2\hat{S}_{[4]}(\theta) + n^{-1}\sum_k{}'G(x_k)^2,$$

where

$$\hat{S}_{[3]}(\theta) \equiv n^{-1}\sum_k{}'\{\hat{g}_{\theta\langle k\rangle}(\theta\cdot x_k)\}^2,$$

$$\hat{S}_{[4]}(\theta) \equiv n^{-1}\sum_k{}'G(x_k)\hat{g}_{\theta\langle k\rangle}(\theta\cdot x_k);$$

and expand $\hat{S}_{[4]}$ as

$$\hat{S}_{[4]}(\theta) = \hat{S}_{[5]}(\theta) + \hat{S}_{[6]}(\theta),$$

where

$$\hat{S}_{[5]}(\theta) \equiv n^{-1}\sum_k G(x_k)\{\hat{g}_{\theta\langle k\rangle}(\theta\cdot x_k) - \tilde{g}_{\theta\langle k\rangle}(\theta\cdot x_k)\},$$

$$\hat{S}_{[6]}(\theta) \equiv n^{-1}\sum_k{}'G(x_k)\tilde{g}_{\theta\langle k\rangle}(\theta\cdot x_k).$$

Put

$$\hat{S}_{[7]}(\theta) \equiv n^{-1} \sum_k{}' \{g_\theta(\theta \cdot x_k | h)\}^2,$$

$$\hat{S}_{[8]}(\theta) \equiv n^{-1} \sum_k{}' G(x_k) g_\theta(\theta \cdot x_k | h).$$

By repeated application of the continuity argument, and use of Lemma 5.1, it may be shown that there exist random variables $T_{n1}, \dots, T_{n4}$ not depending on θ such that, under the stated conditions,

$$nh \sup_{\theta \in \Theta_\varepsilon} \left| \hat{S}_{[2]}(\theta) - T_{n1} \right| \to 0,$$

$$nh \sup_{\theta \in \Theta_\varepsilon} \left| S_{[5]}(\theta) - T_{n2} \right| \to 0,$$

$$nh \sup_{\theta \in \Theta_\varepsilon} \left| \hat{S}_{[3]}(\theta) - \hat{S}_{[7]}(\theta) - T_{n3} \right| \to 0,$$

$$nh \sup_{\theta \in \Theta_\varepsilon} \left| \hat{S}_{[6]}(\theta) - \hat{S}_{[8]}(\theta) - T_{n4} \right| \to 0$$

almost surely. It follows that

$$(5.2) \qquad nh \sup_{\theta \in \Theta_\varepsilon} \left| \hat{S}(\theta) - \left\{ \hat{S}_{[7]}(\theta) - 2\hat{S}_{[8]}(\theta) \right\} - T_{n5} \right| \to 0$$

almost surely, where T_{n5} does not depend on θ. Further algebra shows that

$$(5.3) \qquad nh \sup_{\theta \in \Theta_\varepsilon} \left| \hat{S}_{[7]}(\theta) - 2\hat{S}_{[8]}(\theta) - S(\theta | h) - T_{n6} \right| \to 0$$

almost surely, where T_{n6} does not depend on θ. Theorem 4.1 follows from (5.2) and (5.3). □

PROOF OF THEOREM 4.3. Define

$$\hat{A}(u, \theta) \equiv (nh)^{-1} \sum_{k=1}^{n} G(x_k) K\{(u - \theta \cdot x_k) h^{-1}\},$$

$$\hat{B}(u, \theta) \equiv (nh)^{-1} \sum_{k=1}^{n} K\{(u - \theta \cdot x_k) h^{-1}\} \qquad [= \hat{f}'_\theta(u)],$$

$$\hat{D}(u, \theta) \equiv (nh)^{-1} \sum_{k=1}^{n} \varepsilon_k K\{(u - \theta \cdot x_k) h^{-1}\},$$

$A(u, \theta | h) \equiv E\{\hat{A}(u, \theta)\}$ and $B(u, \theta | h) \equiv E\{\hat{B}(u, \theta)\}$. The argument leading to Lemma 5.1 may be employed to show that for each $\xi > 0$,

$$\sup_{\theta \in \Omega,\, x \in \mathscr{A}} \{ |\hat{A}(\theta \cdot x, \theta) - A(\theta \cdot x, \theta | h)| + |\hat{B}(\theta \cdot x, \theta) - B(\theta \cdot x, \theta | h)| + |\hat{D}(\theta \cdot x, \theta)| \}$$

$$= O\{(nh)^{-1/2} n^\xi\}$$

almost surely. Hence

$$\begin{aligned}\hat{g}_\theta(u) &= \{\hat{A}(u,\theta) + \hat{D}(u,\theta)\}/\hat{B}(u,\theta)\\ &= \{A(u,\theta|h)/B(u,\theta|h)\} + \{\hat{A}(u,\theta) - A(u,\theta|h) + \hat{D}(u,\theta)\}/B(u,\theta|h)\\ &\qquad - \{\hat{B}(u,\theta) - B(u,\theta|h)\}A(u,\theta|h)/B(u,\theta|h)^2 + o\{(nh)^{-1/2}\}\end{aligned}$$

almost surely, uniformly in real numbers u of the form $\theta \cdot x$ for $\theta \in \Omega$ and $x \in \mathscr{A}$.

Let C denote either A or B. Using the fact that $\hat{\theta}_0$ and $\theta_0(h)$ may both be expressed as $\theta_0 + \text{const } h^r\theta_{00} + o(h^r)$, it may be shown that

$$C(\hat{\theta}_0 \cdot x, \hat{\theta}_0|h) - C\{\theta_0(h) \cdot x, \theta_0(h)|h\} = o\{(nh)^{-1/2}\},$$

uniformly in $x \in \mathscr{A}$. Therefore Theorem 4.3 will be proved if we show that, with $\hat{F}$ denoting any one of $\hat{A} - A$, $\hat{B} - B$, $\hat{D}$, we have

$$\sup_{x \in \mathscr{A}} \left|\hat{F}(\hat{\theta}_0 \cdot x, \hat{\theta}_0) - \hat{F}\{\theta_0(h) \cdot x, \theta_0(h)\}\right| = o\{(nh)^{-1/2}\}$$

almost surely. This may be accomplished via the continuity argument, discussed early in the proof of Theorem 4.1. □

PROOF OF THEOREM 4.4. It is readily shown that for any nonrandom $\theta = \theta(n)$ converging to θ_0 [such as $\theta = \theta_0(h)$], $\hat{g}_\theta(\theta \cdot x) - g_\theta(\theta \cdot x|h)$ has the same asymptotic distribution as $\hat{g}_{\theta_0}(\theta_0 \cdot x) - g_{\theta_0}(\theta_0 \cdot x|h)$, this being normal $N\{0, (nh)^{-1}\tau^2(x)\}$ and being derivable via classical arguments associated with univariate nonparametric regression [13], page 239ff. It remains only to determine the asymptotic bias, and that may be accomplished after some tedious algebra. □

Acknowledgments. This is a substantially abbreviated version of an earlier manuscript. Detailed, constructive criticism by two referees and the Associate Editor encouraged me to present the proofs only in outline, and to clarify the discussion in Sections 1 to 4. I am grateful for their care and patience.

REFERENCES

[1] DIACONIS, P. and FREEDMAN, D. (1984). Asymptotics of graphical projection pursuit. *Ann. Statist.* **12** 793–815.

[2] DIACONIS, P. and SHASHAHANI, M. (1984). On nonlinear functions of linear combinations. *SIAM J. Sci. Statist. Comput.* **5** 175–191.

[3] DONOHO, D. L. and JOHNSTONE, I. M. (1989). Projection-based approximation and a duality with kernel methods. *Ann. Statist.* **17** 58–106.

[4] FILL, J. and JOHNSTONE, I. (1984). On projection pursuit measures of multivariate location and dispersion. *Ann. Statist.* **12** 127–141.

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

[6] FRIEDMAN, J. H. and TUKEY, J. W. (1974). A projection pursuit algorithm for exploratory data analysis. *IEEE Trans. Comput.* **C-23** 881–889.

[7] HALL, P. (1988). Estimating the direction in which a data set is most interesting. *Probab. Theory Related Fields* **80** 51–78.

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

[9] HELGASON, S. (1980). *The Radon Transform*. Birkhäuser, Boston.

[10] HUBER, P. J. (1985). Projection pursuit (with discussion). *Ann. Statist.* **13** 435–525.

[11] KRUSKAL, J. B. (1969). Toward a practical method which helps uncover the structure of a set of multivariate observations by finding a linear transformation which optimizes a new "index of condensation." In *Statistical Computation* (R. C. Milton and J. A. Nelder, eds.) 427–440. Academic, New York.

[12] KRUSKAL, J. B. (1972). Linear transformation of multivariate data to reveal clustering. In *Multidimensional Scaling: Theory and Application in the Behavioural Sciences*. I. *Theory* (R. N. Shepard, A. K. Romney and S. B. Nerlove, eds.) 181–191. Seminar, New York.

[13] PRAKASA RAO, B. L. S. (1983). *Nonparametric Functional Estimation*. Academic, New York.

[14] STONE, C. J. (1980). Optimal rates of convergence for nonparametric estimators. *Ann. Statist.* **8** 1348–1360.

[15] STONE, C. J. (1982). Optimal global rates of convergence for nonparametric regression. *Ann. Statist.* **10** 1040–1053.

[16] STONE, C. J. (1984). An asymptotically optimal window selection rule for kernel density estimates. *Ann. Statist.* **12** 1285–1297.

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

[18] SWITZER, P. (1970). Numerical classification. In *Geostatistics* (D. F. Merriam, ed.) 31–43. Plenum, New York.

[19] SWITZER, P. and WRIGHT, R. M. (1971). Numerical classification applied to certain Jamaican eocene nummulitids. *Math. Geol.* **3** 297–311.

DEPARTMENT OF STATISTICS
FACULTY OF ECONOMICS AND COMMERCE
AUSTRALIAN NATIONAL UNIVERSITY
GPO BOX 4
CANBERRA, ACT 2601
AUSTRALIA