The Annals of Statistics
1985, Vol. 13, No. 2, 435–475

INVITED PAPER

PROJECTION PURSUIT[1]

BY PETER J. HUBER

Harvard University

Projection pursuit is concerned with "interesting" projections of high dimensional data sets, with finding such projections by machine, and with using them for nonparametric fitting and other data-analytic purposes. This survey attempts to put the fascinating problems and ramifications of projection pursuit—which range from principal components, multidimensional scaling, factor analysis, nonparametric regression, density estimation and deconvolution of time series to computer tomography and problems in pure mathematics—into a coherent perspective

CONTENTS

Received May 1983; revised December 1984.

[1] This work was facilitated in part by Office of Naval Research Contract N0014-79-C-0512 and National Science Foundation Grants MCS-79-08685, MCS-82-00914.

AMS 1980 *subject classifications.* Primary 62H99.

Key words and phrases. Projection pursuit, multivariate data analysis, principal components, computer tomography, robust multivariate methods, minimum entropy.

I. Generalities on projection pursuit.

1. Introduction. Projection pursuit (PP) techniques were originally proposed and experimented with by Kruskal (1969, 1972). Related ideas occur in Switzer (1970) and Switzer and Wright (1971). The first successful implementation is due to Friedman and Tukey (1974), who also coined the catchy name.

The original purpose of PP was to machine-pick "interesting" low-dimensional projections of a high-dimensional point cloud by numerically maximizing a certain objective function or *projection index.* In its Friedman-Tukey form, this index was the product of a robust measure of scale (trimmed standard deviation) with a measure of clumpiness (a weighted count of the number of close pairs).

After a dormant stage of several years, Friedman and Stuetzle extended the idea behind PP and added projection pursuit regression (PPR: Friedman and Stuetzle, 1981), projection pursuit classification (PPC: Friedman and Stuetzle, 1980) and projection pursuit density estimation (PPDE: Friedman, Stuetzle and Schroeder, 1984).

The most exciting feature of PP is that it is one of the very few multivariate methods able to bypass the "curse of dimensionality" caused by the fact that high-dimensional space is mostly empty. For example, assume that a large number of points is distributed uniformly in the 10-dimensional unit ball. Then the radius of a ball containing 5% of the points is $(0.05)^{0.1} = 0.74$. This implies that kernel smoothers and similar methods will not be able to pick up small features, unless the sample size is gigantic. PP avoids this problem by working in low-dimensional linear projections. The price to be paid is, of course, that PP is poorly suited to deal with highly nonlinear structures (but kernel smoothers are not a viable alternative either).

In addition, the more interesting PP methods are able to ignore irrelevant (i.e. noisy and information-poor) variables. This is a distinct advantage over methods based on interpoint distances like minimal spanning trees, multidimensional scaling and most clustering techniques. These latter methods also avoid (at least to a certain extent) the curse of dimensionality, but all of them can be derailed by noninformative variables.

Many of the methods of classical multivariate analysis turn out to be special cases of PP. Examples are principal component and discriminant analysis, and the quartimax and oblimax methods in factor analysis.

PP emerges as the most powerful method yet invented to lift one-dimensional statistical techniques to higher dimensions. To give a simple example: if we take the 2-sample t-statistic as our projection index, then PP searches for the best discriminating hyperplane in the classical, Fisherian sense. If we replace the t-statistic by a robust 2-sample test statistic, we obtain a robust version of discriminant analysis.

The only known affine equivariant estimators of multivariate location and scatter with high breakdown point (i.e. approaching ½ in large samples) are based on PP ideas (Stahel, 1981; Donoho, 1982).

PP methods have one serious drawback: their high demand on computer time. We may say that PP became both needed and feasible through the advent of inexpensive high-speed computing, and this may explain its simultaneous, multiple invention around 1970.

Among the more remote ramifications of PP, one should mention computer tomography (CT): both PP and CT are concerned with the efficient reconstruction of higher dimensional structure from lower dimensional projections, and there are interesting possibilities for cross-fertilization.

Furthermore, there is an amusing connection between PPR and Hilbert's 13th problem, whose solution, the celebrated Kolmogorov-Arnold-Kahane theorem (see Vitushkin, 1978, page 27 f.), becomes relevant if one should want to do PPR with nonlinearly transformed variables.

2. Why projection pursuit? If we want to check a low-dimensional data set for the possible presence of some unspecified, unanticipated structure, then, as we all know, the most effective approach is to draw pictures—histograms in one or two dimensions, scatterplots in two or three dimensions.

While it is possible to encode several more dimensions into pictures by using time (motion), color, and various symbols (glyphs), the human perceptual system is not really prepared to deal with more than three continuous dimensions simultaneously.

The trouble lies with the dimensionality, not with the method of encoding. This is shown by the empirical fact that we find it easy to transpose two space and one time dimension into three space dimensions (as in kinematic graphics), and we can also perform such a transposition, although less easily, with two space and one color dimension (as with cartographic maps).

Procedures for dealing directly with four or more dimensions, coding the fourth dimension by color or by time (slicing, masking, cf. Tukey and Tukey, 1981), seem to work well only if the data are either clustered in such a way that the superb power of color for encoding categorical variables bears fruit, or if the data set is essentially three-dimensional, so that slicing the data with regard to one variable cuts two-dimensional sections through the space spanned by the other three variables.

Thus, if we want to put the human ability for essentially instantaneous pattern discovery to good use with four and higher dimensional data, we should first reduce the dimensionality. For obvious reasons, we shall usually want to look first at the projections onto the spaces spanned by one, two or three of the coordinates. But it seems that few people muster the patience and concentration for a careful scrutiny of all $\binom{d}{2}$ and $\binom{d}{3}$ scatterplots of pairs or triples of variables, if the dimension d is larger than about 10 or 7, respectively.

If we want to consider arbitrary one- to three-dimensional linear projections, the problem gets even worse. When we rotate a point cloud in three-space in order to view its two-dimensional projections, then the "interesting" features

often are recognizable only over a relatively narrow range of "squint angle" (Tukey and Tukey, 1981, page 215). In our experience, a rotation angle of about 10° to 20° may be typical, but for example the infamous planes produced by the random number generator RANDU are visible only in a range narrower than 5° (see Figure 2.1). This generator has the property that any three consecutively generated uniform pseudo-random numbers satisfy $x_{n+2} - 6x_{n+1} + 9x_n \equiv 0 (\text{mod } 1)$, and thus the triplets (x_n, x_{n+1}, x_{n+2}) all lie on 15 parallel planes through the unit cube.

A crude order-of-magnitude estimate based on a squint angle of 10° suggests that we need to look at about 10^{d-1} one-dimensional projections of d-dimensional data, and at something like 10^{2d-4} two-dimensional projections. The fastest way for doing so is to rotate smoothly from one projection to the next (this is less tiring and quicker, since we do not have to reorient ourselves in each new projection), and then we may inspect about one such projection per second. Thus, a reasonably complete visual search in four dimensions (a "Grand Tour" in the sense of Dan Asimov) takes about three hours. It is evident that an exhaustive visual search is out of the question if d exceeds 4. Unless we are willing to rely on happy serendipity, we need an automated procedure that ferrets out projections likely to be of interest to the data analyst. This raises the problem of how to characterize "interestingness" in a numerical fashion.

Unfortunately, the squint angle argument applies also to machine search; we cannot hope that automated exhaustive search will carry us more than two or three dimensions beyond the limits set by human endurance.

Perhaps, more modestly, we should regard PP as a method to increase the likelihood of finding interesting projections.

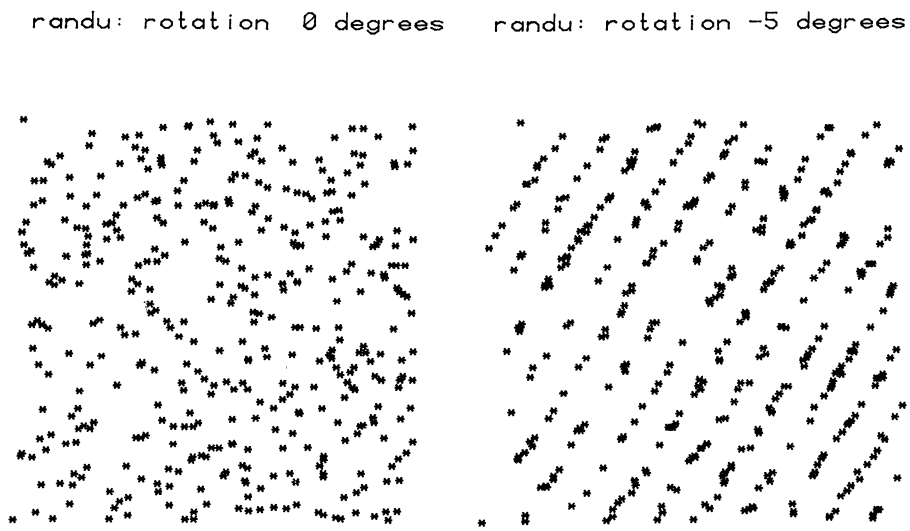

FIG. 2.1.

3. Some concepts and definitions. We observe a sample, but we really are concerned with elucidating an underlying structure. Thus, it is conceptually convenient to separate PP into an "abstract" version which operates on d-dimensional probability distributions (mostly densities) and a "practical" version that is applied to samples (i.e. empirical measures or "point clouds"). The two versions might be identical, but often the abstract version will work on smooth distributions only; so, in order to translate it into a practical one we must insert a suitable smoother at an appropriate place. For notational convenience, we shall use random variable terminology. The letter X shall be used indiscriminately for a point cloud, that is, an n-tuple of points $(x_1, \cdots, x_n)$ in $\mathbb{R}^d$, or for a random variable with values in $\mathbb{R}^d$. By ave$\{X\}$ we shall equally, indiscriminately denote either the sample mean $(1/n) \sum x_i$ or the expectation $E(X)$. Initially, we shall be concerned exclusively with the abstract version and shall postpone sampling questions.

A *linear projection* from $\mathbb{R}^d$ to $\mathbb{R}^k$ is any linear map A, or $k \times d$ matrix of rank k:

$$Z = AX, \quad X \in \mathbb{R}^d, \quad Z \in \mathbb{R}^k.$$

We speak of an orthogonal projection if the row vectors of A are orthogonal to each other and have length 1.

If X is a d-dimensional random variable with distribution F, then $Z = AX$ is a k-dimensional random variable with distribution F_A. If $k = 1$, A reduces to a row vector a^T, and we use lower case letters F_a, etc.

In passing, we note that any d-dimensional distribution is uniquely characterized by its one-dimensional projections F_a. This follows trivially from the fact that F is uniquely determined by its characteristic function ψ and that the characteristic function ψ_a of the one-dimensional projection F_a in direction a equals the section of ψ along the same direction:

$$\psi_a(t) = E(e^{ita^TX}) = \psi(ta). \tag{3.1}$$

By definition, PP searches for a projection A maximizing (or minimizing) a certain objective function or *projection index* $Q(F_A)$. We are specifically interested not only in absolute, but also in local, extrema. While Q is a functional on the space of distributions on $\mathbb{R}^k$, we find it more convenient also here to use random variable terminology and, by abuse of notation, to write $Q(X)$ and $Q(AX)$ instead of $Q(F)$ and $Q(F_A)$. Primarily, we shall be concerned with one-dimensional projections, and for obvious representational reasons we shall rarely want to go beyond three-dimensional projections.

II. Projection pursuit applied to point clouds—abstract version.

4. A classification of projection indices. We single out a few classes of objective functions according to their invariance properties. For simplicity, we consider only one-dimensional orthogonal projections, but the ideas generalize.

In the following, Z is a real random variable, while s, t are (nonrandom) real

numbers. We distinguish three classes of objective functions Q:

CLASS I. Location-scale equivariance:

$$Q_{\mathrm{I}}(sZ + t) = sQ_{\mathrm{I}}(Z) + t.$$

CLASS II. Location invariance, scale equivariance:

$$Q_{\mathrm{II}}(sZ + t) = |s| Q_{\mathrm{II}}(Z).$$

CLASS III. Affine invariance:

$$Q_{\mathrm{III}}(sZ + t) = Q_{\mathrm{III}}(Z), \quad s \neq 0.$$

We note that the absolute difference of two Class I functionals is a Class II functional:

$$|Q'_{\mathrm{I}}(Z) - Q''_{\mathrm{I}}(Z)| = Q_{\mathrm{II}}(Z),$$

and the quotient of two Class II functionals is of Class III:

$$Q'_{\mathrm{II}}(Z)/Q''_{\mathrm{II}}(Z) = Q_{\mathrm{III}}(Z).$$

Let X be a d-dimensional random variable or point cloud. We note that both the mean vector $\mu = \mathrm{ave}\{X\}$ and the principal components, i.e. the eigenvalue/eigenvector representation of the covariance matrix $\Sigma = \mathrm{ave}\{(X - \mu)(X - \mu)^T\}$, can be captured by PP methods as follows.

EXAMPLE 4.1. Let $Q = \mathrm{ave}$. This is Class I functional; $Q(a^TX) = \mathrm{ave}\{a^TX\}$ with $\|a\| = 1$ is maximized by $a_0 = \mu/\|\mu\|$, and the value at the maximum is $Q(a_0^TX) = \|\mu\|$. Hence we may define the d-dimensional mean μ via PP as $a_0Q(a_0^TX)$.

EXAMPLE 4.2. Let Q be the standard deviation, that is,

$$Q(a^TX) = [\mathrm{ave}\{[a^t(X - \mu)]^2\}]^{1/2},$$

with $\|a\| = 1$. This Q is a Class II functional; the maximum value of this objective function is the largest singular value of X (i.e. the square root of the largest eigenvalue of Σ), and it is reached at any eigenvector belonging to this eigenvalue. The other eigenvalues and eigenvectors can be found successively by restricting a to the orthogonal component of the space spanned by the previously found eigenvectors.

Class I functionals are one-dimensional location estimators, and PP with a Class I functional Q_{I} will yield a kind of d-dimensional location estimator in a manner analogous to Example 4.1. We said "kind of" because the resulting estimator in general is neither uniquely defined nor location equivariant. In somewhat more detail this works as follows. Let X be given. Assume that $a_0 \in \mathbb{R}^d$, with $\|a_0\| = 1$, maximizes $Q_{\mathrm{I}}(a^TX)$. Put $T(X) = a_0Q_{\mathrm{I}}(a_0^TX)$.

PROPOSITION 4.3. *The functional T is uniquely defined and location equivariant for the translation family generated by X:*

$$T(X + t) = T(X) + t \quad \text{for all} \quad t \in \mathbb{R}^d,$$

iff there is a $\mu \in \mathbb{R}^d$ such that

$$Q_I(a^T(X - \mu)) = 0 \quad \text{for all} \quad a \in \mathbb{R}^d, \tag{4.1}$$

and then $T(X) = \mu$. Condition (4.1) holds in particular if X is centro-symmetric about μ (i.e. if $X - \mu$ and $-(X - \mu)$ have the same distribution).

PROOF. If X is centro-symmetric about μ, then $Z = a^T(X - \mu)$ is symmetric about 0, and

$$Q_I(-Z) = -Q_I(Z) = Q_I(Z) = 0,$$

which establishes the last assertion of the proposition.

Condition (4.1) is sufficient; if it holds, then

$$\begin{aligned} Q_I(a^T(X + t)) &= Q_I(a^T(X - \mu) + a^T(\mu + t)) \\ &= Q_I(a^T(X - \mu)) + a^T(\mu + t) \\ &= a^T(\mu + t) \end{aligned}$$

which is maximized for $a = a_t = (\mu + t)/\|\mu + t\|$. Then $Q_I(a_t^T(X + t)) = \|\mu + t\|$, and it follows from $T(X + t) = a_t Q_I(a_t^T(X + t)) = \mu + t$ that T is translation equivariant.

Conversely, if T is translation equivariant, put $\mu = T(X)$. Take an arbitrary, but fixed value $t \in \mathbb{R}^d$ and let $a_t = T(X + t)/\|T(X + t)\| = (\mu + t)/\|\mu + t\|$. Then

$$\sup_a Q_I(a^T(X + t)) = \|T(X + t)\| = \|T(X) + t\| = \|\mu + t\|.$$

On the other hand,

$$\begin{aligned} \sup_a Q_I(a^T(X + t)) &= Q_I(a_t^T(X + t)) \\ &= Q_I(a_t^T(X - \mu) + a_t^T(\mu + t)) \\ &= Q_I(a_t^T(X - \mu)) + a_t^T(\mu + t) \\ &= Q_I(a_t^T(X - \mu)) + \|\mu + t\|. \end{aligned}$$

Hence $Q_I(a_t^T(X - \mu)) = 0$. Since t was arbitrary, a_t can be any arbitrary unit vector, and it follows that (4.1) holds.

Clearly, if (4.1) holds, then it also follows that T is uniquely defined on the translates of X, namely $T(X + t) = \mu + t$. □

Actually, Q_I = ave is the *only* location functional leading to a translation equivariant estimator in dimensions $d > 1$. This was proved by Critchlow (1981,

unpublished) and independently by Fill and Johnstone (1984). Similarly, Q_{II} = {standard deviation} is the only scale functional leading to an affine equivariant estimator of a dispersion matrix, see Fill and Johnstone (1984).

5. What is an "interesting" projection? We cannot expect universal agreement on what constitutes an "interesting" projection. A projection in which the data separate into distinct, meaningful clusters would certainly be interesting. But there are also interesting features that are not of the distinct cluster type (e.g. an edge, or jump of density, at the boundary of some region). Rather than trying to identify the kind of features we might regard as potentially interesting, we should perhaps better begin by trying to understand why people have had some degree of success with certain specified techniques.

5.1 *Principal components and other Class II approaches.* The prevalent classical approach is to reduce dimensions by the method of principal components: calculate the eigenvalues and eigenvectors of the covariance or correlation matrix, and project the data orthogonally into the space spanned by the eigenvectors belonging to the largest eigenvalues. Often, these projections show interesting structure. Why? There seem to be at least two, loosely related reasons.

First, if a population is an aggregate of several clusters, then these can become individually visible only if the separation between clusters is larger than the internal scatter of the clusters. Thus, if there are only a few clusters, the leading principal axes will tend to pick projections with good separations. Of course, principal components can go astray, either if there are too many isotropically distributed clusters (compare the Friedman-Tukey, 1974, example with clusters at the corners of a regular simplex), or if there are meaningless variables with a high noise level.

The second reason is more germane to principal component analysis performed on correlation matrices. Assume that we have an intrinsic structure describable by a few (unobservable) variables, and that we observe many, possibly differently scaled (linear) functions of these variables, with independent random noise added. Then principal component analysis tends to act as a variation reducing technique (not unlike the sample mean), relegating most of the random noise to the trailing components, and collecting the systematic structure into the leading ones.

Principal components are quite sensitive to outliers (see, e.g. Devlin, Gnanadesikan and Kettenring, 1981), and while sometimes the outliers are part of the structure to be described, one sometimes would prefer to set them aside. This might be achieved by using the PP version of principal components (see Example 4.2) with a robust Class II functional as projection index. See Chen and Li (1981) and Li and Chen (1981).

5.2 *Class III approaches.* When drawing scatterplots and other graphs, we usually locate and scale the picture so that it nicely fills the available space (see for example the explicit recommendations of Cleveland and McGill, 1984). This may indicate that visual interestingness is an affine invariant notion. The argument does not extend to quantitative aspects: the "judged association" of

scatterplots is not invariant (cf. Cleveland, Diaconis and McGill, 1982), and we usually carry along the eliminated location/scale information in numerical form, namely as marginal annotations of the graphs. But in any case, it suggests to separate off location/scale and to investigate the affine invariant aspects in separation.

Therefore, I shall from now on concentrate on Class III functionals.

Below, I shall adduce heuristic arguments to the effect that a projection is less interesting the more nearly normal it is. Intuitively, the central limit theorem says that convolution makes distributions more normal, hence the convolution of two distributions should be more normal (and less interesting) than the less normal among the two convolution factors. In other words, it is desirable that the projection index Q satisfies the following requirements:

(5.1) Q should be affine invariant (= Class III), and if X and Y are independent random variables with finite variances, then $Q(X + Y) \leq \max(Q(X), Q(Y))$.

Some heuristic arguments that interestingness goes together with nonnormality, are as follows.

- A multivariate distribution is normal, iff all of its one-dimensional projections are normal (this is one of the well-known characterizations of multivariate normality). So all of them are equally (un-)interesting.
- In particular, if the least normal projection is normal, we need not look at any other projection.
- For most high-dimensional point clouds most low-dimensional projections are approximately normal. This statement has recently been made precise by Diaconis and Freedman (1984). See also Figure 5.1, which compares a random projection of the corners of a seven-dimensional cube with a symmetrized normal sample.

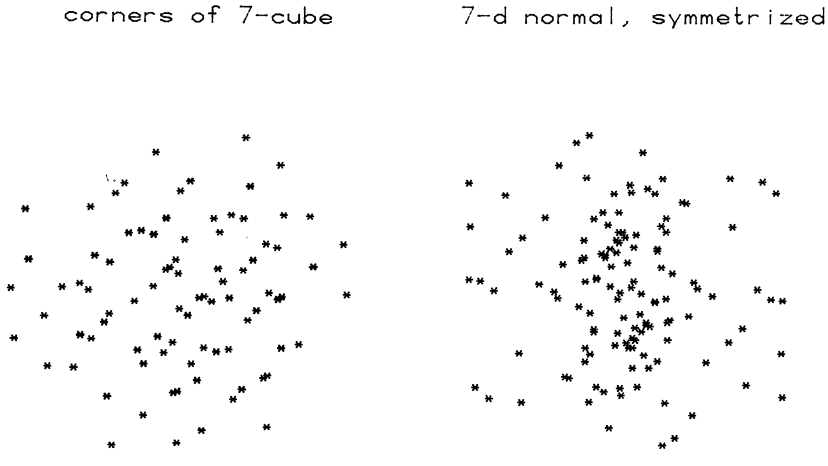

FIG. 5.1.

NOTE. Diaconis has pointed out that PP generalizes from Euclidean space to more general groups and their quotient spaces. Naturally, on a compact group the place of the least interesting distribution is taken by the uniform one.

All examples of functionals satisfying (5.1) I know of are of the form

$$Q(X) = h(S_1(X)/S_2(X)), \tag{5.2}$$

where h is monotone increasing function, and S_1, S_2 are scale functionals (= Class II) satisfying

$$S_1^2(X + Y) \leq S_1^2(X) + S_1^2(Y), \quad \text{(subadditivity)}, \tag{5.3}$$

and

$$S_2^2(X + Y) \geq S_2^2(X) + S_2^2(Y), \quad \text{(superadditivity)}, \tag{5.4}$$

respectively, when X and Y are independent random variables. Property (5.1) then follows easily from the well-known inequality

$$\frac{a + c}{b + d} \leq \max\left\{\frac{a}{b}, \frac{c}{d}\right\} \tag{5.5}$$

valid for arbitrary positive real numbers.

We first give three examples of sub- and superadditive scale functionals. All are closely related to f-divergency, see Cziszar (1967).

EXAMPLE 5.1. Let c_m be the mth cumulant of X:

$$c_m = (d/idt)^m \log(E(e^{itX})), \tag{5.6}$$

and let

$$S_1(X) = |c_m|^{1/m}, \quad m \geq 2. \tag{5.7}$$

Then S_1^2 is subadditive. This follows from the Minkowski inequality and the fact that c_m is an additive functional.

EXAMPLE 5.2. Let

$$S_2^2 = \frac{1}{\int (f'/f)^2 f\, dx} \tag{5.8}$$

be the inverse Fisher information, then S_2^2 is superadditive. This superadditivity is intuitively evident from the remark that $S_2^2(X) + S_2^2(Y)$ is the asymptotic variance of the sum of two asymptotically efficient location estimators based on X and Y, respectively, while $S_2^2(X + Y)$ is the asymptotic variance of the best estimator based on $X + Y$. For a formal proof, see Blachman (1965).

EXAMPLE 5.3. Let

$$S_2 = \exp\left\{-\int \log(f) f\, dx\right\} \tag{5.9}$$

be exponential Shannon entropy, then S_2^2 is superadditive. For a proof, see Blachman (1965).

This yields the following three examples of projection indices satisfying (5.1).

EXAMPLE 5.4. Standardized absolute cumulants:

$$Q(X) = |c_m(X)|/c_2(X)^{m/2}, \quad m > 2. \tag{5.10}$$

In particular, for $m = 3$ we obtain (absolute) skewness, for $m = 4$ (absolute) kurtosis.

EXAMPLE 5.5. Standardized Fisher information:

$$Q(X) = \sigma^2(X) \int \left(\frac{f'}{f}\right)^2 f\,dx - 1. \tag{5.11}$$

EXAMPLE 5.6. Standardized negative Shannon entropy:

$$Q(X) = \int \log(f) f\,dx + \log((2\pi e)^{1/2}\sigma(X)). \tag{5.12}$$

It is straightforward to see that in all three examples $Q(X) \geq 0$, with equality if X is normal. In Example 5.4, this follows trivially from the fact that the higher cumulants of the normal distribution are 0. In Example 5.5, we note that Q can be rewritten as

$$Q(X) = \sigma^2(X) \int \left(\frac{f'}{f} - \frac{\phi'}{\phi}\right)^2 f\,dx \geq 0, \tag{5.13}$$

where ϕ is a normal density with the same mean and variance as f. In Example 5.6, we can rewrite Q with the same ϕ as

$$Q(X) = -\int \log\left(\frac{\phi}{f}\right) f\,dx, \tag{5.14}$$

and $Q(X) \geq 0$ follows from Jensen's inequality. In fact, in the last two examples (but not in Example 5.4), $Q(X) = 0$ conversely implies that X is normal.

More generally, if Q is any functional satisfying (5.1), and if $X_1, \cdots, X_n$ are independent copies of any random variable X with finite variance, then (5.1) implies by induction

$$Q(X_1 + \cdots + X_n) \leq Q(X), \tag{5.15}$$

and if Q is weak-star lower semicontinuous, it follows by passing to the limit that

$$Q(N) \leq Q(X), \tag{5.16}$$

where N is a normal random variable.

Note that any Q satisfying (5.1) essentially amounts to a test statistic for testing normality. According to Ferguson (1961), skewness and kurtosis (Example 5.4) are most powerful for testing normality against the presence of outliers. A sample version of Example 5.5 amounts in essence to a test statistic for testing whether the score function $-f'/f$ is a straight line, compare (5.13). Finally, if we write the standardized Shannon entropy of Example 5.6 in the form (5.14) and

expand it into a Taylor series in powers of $\Delta = f - \phi$, we obtain the approximations

$$Q(X) \simeq \frac{1}{2}\int \frac{\Delta^2}{f}\,dx \simeq \frac{1}{2}\int \frac{\Delta^2}{\phi}\,dx. \tag{5.17}$$

If we approximate further by taking a finite Riemann sum and by inserting a histogram type density estimate for f, we see that the entropy index asymptotically amounts to a χ^2-test statistic used for testing normality.

In passing, I should mention that we have found empirically that all the usual test statistics for normality (Kolmogorov-Smirnov, Durbin-Watson, etc.) give about the same results when they are used as projection indices; that is, they tend to find very similar directions. Major exceptions to this rule are skewness and kurtosis, which are very outlier sensitive. Incidentally, this makes one wonder about the quartimax and oblimax method of factor analysis (see Harman, 1967), which are, essentially, PP-methods based on kurtosis. Though, only test statistics that increase under deconvolution (i.e. satisfy (5.1)) are conceptually satisfactory for finding least normal projections.

Finally, I should remark that the original Friedman-Tukey (1974) index mentioned in the introduction can be described in our framework as being a finite sample version of the "abstract" projection index.

$$Q(X) = \sigma_\alpha(X)\int f^2\,dx, \tag{5.18}$$

where σ_α stands for the α-trimmed standard deviation, and f is the density of X. This is a Class III functional, but it is not a consistent test of normality; it reaches its minimum at a density of the form $(a - bx^2)_+$, for some constants $a > 0$, $b > 0$.

To obtain a sample version of (5.18), we replace σ_α by the sample α-trimmed standard deviation $\hat{\sigma}_\alpha$, and the density f by the kernel estimate

$$\hat{f}(x) = (1/n)\sum k(x - x_i) \tag{5.19}$$

with

$$\begin{aligned} k(x) &= 1/R \quad \text{for} \quad |x| < R/2, \\ &= 0 \quad\quad \text{otherwise.} \end{aligned}$$

Note that

$$\int \hat{f}(x)^2\,dx = \frac{1}{R^2n^2}\sum_{i,j}(R - |x_i - x_j|)_+. \tag{5.20}$$

We obtain—apart from a proportionality factor—the original 1974 Friedman-Tukey index

$$\hat{\sigma}_\alpha \sum_{i,j}(R - |x_i - x_j|)_+ \tag{5.21}$$

if we take R to be 0.1 times the sample standard deviation in the direction of the largest principal component of the unprojected point cloud. Clearly, this choice

of R is not affine equivariant, and thus (5.21) is not of Class III (but this was changed in later implementations; personal communication by Friedman). To obtain a Class III sample version of (5.18), we might determine R in an equivariant fashion from the projected sample (e.g. by putting $R = c\hat{\sigma}_\alpha$, with say $c = 0.1$).

6. Two-sample PP and robust multivariate estimators. In this section it is more convenient to state the results in terms of finite samples X and Y in d-space. We already mentioned that the classically best discriminating hyperplane between X and Y can be found by doing PP with the 2-sample t-statistic as the projection index, or equivalently, by maximizing the projection index

$$\frac{\text{ave}\{a^T X\} - \text{ave}\{a^T Y\}}{\text{sdv}\{a^T(X \cup Y)\}}, \tag{6.1}$$

where sdv is the standard deviation. Note that (6.1) is a monotone function of the usual 2-sample t-statistic.

This might be robustified for example by replacing the sample average by the median, and the standard deviation by the median absolute deviation:

$$\frac{\text{med}\{a^T X\} - \text{med}\{a^T Y\}}{\text{mad}\{a^T(X \cup Y)\}}. \tag{6.2}$$

However, I would not advocate using this expression as it stands: I conjecture that a modified denominator, for example $\text{mad}\{(a^T X - \text{med}(a^T X)) \cup (a^T Y - \text{med}(a^T Y))\}$, should lead to better results.

The supremum over a of (6.2) (or of one of its variants) provides a very robust, affine invariant measure of the separation between X and Y. We can put this to good use for measuring the outlyingness of an observation x_i in a single sample X: put

$$r_i = \sup_a \frac{a^T x_i - \text{med}\{a^T X\}}{\text{mad}\{a^T X\}}. \tag{6.3}$$

This can be used to construct highly robust multivariate estimators. Let $w(r)$ be a strictly positive, decreasing function of $r \geq 0$, such that $rw(r)$ is bounded, and define weights

$$w_i = w(r_i).$$

Then the statistic

$$T_w = \sum w_i x_i / \sum w_i \tag{6.4}$$

is an affine equivariant estimator of location, and

$$C_w = \sum w_i^2 (x_i - T_w)(x_i - T_w)^T / \sum w_i^2 \tag{6.5}$$

is an affine equivariant estimator of the scatter matrix. If the points of X are in general position (i.e. no $d + 1$ of them lie in a $(d - 1)$-dimensional hyperplane)

then the breakdown point of both T_w and C_w is

$$\varepsilon^* \geq (n - 2d + 1)/(2n - 2d + 1) \tag{6.6}$$

(with equality in dimensions $d > 2$, and $\varepsilon^* = \frac{1}{2}$ for $d = 1$ or 2). This result is due to Donoho (1982); similar results (for infinite samples) have been obtained somewhat earlier by Stahel (1981).

PROOF (Donoho). We must show that T_w and C_w remain bounded (and nonsingular) unless we add at least $n - 2d + 1$ bad points to the sample X, where n is the size of X. Let Y be a set of m bad points, then so long as $m < n$,

$$\min\{a^T X\} \leq \operatorname{med}\{a^T(X \cup Y)\} \leq \max\{a^T X\};$$

hence

$$\sup_{|a|=1} |\operatorname{med}\{a^T(X \cup Y)\}| \leq \max_i |x_i|.$$

A similar argument gives

$$\sup_{|a|=1} \operatorname{mad}\{a^T(X \cup Y)\} \leq 2 \max_i |x_i|.$$

For a lower bound on the mad, note that $\operatorname{mad}\{a^T(X \cup Y)\} = 0$ only if strictly more than half of the elements of $a^T(X \cup Y)$ have the same value, that is, only if more than $(n + m)/2$ points of $X \cup Y$ lie in some $(d - 1)$-dimensional hyperplane. Since no more than d points from X can lie in such a plane by assumption, the number of contaminating points then must satisfy $d + m > (n + m)/2$. So if X is in general position and $n \geq m + 2d$,

$$\inf_{\#Y=m} \inf_{|a|=1} \operatorname{mad}\{a^T(X \cup Y)\} > 0.$$

These inequalities imply after some further algebra that the weights w_i of the points in X are bounded away from 0 (uniformly in Y), and that the $w_i |x_i|$ and the $w_j |y_j|$ are bounded, so long as $n \geq m + 2d$. Hence T_w and C_w are bounded and nonsingular, and this implies the inequality (6.6). □

These PP estimators are the only known estimators of multivariate location and scatter that are both affine equivariant and whose breakdown point approaches ½ in large samples.

7. Questions of k-dimensional projections. In the preceding sections we were concerned with one-dimensional projections of d-dimensional point clouds. The same approach, maximizing some functional Q of distributions in $\mathbb{R}^k$, applies to higher dimensional projections, but it has drawbacks:

(1) computations get harder (maximization over approximately kd instead of d variables);
(2) it yields only a k-dimensional subspace, but for interpretational reasons, one would prefer to get an ordered set of k directions.

Therefore, stepwise approaches are attractive; fix the first $k - 1$ directions found and optimize among projections onto the k-space spanned by the fixed

$k-1$ plus one additional variable direction. But also this does not give a sequence of directions, merely a nested sequence of subspaces. Note that orthogonal directions do not suffice, the interesting directions may be oblique to each other. (The approach can of course also be reversed: find first a k-dimensional projection, then reduce dimensions one by one.)

If we want to find a sequence of directions, recursive approaches are more appealing: find the most interesting direction, remove the structure that makes this direction interesting, and iterate. In essence, this amounts to PP density estimation (PPDE, Friedman, Stuetzle and Schroeder, 1984); so long as we are concerned not with the sample, but with the population version, we should better call it PP density approximation (PPDA). We postpone this problem until Section 11.

It is conceivable that stepwise approaches may miss structure that a direct k-dimensional search would find easily. After all, it is an empirical fact that a two-dimensional scatterplot may show striking features that would pass unnoticed in any one-dimensional projection. For example, holes (empty regions) are very hard to discover in low-dimensional projections. We have no reason to assume that machine search behaves much differently from visual search, and Example 14.1 below may give analytical support to this assertion.

8. What next? After one has found some "interesting" projections, what does one do next? Typically, the next action is one of the following list (part (1) corresponds to the operational paradigm behind Friedman and Tukey (1974) and the PRIM-9 system; parts (2) and (3) correspond roughly to PPC and PPR):

(1) Identify clusters, isolate them and investigate them separately.
(2) Identify clusters and locate them (i.e. replace them by, say, their center and classify points according to their membership to a cluster).
(3) Find a parsimonious description (separate structure from random noise in a nonparametric fashion).

Clearly, there is a floating boundary between the entries in this list, and the details need investigation.

We note that often a cluster can be characterized by the location of its center and the scatter matrix of the points forming the cluster.

Assume for the moment that we would like to optimize a PP procedure for finding clusters. Then, even in the relatively simple case of (possibly overlapping) elliptical clusters with different centers and covariance structures, it is far from clear how we should optimize the choice of objective functions Q. In view of Section 5, the problem of detecting such clusters may be formalized as a test of normality whose power is optimized for a particular class of nonparametric alternatives. Clearly, PP here infuses some new ideas and problems into the old field of nonparametric tests. On the other hand, determining the shape of the (elliptical) clusters is a problem of robust estimation of scatter matrices (with a twist—we typically will want to concentrate on a *minority* of the points and ignore the *majority*).

If the clusters are not nearly elliptical, a description in terms of scatter

matrices becomes inappropriate. For nonconvex clusters (e.g. curved and twisted "sausages") low-dimensional projections should still be able to reveal the presence of structure, but they may be of little help in unravelling it, mainly because each projection may show confusing overlapping effects. Compare also Tukey (1982).

In such cases, a separation of structure from noise ("sharpening") may reduce overlapping effects and thus help with the interpretation. It recently has emerged that PP methods are able to yield one of the most general and theoretically clearest approaches to sharpening by deconvoluting the underlying distribution (see Section 18). But first we must discuss some representational problems.

III. Projection pursuit regression (PPR)—abstract version.

9. Projection pursuit regression. Let (X, Y) be a pair of random variables such that X is $\mathbb{R}^d$-valued and Y is $\mathbb{R}$-valued. The problem is to estimate the response surface

$$f(x) = E(Y \mid X = x)$$

from n observations $(X_1, Y_1), \cdots, (x_n, Y_n)$ of (X, Y). A straightforward nonparametric approach to this problem consists in estimating $f(x)$ from the values Y_i observed at the k points X_i nearest to x, for example, by fitting a constant, or preferably, a linear function to them and evaluating it at x. Under weak assumptions this approach will estimate $f(x)$ consistently, compare Stone (1977). Though, if d is large, the curse of dimensionality causes trouble, and it may be more attractive to approximate the response surface by a sum of *ridge functions*:

$$f(x) \sim \sum_1^m g_j(a_j^T x). \tag{9.1}$$

Note that ridge functions may be thought of as generalizations of linear functions: like the latter they are constant on hyperplanes.

The projection pursuit estimation and approximation process, proposed by Friedman and Stuetzle (1981), works roughly as follows. Assume that we already have determined the first $m - 1$ terms, that is, vectors a_j and functions g_j of one real variable. Let

$$r_i = Y_i - \sum_1^{m-1} g_j(a_j^T X_i) \tag{9.2}$$

be the residuals of this approximation. Let $a \in \mathbb{R}^d$ be any unit vector, plot r_i against $a^T X_i$, and fit a smooth curve g to this scatterplot (Figure 9.1; see Section 20 for an explanation of the wiggly appearance of the curve).

Calculate the sum of the squared residuals relative to this g:

$$\sum_i (r_i - g(a^T X_i))^2, \tag{9.3}$$

and then minimize this sum over all possible choices of directions a. The minimizing direction a_m and the corresponding smooth function g_m then are inserted as the next term into the approximating sum (9.2). The process is iterated until the improvement in (9.3) becomes small.

This procedure has some definite advantages over its closest competitors: in

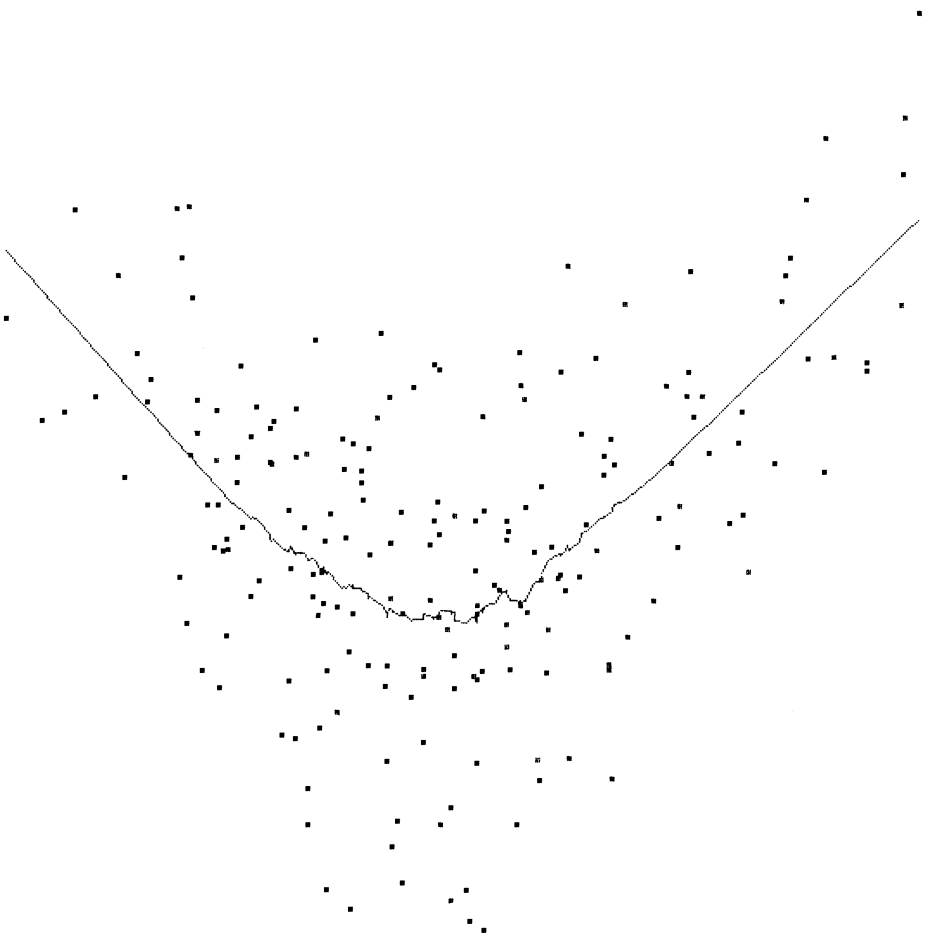

FIG. 9.1.

distinction to nearest neighbor techniques it is able to ignore information-poor variables, and it appears to be much better suited to the representation of intrinsically smooth response surfaces than methods based on recursive partitioning. Once the directions a_j and the functions g_j have been determined, the right-hand side of (9.1) can be evaluated very quickly.

On the other hand, there are considerable technical difficulties. In particular, the choice of the bandwidth of the smoother used to find g is very delicate. The sampling theory of PPR is practically nonexistent. The interpretations of the individual terms in the approximating sum is far from easy.

We shall disregard sampling aspects for the moment (they shall be taken up again in Section 20) and shall concentrate on the problem of approximating a given function f by an expansion of the form (9.1).

First, in what sense should the series (9.1) converge to f? If the dimension is $d > 1$, then the summands are not integrable in $\mathbb{R}^d$, unless they are zero almost

everywhere, so L_2-convergence makes sense only with respect to some bounded (i.e. probability) measure P in $\mathbb{R}^d$.

$$\int (f(x) - \sum_1^m g_j(a_j^T x))^2 \, dP \to 0. \tag{9.4}$$

To fix the idea, we may take P to be the uniform measure on the unit cube. Then it is clear that every square integrable function can be approximated in the sense of (9.4); indeed, the ordinary Fourier series expansion of f is of this form.

Assume that we already have determined projection directions a_j and functions g_j for $j < m$. Now we want to determine a_m, g_m such that the norm

$$\int r^2 \, dP \tag{9.5}$$

of the residual function

$$r(x) = f(x) - \sum_1^{m-1} g_j(a_j^T x) \tag{9.6}$$

is decreased by the maximum possible amount when $g_m(a_m^T x)$ is added to the sum in (9.6).

For fixed a_m, the solution is given by

$$g_m(z) = E(r(X) \mid a_m^T X = z), \tag{9.7}$$

where the conditional expectation is taken under the assumption that X is distributed according to the probability measure P.

PROOF OF (9.7). Let E' denote the conditional expectation, given $a_m^T X = z$. Then, among functions g of z, $E'[(r - g)^2]$—and thus $E[(r - g)^2]$—clearly is minimized for $g_m = E'(r)$. □

Moreover, we note that $E[(r - g_m)^2] = E(r^2) - E(g_m^2)$, so the residual norm is decreased most by choosing the direction a_m so that it maximizes the marginal norm $E(g_m^2)$.

Under mild smoothness conditions, $E(g_m^2)$ depends continuously on the direction a_m, so a standard compactness argument yields that a maximizing direction a_m exists.

By induction, the residual $r = r_m$ in (9.6) satisfies

$$E(r_m^2) = E(f^2) - \sum_1^{m-1} E(g_j^2) \geq 0.$$

It follows in particular that the maximal marginal norm $E(g_m^2)$ of the residual r_m converges to 0 as $m \to \infty$.

This does not imply that $E(r_m^2) \to 0$. However, since $E(|g_m|) \to 0$, it follows that the Fourier transform $\hat{r}_m(s) = E(r_m \exp(is^T X))$ converges uniformly to 0 (to show this, use a relation similar to (3.1)). It is evident from this remark that in order to prove L_2-convergence, it would suffice to establish a tightness condition on the frequency spectrum of r_m. Since projections are a kind of smoothers, they should not dissipate spectral power to higher frequencies; therefore, I conjecture

that no additional regularity conditions are needed, but I do not have a proof. The successive approximations

$$f_m(X) = \sum_1^m g_j(a_j^T x), \quad m = 1, 2, \cdots$$

to f need not be the best possible for m summands. In general, it is possible to improve the fit by various versions of *backfitting*: omit one of the earlier summands g_j, determine the best possible replacement and then iterate. Usually, the directions a_j are kept constant in this process.

10. Exact representations by finite sums. It is of some interest to know the structure of the approximating sums (9.1), or in other words, of the functions f that can be represented *exactly* by a finite sum of ridge functions. For simplicity we shall only consider the case $d = 2$, and shall assume that all functions are smooth (but this assumption could easily be removed). A detailed discussion can be found in Diaconis and Shahshahani (1984).

First, we note that the representation

$$f(x) = f(x_1, x_2) = \sum_1^n g_j(a_{1j}x_1 + a_{2j}x_2) \tag{10.1}$$

need not be unique. For example, in view of the identity

$$x_1x_2 = (1/4ab)[(ax_1 + bx_2)^2 - (ax_1 - bx_2)^2], \tag{10.2}$$

$f(x) = x_1x_2$ has infinitely many representations as a sum of two ridge functions.

This example involves a homogeneous polynomial, and, in fact, this kind of indeterminacy is the only one that occurs. More precisely, assume that f has two representations:

$$f(x) = \sum_1^n g_j(a_j^T x) = \sum_1^m h_j(b_j^T x), \tag{10.3}$$

where $(a_1, \cdots, a_n)$ are pairwise linearly independent two-dimensional vectors, and similarly for $(b_1, \cdots, b_m)$. Standardize these vectors such that $\| a_j \| = \| b_k \| = 1$, and that the first nonzero component of each a_j, b_k is > 0.

PROPOSITION 10.1 (Diaconis and Shahshahani (1984)). *For each j, either g_j is a polynomial, or else there is a k such that $a_j = b_k$, and $g_j - h_k$ is a polynomial.*

PROOF. It suffices to show that if

$$\sum_1^n g_j(a_j^T x) = 0, \tag{10.4}$$

with $(a_1, \cdots, a_n)$ pairwise linearly independent vectors, then g_1 is a polynomial. Note that a ridge function can be annihilated by taking the derivative in direction of the ridge. Thus, if we successively annihilate all summands in (10.4), except the first one, by taking derivatives in directions orthogonal to $a_2, a_3, \cdots, a_n$, we find that the $(n-1)$st derivative of g_1 must vanish identically. Hence g_1 is a polynomial of degree $\leq n - 2$. □

There are functions that cannot be represented in the form (10.1) for any

finite n. A simple example is

$$f(x) = e^{x_1 x_2}. \tag{10.5}$$

Obviously, this function is not annihilated by any finite number of directional derivatives, hence it cannot be of the form (10.1).

IV. Projection pursuit density approximation (PPDA).

11. Multiplicative expansions. If f is not just any arbitrary function on $\mathbb{R}^d$, but a probability density, then additive decompositions in the style of Section 9 are distinctly awkward; the approximating sums will not be probability densities themselves, unless one resorts to ad hoc tricks like taking positive parts, truncating (to ensure L_1 integrability) and rescaling. Multiplicative decompositions make more sense; approximate f by

$$f_k(x) = \prod_1^k h_j(a_j^T X). \tag{11.1}$$

Note that if $k = d$, and if the a_j are linearly independent, then (11.1) amounts to approximating f by a product density (in a coordinate system with basis vectors a_j).

If $k < d$, then (11.1) is not integrable; therefore, we shall prefer representations of the form

$$f_k(x) = f_0(x) \prod_1^k h_j(a_j^T x), \tag{11.2}$$

where f_0 is some standard probability density in $\mathbb{R}^d$ (e.g. a normal density with the same mean and covariance as f).

We can view the sequence $\{h_j, a_j\}$ either "synthetically," as a sequence of modifications to f_0 that builds up the structure of f, such that f_k converges to f in a suitable sense. Or else, we can view it "analytically," as a sequence of modifications to f that strips away its structure, step by step, such that the sequence

$$f_{-k}(x) = f(x) \prod_1^k h_j(a_j^T x)^{-1} \tag{11.3}$$

converges to f_0 for $k \to \infty$.

The two viewpoints bear some relevance on how we would determine the sequence $\{h_j, a_j\}$. Assume that δ is a metric in the space of probability densities, and that we have determined the sequence $\{h_j, a_j\}$ up to $k - 1$. Then according to the synthetic viewpoint, we would want to determine h_k, a_k such that $\delta(f, f_k)$ is minimized; according to the analytic viewpoint, we would minimize $\delta(f_0, f_{-k})$. The two approaches are dual to each other: they interchange the roles of f_0 and of f. Unless we use special properties of either f_0 or f (e.g. that f_0 is normal or that f is estimated from a sample), it therefore suffices to treat one of the two approaches.

The quality of the approximation of a density g to a density f can be measured in many ways, for example by

(1) relative entropy

$$E(f, g) = \int \log\left(\frac{f}{g}\right) f \, dx, \tag{11.4}$$

(which is not a metric, since $E(f, g) \neq E(g, f)$ in general), or by

(2) Hellinger distance

$$H(f, g) = \int (\sqrt{f} - \sqrt{g})^2 \, dx, \tag{11.5}$$

or by any other measure of distance beween distributions (Prohorov distance, bounded Lipschitz metric, etc.). Since we are working with densities, we naturally are more attracted to discrepancy measures defined in terms of the densities like (1) or (2) than to the other distances mentioned.

Among the two, (1) is particularly well suited to an additive decomposition of $\log f$, implicit in (11.2) and (11.3), while (2) is better matched to an additive decomposition of $\sqrt{f}$; incidentally, this is another way of forcing positivity of the approximating densities

$$f_k(x) = (\textstyle\sum_1^k h_j(a_j^T x))^2.$$

12. Properties of relative entropy. We begin with a few auxiliary lemmas.

Let

$$\begin{aligned} p(z) &= \tfrac{1}{2}z^2 \qquad \text{for} \quad |z| \leq 1, \\ &= |z| - \tfrac{1}{2} \quad \text{for} \quad |z| > 1. \end{aligned} \tag{12.1}$$

We note that p is a continuously differentiable convex function.

LEMMA 12.1. *For $z > -1$, we have*

$$z - \log(1 + z) \geq \tfrac{1}{2}p(z).$$

PROOF. Let $\ell(z) = z - \log(1 + z) - \tfrac{1}{2}p(z)$. We have $\ell(0) = 0$, and we easily verify that

$$\begin{aligned} \ell'(z) &= \frac{z(1 - z)}{2(1 + z)} \quad \text{for} \quad |z| < 1, \\ &= \frac{z - 1}{2(1 + z)} \quad \text{for} \quad z \geq 1. \end{aligned}$$

Hence $\ell'(z) < 0$ for $-1 < z < 0$, and $\ell'(z) > 0$ for $z > 0$, thus ℓ reaches its (unique) minimum at $z = 0$, and the assertion of the lemma follows. □

LEMMA 12.2. *Relative entropy satisfies*

$$E(f, g) = -\int \log\left(\frac{g}{f}\right) f \, dx \geq \frac{1}{2} \int p\left(\frac{g}{f} - 1\right) f \, dx \geq 0,$$

and in particular, $E(f, g) = 0$ implies $f = g$ a.e.

PROOF. Put $h = g/f - 1$. Then

$$-\int \log\left(\frac{g}{f}\right) f\,dx = \int [h - \log(1+h)]f\,dx - \int hf\,dx.$$

Since $\int hf\,dx = \int_{f>0} (g-f)\,dx \le 0$, the inequalities of the lemma follow from Lemma 12.1. Evidently, $E(f, g) = 0$ implies $p(g/f - 1) = 0$ a.e. $[f]$; thus $g/f = 1$ a.e. $[f]$, and since both g and f are probability densities this implies $f = g$ a.e. □

LEMMA 12.3.

$$\int (\sqrt{f} - \sqrt{g})^2\,dx \le \int |f - g|\,dx \le [2E(f, g)]^{1/2}.$$

In particular, if $E(f, f_n) \to 0$, then $f_n \to f$ in L_1 and in Hellinger distance.

PROOF. The first inequality is trivial:

$$(\sqrt{f} - \sqrt{g})^2 \le |\sqrt{f} - \sqrt{g}|\,|\sqrt{f} + \sqrt{g}| = |f - g|.$$

For the second, see Kemperman (1969, page 162 f.) and Cziszar (1975). □

We note that relative entropy is invariant under arbitrary affine transformations (in fact, under arbitrary differentiable 1-1-transformations).

The following lemma must be known, but I do not have a ready reference.

LEMMA 12.4. *Assume that f is a probability density in $\mathbb{R}^d$ which has finite second moments. Then the best Gaussian approximation g to f in the relative entropy sense (i.e. minimizing $E(f, g)$) has the same mean vector and the same covariance matrix as f.*

PROOF. In view of the preceding remark we may, without loss of generality, choose the coordinate system such that f has mean zero and unit covariance matrix. Let g_0 be the standard normal density in $\mathbb{R}^d$, and let g be any other normal density in $\mathbb{R}^d$, with mean vector μ and covariance matrix Σ. Then

$$\begin{aligned} E(f, g) - E(f, g_0) &= \int \log\left(\frac{g_0}{g}\right) f\,dx \\ &= \tfrac{1}{2}[\log(\det \Sigma) + E_f\{(x-\mu)^T\Sigma^{-1}(x-\mu)\} - E_f\{x^t x\}] \\ &= \tfrac{1}{2}[\log(\det \Sigma) + \operatorname{tr}(\Sigma^{-1}) + \mu^T\Sigma^{-1}\mu - d]. \end{aligned}$$

Assume that the eigenvalues of Σ^{-1} are $\lambda_1, \cdots, \lambda_d$, then this can be written

$$= \tfrac{1}{2}[\Sigma(\lambda_i - \log \lambda_i) + \mu^T\Sigma^{-1}\mu - d].$$

Since $\lambda - \log \lambda \ge 1$, with equality iff $\lambda = 1$, we obtain that $E(f, g) - E(f, g_0) > 0$, unless $g = g_0$. □

13. Minimization of relative entropy and PPDA. In this section we are concerned with optimal choices for the directions a_j and the functions h_j in a

decomposition of the form (11.1). Assume first that $k = d$, and that the a_j are fixed, linearly independent vectors in $\mathbb{R}^d$. Without loss of generality, we may choose a_j to be the jth coordinate direction (cf. the remark preceding Lemma 12.4). In other words, the problem is to find the best approximation of a given density f by a product density

$$g(x) = \prod_1^d g_j(x_j), \tag{13.1}$$

where the g_j are one-dimensional probability densities.

The quality of the approximation shall be measured in terms of relative entropy

$$E(f, g) = \int [\log f - \sum \log g_j(x_j)] f(x_1, \cdots, x_d)\, dx_1 \cdots dx_d. \tag{13.2}$$

Clearly, this is minimized by minimizing

$$-\sum \int \log[g_j(x_j)] f(x_1, \cdots, x_d)\, dx_1 \cdots dx_d,$$

which in turn is minimized by minimizing

$$-\int \log[g_j(x_j)] f_j(x_j)\, dx_j$$

for each j separately, where

$$f_j(x_j) = \int f(x_1, \cdots, x_d)\, dx_1 \cdots d_{x_{j-1}}\, d_{x_{j+1}} \cdots d_{x_d}$$

is the jth marginal density.

Since $E(f_j, g_j) > 0$ for $g_j \neq f_j$, (Lemma 12.2), the minimum clearly is achieved for the unique choice $g_j = f_j$.

We note in passing that this calculation at the same time proves that Shannon entropy

$$E_{\mathrm{Sh}}(f) = -\int \log(f) f\, dx$$

satisfies

$$E_{\mathrm{Sh}}(f) \leq \sum E_{\mathrm{Sh}}(f_j)$$

with equality iff $f = \prod f_j$.

By letting the d directions a_j vary simultaneously and freely, we may more generally approximate f by the best possible product density (we do not worry about existence of the minimum for the moment):

$$g(x) = \prod_1^d g_j(a_j^T x).$$

I do not know whether the best possible approximation can be constructed by a stepwise approach: first solve a minimum problem to find a_1 and g_1, then find a_2 and g_2, and so on.

But if f is an exact product density in a suitable coordinate system, then there is a stepwise approach that will sequentially pick up the unique factors one at a

time. This is a nontrivial result involving some subtle properties of entropy; it shall be sketched briefly.

The basic idea is to use the "analytic" approach mentioned in Section 11. Let g be the Gaussian density with the same mean and covariance matrix as f, and let f_a, g_a be the one-dimensional marginal densities of $a^T x$ under f and g, respectively. Note that the relative entropy $E(f_a, g_a)$ coincides with the Class III projection index of Example 5.6:

$$E(f_a, g_a) = \int \log\left(\frac{f_a}{g_a}\right) f_a \, dx$$

$$= Q(f_a) = \int \log(f_a) f_a \, dx + \log(\sqrt{2\pi e}\sigma(f_a)).$$

Now assume that f is a product density; without loss of generality we may assume that

$$f(x) = \prod f_j(x_j),$$

and that the factors are ordered such that

$$E(f_1, g_1) \geq E(f_2, g_2) \geq \cdots \geq E(f_d, g_d) \geq 0.$$

Since $Q(f_a)$ satisfies condition (5.1) of Section 5.2, it follows that Q reaches its maximum at a factor of f, namely at the factor f_1 with the largest relative entropy $E(f_1, g_1)$. We divide out this factor and replace f by

$$f^*(x) = f(x)g_1(x_1)/f_1(x_1).$$

We note that f^* still is a product density,

$$f^*(x) = \prod f_j^*(x_j)$$

with $f_1^* = g_1$, $f_j^* = f_j$ for $j > 1$. Thus, if f^* is subjected to the same process as f before, the second factor f_2 is picked out, and so on. If $E(f_j, g_j) = E(f_{j+1}, g_{j+1}) > 0$, the order in which the two factors are picked is indeterminate. The process continues until $f^* = g$ is a Gaussian density, that is, until either $j = d$ or $E(f_j, g_j) = 0$, whichever happens first.

Now let g be any approximation to any given density f in $\mathbb{R}^d$. We shall attempt to improve the approximation by replacing $g(x)$ by a density of the form $g^*(x) = g(x)h(x_1)$, where h depends on the first coordinate only. Note that g and g^* determine the same conditional density given x_1. An intuitively attractive choice thus is to determine h such that the one-dimensional marginal distribution g_1 of g^* in direction x_1 agrees with the corresponding marginal distribution f_1 of f. We shall show that this indeed minimizes relative entropy.

LEMMA 13.1. *Relative entropy $E(f, g^*)$ is minimized by the choice*

$$h(x_1) = f_1(x_1)/g_1(x_1),$$

where f_1 and g_1 are the marginal densities of f and g in direction x_1, and for this choice, the decrease in relative entropy is

$$E(f, g) - E(f, g^*) = E(f_1, g_1).$$

PROOF. We note that the conditional density, given x_1, is the same for g and g^*, however h_1 is chosen, namely

$$g(\cdot \mid x_1) = g(x_1, \cdots, x_n)/g_1(x_1)$$

and that $g_1(x_1)h(x_1)$ is the marginal density of g^*. Thus

$$E(f, g^*) = \int (\log f - \log g^*) f \, dx$$

$$= \int [\log f(\cdot \mid x_1) - \log g(\cdot \mid x_1) + \log f_1(x_1) - \log(g_1(x_1)h(x_1))] f \, dx_1$$

which is minimized by minimizing

$$\int [\log f_1 - \log(g_1 h)] f \, dx = E(f_1, g_1 h),$$

and this clearly is achieved by the unique choice $g_1 h = f_1$. This proves the first assertion of the lemma. The second assertion follows from a composition of the above expression for $E(f, g^*)$ for the two choices $h = 1$ and $h = f_1/g_1$:

$$E(f, g) - E(f, g^*) = -\int \log(g_1(x_1)) f \, dx + \int \log(f_1(x_1)) f \, dx$$

$$= E(f_1, g_1). \quad \square$$

According to this lemma, if we may choose the projection direction a, then the largest possible improvement in relative entropy that can be achieved through replacing $g(x)$ by $g^*(x) = g(x)h(a^T x)$ clearly is obtained with

$$h = f_a/g_a,$$

where f_a and g_a are the marginal densities of f and g, respectively, in direction a, and where a is chosen such that it maximizes

$$E(f_a, g_a) = E(f, g) - E(f, g^*). \tag{13.3}$$

At the moment, we are not concerned with the existence of such an a; maximization within a prescribed relative error tolerance is in fact good enough for all practical purposes.

The procedures just described shall be referred to as the projection pursuit density approximation (PPDA) method: find a direction a maximizing $E(f_a, g_a)$, and then either replace f by $f^* = fg_a/f_a$ ("analytic" version) or g by $g^* = gf_a/g_a$ ("synthetic" version), then iterate.

We already noted that this procedure (with a normal g) finds the least normal projection of f, and if the analytic version is applied iteratively to a product density, it will find the unique factors in descending order of nonnormality. It would be interesting to know (cf. Section 15) whether these results remain true if we interchange the arguments and maximize $E(g_a, f_a)$ instead.

14. Maximum marginal relative entropy. Let f and g be arbitrary probability densities in $\mathbb{R}^d$, and let f_a and g_a be their one-dimensional marginals

in direction a. We can use *maximum marginal relative entropy*

$$E^*(f, g) = \sup_a E(f_a, g_a)$$

as a measure of discrepancy between f and g. Clearly, because of Lemma (13.1), $E^*(f, g) \leq E(f, g)$. Since any distribution is uniquely characterized by the family of its marginals in all possible directions (Cramér-Wold theorem), we have

$$E^*(f, g) = 0 \Rightarrow f = g \Rightarrow E(f, g) = 0.$$

If we determine a sequence $\{g^{(k)}\}$ of successive approximations by PPDA, then each step decreases $E(f, g^{(k)})$ by $E^*(f, g^{(k)})$. Hence, if $E(f, g) < \infty$, it follows that $E^*(f, g^{(k)})$ converges to 0; in fact, for any given $\varepsilon > 0$, it takes at most $k = E(f, g)/\varepsilon$ steps to reach a density $g^{(k)}$ for which $E^*(f, g^{(k)}) \leq \varepsilon$.

Maximum marginal relative entropy is a concept particularly well suited to PPDA, and it deserves a closer study. Let f be a fixed probability density in $\mathbb{R}^d$, while $g^{(k)}$ is an arbitrary sequence of probability densities.

Clearly, $E(f, g^{(k)}) \to 0$ implies $E^*(f, g^{(k)}) \to 0$. The reverse implication is false.

EXAMPLE 14.1 (D. Critchlow). Let f be the uniform density in the unit disk in $\mathbb{R}^2$:

$$\begin{aligned} f(x) &= 1/\pi \quad \text{for} \quad \|x\| < 1, \\ &= 0 \quad \text{otherwise.} \end{aligned}$$

Let $g^{(k)}$ be defined as follows (see Figure 14.1):

$$\begin{aligned} g^{(k)}(x) &= 1/\pi \quad \text{for} \quad 1/k \leq \|x\| < 1, \\ &= 2/\pi \quad \text{for} \quad \|x\| < 1/k, \quad x_1 > 0, \\ &= 0 \quad \text{otherwise.} \end{aligned}$$

It is straightforward to verify that $E(f, g^{(k)}) = \infty$ for all k, but $E^*(f, g^{(k)}) \to 0$.

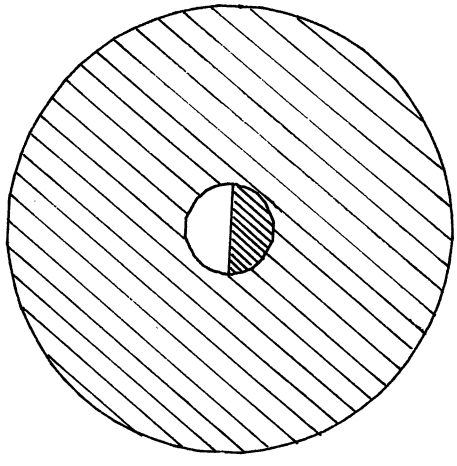

FIG. 14.1.

We shall now derive a few consequences of E^*-convergence.

PROPOSITION 14.2. *$E^*(f, g^{(k)}) \to 0$ implies that $g^{(k)} \to f$ in the sense of weak(-star) convergence of the underlying measures.*

Basically, this proposition is just another version of the Cramér-Wold theorem, compare Billingsley (1968, page 48).

PROOF. In view of Lemma 12.3, $E^*(f, g^{(k)}) \to 0$ implies that the marginal densities show uniform L_1-convergence:

$$\sup_a \int |f_a - g_a^{(k)}| \to 0.$$

Hence, the characteristic functions ψ_a of the one-dimensional marginals converge uniformly, and since the characteristic function ψ of any density f is related to the characteristic functions ψ_a of its marginals f_a by

$$\psi(ta) = E(e^{ita^T x}) = \psi_a(t), \quad t \in \mathbb{R}, \quad a \in \mathbb{R}^d$$

it follows that the characteristic functions $\psi^{(k)}$ of $g^{(k)}$ converge uniformly to ψ:

$$|\psi(ta) - \psi^{(k)}(ta)| \leq \sup_a \int |f_a - g_a^{(k)}|.$$

Hence, $g^{(k)}$ converges weakly. □

I conjecture that if f is sufficiently smooth, so that its characteristic function is absolutely integrable, and if the sequence $g^{(k)}$ is generated by PPDA, then $g^{(k)} \to f$ uniformly and in the L_1-sense. Actually, I can prove only a very special case (which however covers the intended applications).

PROPOSITION 14.3. *Assume that the density f in $\mathbb{R}^d$ can be deconvoluted with a Gaussian component:*

$$f = \bar{f} * \phi,$$

where $\bar{f}$ is some density, and ϕ is normal $\mathcal{N}(0, \sigma^2 I_d)$ for some $\sigma^2 > 0$. Let $g^{(0)}$ be the normal density with the same mean and covariance matrix as f. Then the sequence $g^{(k)}$, $k = 0, 1, 2, \cdots$, of approximating densities, constructed by PPDA (Section 13, end), converges uniformly and in L_1 to f.

PROOF. Each of the $g^{(k)}$ allows a deconvolution $g^{(k)} = \bar{g}^{(k)} * \phi$. This shall be proved by induction. It clearly is true for $k = 0$, since $g^{(0)}$ itself is Gaussian.

Thus, assume that $g = g^{(k)}$ can be deconvoluted. For the following argument it is essential that ϕ is a product density in all orthogonal coordinate systems, so that convolutions can be calculated coordinate-wise. Choose the coordinate system such that a_k is the first coordinate direction. Then the product representation of $\bar{g}$:

$$\bar{g} = \bar{g}(\cdot \mid x_1)\bar{g}_1(x_1)$$

induces a product representation of g:

$$g = g(\cdot \mid x_1)g_1(x_1),$$

with $g(\cdot \mid x_1)$ and $g_1(x_1)$ being obtained by convoluting their barred counterparts with $(d-1)$- and one-dimensional normal densities, respectively. If in the last displayed equation we replace g_1 by $f_1 = \bar{f}_1 * \mathcal{N}(0, \sigma^2)$, we obtain the desired deconvolution of the next term $g^{(k+1)}$.

The characteristic functions of $g^{(k)}$ and $\bar{g}^{(k)}$ are related by $\psi^{(k)}(s) = \bar{\psi}^{(k)}(s)\exp(-\sigma^2 \mid s \mid^2/2)$, and a similar relation holds for the characteristic functions ψ, $\bar{\psi}$ of f, $\bar{f}$, hence ψ and $\psi^{(k)}$ are majorized by $\exp(-\sigma^2 \mid s \mid^2/2)$.

It follows that the $\psi^{(k)}$ are absolutely integrable, and $g^{(k)}$ can therefore be represented as

$$g^{(k)}(x) = (2\pi)^{-d} \int \psi^{(k)}(s)e^{-isTx}\, ds.$$

Since the sequence $\psi^{(k)}$ converges uniformly, it now follows from the majorization of ψ and $\psi^{(k)}$ that the sequence

$$\mid f(x) - g^{(k)}(x) \mid \leq (2\pi)^{-d} \int \mid \psi(s) - \psi^{(k)}(s) \mid\, ds$$

converges uniformly to 0. L_1-convergence follows trivially (from uniform convergence and tightness of the weakly convergent sequence of measures $g^{(k)}$). □

V. Projection pursuit density estimation (PPDE).

15. General remarks on PPDE. It is straightforward to change the PPDA procedures of Section 13 into density estimators. The first step is to standardize the point cloud in $\mathbb{R}^d$ by an affine transformation so that it is centered at 0 and that its covariance matrix is the unit matrix. Note that this raises some very delicate robustness questions; our density estimates should not be sensitive to occasional outliers, but they should be able to pick up long tails in the underlying distribution. These two requirements are contradictory; we lack a rational basis for separating between spurious outliers and genuine long tails. But from a pragmatical point of view, we note that all density estimators have trouble coping with isolated points in the tails—at best, these points produce equally isolated bumps in the estimate, and at worst, (especially if we use adaptive cross-validation), they may act as leverage points, messing up the estimate elsewhere. So it is probably wise to identify and to set aside such isolated points (for example, with the help of the measure (6.3) of outlyingness), and to disregard them in the estimation procedures, but to show them as remarkable points in (at least some of) the pictures we produce.

Since it seems to be better if the initial estimate has too heavy tails than if it has too light tails, we would seem to have the choice of either using the classical, nonrobust mean and covariance matrix together with a Gaussian $g^{(0)}$, or else robust location and covariance estimates together with a density $g^{(0)}$ that is

heavier tailed than the Gaussian (but we should not combine a robust covariance estimate with a Gaussian $g^{(0)}$). I believe the simpler first version to be good enough in most cases.

The zeroth order density estimate $g^{(0)}$ thus ordinarily is the d-dimensional standard normal. The approximation steps now can be described as follows.

Let

$$g(x) = g^{(k)}(x) = g^{(0)}(x) \prod_1^k h_j(a_j^T x)$$

be the current density estimate.

According to Section 13, we should determine a next direction $a = a_{k+1}$ such that it maximizes $E(f_a, g_a)$, and then put $h_{k+1} = f_a/g_a$.

For a given a, f_a is straightforward to estimate: project the sample in direction a, yielding $z_i = a^T x_i$, $i = 1, \cdots, n$, and then calculate a one-dimensional density estimate $\hat{f}_a$ based on $(z_1, \cdots, z_n)$.

The projection g_a of the current density estimate is a well-defined quantity, and from the point of view of theory does not present any problem. However, we may run into trouble with its calculation, in particular since it has to be calculated inside a minimization loop. Direct numerical integration almost certainly is too slow. There are several appealing Monte Carlo approaches (cf. Friedman, Stuetzle and Schroeder 1984). A first one is to replace g by a sample $y_1, \cdots, y_n$ from g, and then to estimate $\hat{g}_a$ in the same way as $\hat{f}_a$. This may not be easy to implement (how does one sample efficiently from g?). A second one is to take a random sample $y_1, \cdots, y_N$ from some cleverly chosen distribution with density $h^{(0)}(x)$ in $\mathbb{R}^d$ (for example, a truncated normal one, if outliers have been purged away from the original data sample). Put

$$v_i = \frac{g^{(0)}(x)}{h^{(0)}(x)} \prod_1^k h_j(a_j^T y_i),$$

and create a histogram with bin width Δ and value

$$\ell_j = \sum \{v_i \mid z_j - \Delta/2 \le a^T x_i < z_j + \Delta/2\}$$

for the bin with midpoint z_j. Then apply a kernel smoother to this histogram to obtain the estimate $\hat{g}_a$.

A possibly even more appealing approach is not to maximize $E(f_a, g_a)$, but $E(\hat{f}_a^{(-k)}, g_a^{(0)})$, where $\hat{f}_a^{(-k)}$ is a sample version of (11.3), defined as follows. Let

$$w_i = \prod_1^k h_j(a_j^T x_i)^{-1},$$

and create a histogram with bin width Δ and value

$$\ell_j = \sum \{w_i \mid z_j - \Delta/2 \le a^T x_i < z_j + \Delta/2\}$$

for the bin with midpoint z_j. Then apply a kernel smoother to this histogram to obtain the estimate $\hat{f}_a^{-k}$. This corresponds to the procedure mentioned in the last sentence of Section 13.

We proposed here to use marginal entropy as the criterion to be maximized when determining a new projection direction. This certainly is the conceptually

purest approach. But it is conceivable that other measures of discrepancy might offer advantages from a sampling or computational point of view; the matter needs further investigation.

16. Consistency of PPDE. By reinterpreting some of the results of Section 14, consistency of PPDE is almost trivial to prove. But the proof at the same time shows why consistency per se is a rather useless concept.

Assume that we are given a sample $x_1, \cdots, x_n$ in $\mathbb{R}^d$, and let

$$\mu = (1/n) \sum_1^n \delta_{x_i} \tag{16.1}$$

be the empirical measure. We now apply a spherically symmetric normal kernel smoother to obtain the density estimate

$$\hat{f} = \mu * \mathcal{N}(0, \sigma^2 I_d). \tag{16.2}$$

Note that the marginal density $\hat{f}_a$ of $\hat{f}$ in direction a is obtained by applying the one-dimensional kernel $\mathcal{N}(0, \sigma^2)$ to the projection

$$\mu_a = (1/n) \sum \delta_{a^T x_i} \tag{16.3}$$

of the original data.

It follows that if we iterate a PPDE (using relative entropy as the criterion), with a Gaussian kernel smoother in the projections, it numerically converges in the E^*-sense to the d-dimensional kernel estimate $\hat{f}$.

Since $\hat{f}$ can be deconvoluted with a normal component, it follows from Proposition 14.3 that the PPDE converges uniformly and in L_1 to $\hat{f}$, if the number of iterations tends to infinity.

The d-dimensional kernel estimate $\hat{f}$ is consistent under very weak assumptions on the true underlying density f if σ tends to 0 slowly, while the sample size n goes to ∞. It follows that PPDE is consistent too, provided we iterate it enough so that it approximates $\hat{f}$ sufficiently closely.

This result is not very helpful, however. After all, the main reason for using PPDE is that the sample size is too small for a d-dimensional kernel estimator to make sense. We certainly do not intend to iterate the PP density estimation so far that it approximates the latter. The following example may illustrate the issue.

EXAMPLE 16.1. Take a sample of size n from the standard normal $\mathcal{N}(0, I)$ in d dimensions. Note that a half-and-half mixture of two one-dimensional normal densities $\mathcal{N}(\pm\alpha, \sigma^2)$ is unimodal for $\alpha \leq \sigma$, bimodal otherwise, and that the same holds true for a pair of d-variate normal densities $\mathcal{N}(\pm\mu, \sigma^2 I)$ with $\sigma = \|\mu\|$. Thus, we may say that two sample points are merged in the d-dimensional kernel estimate (16.2), if their Euclidean distance is $\leq 2\sigma$, and that they are separated otherwise. The expected number of merged point pairs can be calculated as

$$m = \tfrac{1}{2}n(n-1)q,$$

where

$$q = P\{\| x_i - x_j \| \leq 2\sigma\} = \chi_d^2(2\sigma^2)$$

is the probability that a specified pair of points is merged. Numerically, we obtain with $d = 10$, $\sigma = 0.1$ and $n = 10^6$ that $m = 0.4$. In other words, we have a better than even chance that all 10^6 sample points are separated. The one-dimensional marginal estimates (with the same kernel) on the other hand will be quite smooth. Note that for one-dimensional estimates a kernel width $kn^{-1/5}$ minimizes the asymptotic mean square error, and that the constant k is such that for $n = 10^6$ the choice $\sigma = 0.1$ is close to optimal (see, e.g., Wegman 1972, page 536).

Note that in this example, the zeroth order PPDA to the underlying density is exact, and in the sampling case the starting density $g^{(0)}$ already is the best PPDE; more generally, if the kth order PPDA $g^{(k)}$ is exact, then there is no reason to go much beyond order k in the PPDE $\hat{g}^{(k)}$, and depending on the sample size, it may be preferable to stop much earlier.

More meaningful consistency results should therefore be concerned with the convergence and the speed of convergence of $\hat{g}^{(k)} - g^{(k)} \to 0$, for *fixed* k.

Since we would not know the true f in practice, we would also need an analogue of Mallows C_p-statistic, telling us when to stop the projection pursuit approximation process.

VI. Connections to computer tomography.

17. Fixed projection directions. Computer tomography (CT), just like PP, is concerned with the reconstitution of a higher dimensional structure from lower dimensional projections. For an introductory survey of CT, see Shepp and Kruskal (1978).

But there are many differences. The most conspicuous one is the absence of a search for informative projections in CT. CT aims for an accurate reconstruction of a not directly observable two-dimensional density from the set of all one-dimensional projections; in practice, one only has a finite, but fairly dense and equispaced set of projections, and they are affected by random observational errors. In PP, on the other hand, the higher dimensional information is directly accessible, but it consists only in a random sample from the density, and the latter should be approximated on the basis of a few selected projections of the random sample.

Nevertheless, some of the mathematics is closely related. For example, the algebraic reconstruction techniques (Gordon, Bender and Herman 1970), whose applications to CT have now been superseded by Fourier techniques, but which have advantages if the data is not equispaced, use the same iterative improvement techniques as PPR and PPDA.

The questions of common relevance to both CT and PP concern, in particular, approximations based on a finite number of discrete projections. Assume that we are to approximate a density function f in $\mathbb{R}^d$, and that we are given m fixed directions $a_1, \cdots, a_m$. What is the "best" additive approximation

$$g(x) = \sum_1^m h_j(a_j x), \tag{17.1}$$

and what is the "best" multiplicative approximation

$$g(x) = \prod_1^m h_j(a_j x) \tag{17.2}$$

to f?

Second, given only the projections f_j (i.e. either the conditional expectations or the marginal densities) of f in the directions a_j, what is the "best" choice of g under the side condition that the projections g_j of g in the directions a_j agree with those of f?

In each case, the notion of "best" needs to be made precise. For additive approximations (17.1), it appears appropriate to formalize "best" so as to minimize the L_2-norm

$$\int (f - g)^2 \, dP; \tag{17.3}$$

for multiplicative approximations (17.2), so as to minimize relative entropy,

$$E(f, g) = \int \log\left(\frac{f}{g}\right) f \, dx. \tag{17.4}$$

For the second type of problem, we may define the "best" choice of g to be the one with the least variability:

$$\int g^2 \, dP = \min!, \tag{17.5}$$

or the one with the largest entropy

$$H(g) = -\int \log(g) g \, dx = \max!. \tag{17.6}$$

Assume that the space M of functions of the form (17.1) is a closed subspace of the Hilbert space of square integrable functions. Then (17.1), (17.3) and (17.5) are nicely matched up: the solution g is uniquely described by the property that it is of the form (17.1) and satisfies $g_j = f_j$.

Geometrically, this is obvious:

(1) If g minimizes $E(f - g)^2$ among all functions in M, then $f - g \perp M$; this orthogonality relation is equivalent to $E(f - g \mid a_j X) = 0$, or $g_j = f_j$ for $j = 1, \cdots, m$.

(2) The orthogonal projection of f to M minimizes Eg^2 among all functions satisfying $g_j = f_j$ for $j = 1, \cdots, m$.

The relations (17.2), (17.4) and (17.6) are matched up in an analogous fashion.

Unfortunately, it is not at all clear whether M is closed; see the remark in Shepp and Kruskal (1978, page 428), and consult Hamaker and Solmon (1978) for a laborious proof in a special case. See furthermore Logan and Shepp (1975) and Logan (1975) for a detailed study of the number of terms required in (17.1) (in terms of the energy distribution of the Fourier transform $\hat{f}$).

VII. Projection pursuit and time series problems.

18. Minimum entropy deconvolution as a sharpening technique. Deblurring and sharpening are ubiquitous problems; just think of how to make a sharp picture from a snapshot blurred by camera motion during exposure, or of how to improve the sound of a historic phonograph record. Our own visual system is surprisingly good at it (this is exploited for example in the anti-aliasing techniques of computer graphics: by suitably blurring a staircase line we can trick our eyes into reconstructing a sharp straight line).

Often, the blurring process is known in detail, and deblurring amounts to the undoing of a known (not necessarily linear) filter. Here we are interested in the other extreme case, where the filter is not known and has to be reconstructed together with the underlying process from the data. Already the case of linear filters (the only case we are going to consider) amounts to the seemingly unsolvable task of factoring an observed process y into a convolution product of two unobservable factors:

$$y = f * x,$$

where f is the unknown filter which has blurred the underlying process of interest x.

To fix the idea, assume that y is a time series, observed at equidistant points: $y = (\cdots, y_t, y_{t+1}, \cdots)$. Thus,

$$y_t = \sum_s f_s x_{t-s},$$

and the problem is to find a filter q inverse to f, so that $q * y = x$.

These are several conceptually different approaches; a common theme behind many of them is to view x as a bottommost, not further reducible causative process. More or less, this amounts to assuming that knowing the past values x_r, $r \leq t$, does not help you in predicting a future value x_s, $s > t$, and that the future values of x do not influence the past values of y, so that $f_s = 0$ for $s < 0$. Assuming stationarity, the first requirement means that the x_t are independent, identically distributed random variables, and we shall make the gratuitous assumption that they have finite variances.

Now, for any filter q, we have $q * y = (q * f) * x$, so $(q * y)_t$, being a linear combination of several x_s, is more normal than a single x_t (in the sense of Section 5.2).

Thus, the filter q inverse to f has the property that it produces a least normal $q * y$. Clearly, q is not unique since it can be shifted in time (replace q_t by q_{t-k} and x_t by x_{t+k}), and if y is a Gaussian process, then q is completely indeterminate.

In the PP framework of Section 5.2, we may phrase the problem as follows: restrict the maximum length of the filter q to d. Consider the segments $(y_t, y_{t+1}, \cdots, y_{t+d-1})$ of length d as points in $\mathbb{R}^d$. Find a least normal one-dimensional projection; the corresponding direction q may be taken as an approximation to f^{-1}.

A concrete application of these notions can be made in geophysics. Donoho

(1981) pointed out the usefulness of the considerations of Section (5.1) in a time series context, and related them to current work on deconvolution in exploration seismology. What geophysicists call "Minimum Entropy Deconvolution"—introduced by Wiggins (1978)—is actually a PP method in the present sense, with kurtosis (not entropy) as a projection index. The point of MED is to recover a convolution component which for geological reasons is supposed to be "impulsive" or "spiky." Modelling such a component as "non-Gaussian i.i.d." one obtains just such impulsive series; and the PP approach described above is the optimal deblurring procedure under that model—if one uses the right projection index. In this case it turns out that the right index actually is standardized entropy, which the MED nomenclature might have suggested; this results from large-sample statistical considerations not employed by Wiggins in naming the method.

19. A time series version of PPR. For stationary Gaussian processes, or more precisely, for processes allowing a harmonic decomposition

$$X_t = \int e^{2\pi i\lambda t} Y(d\lambda) \tag{19.1}$$

in terms of a process Y_λ with independent increments, spectrum analysis clearly is the approach of first choice; it separates the process into irreducible components.

For other processes, for example for those generated by the superposition of nonsinusoidal periodic wave forms, other approaches may be more appropriate. In particular, one might then prefer not to leave the time domain.

In concrete terms, suppose that the process X_t is of the form

$$X_t = \sum_j f_j(t/p_j) \tag{19.2}$$

(plus some noise, which we shall ignore for the moment), where p_j is the period and f_j the shape of the jth periodic component. The function f_j is assumed to be smooth and periodic with period 1. Both p_j and f_j are unknown, but we assume that each f_j averages to 0 over time.

We note that if the p_j are linearly independent over the field of rational numbers, then the representation (19.2) is unique, and it is in principle possible to extract the jth component by averaging over points spaced $p = p_j$ apart; if we put

$$Z_{p,t} = \text{ave}_k\{X_{t+kp}\}, \tag{19.3}$$

then

$$Z_{p_j,t} = f_j(t\ /p_j).$$

Note that

$$\text{ave}_t\{(X_t - Z_{p,t})^2\} = \text{ave}_t\{X_t^2\} - \text{ave}_t\{Z_{p,t}^2\}, \tag{19.4}$$

so it looks attractive to do projection pursuit with regard to the projection operator (19.3) and to search for a period p maximizing $Q(p) = \text{ave}_t\{Z_{p,t}^2\}$.

Unfortunately, this will pick bewilderingly many periods p. Note that $Z_{kp,t}$ is the same function for all $k = 1, 2, \cdots$, and if the f_j are nonsinusoidal, they have higher harmonics with periods p_j/ℓ, so every single component f_j will create spikes in $Q(p)$ at $p = kp_j/\ell$, for all $k \geq 1$ and for at least some $\ell \geq 1$.

Without doubt (19.3) provides a nice method for looking at the shape of periodic components with given periods, and it has been used successfully for example in the investigation of circadian rhythms (cf. Enright, 1981). But it is far from clear whether projection pursuit with Q as a method for uncovering hidden periods is preferable to more conventional methods based on the periodogram or complex demodulation, which search for periods p that yield large values of

$$|C(p)|^2 = |\operatorname{ave}_t\{e^{2\pi it/p}X_t\}|^2. \tag{19.5}$$

While this latter approach picks up the higher harmonics $p = p_j/\ell$, it ignores the spurious subharmonics ($p = kp_j/\ell$ with $k > 1$).

A comparison of the sample versions of these procedures is interesting. For every given value of p, plot X_t against $t(\text{mod } p)$. Then fit a smooth curve $\hat{Z}_{p,t}$ to this scatterplot to obtain a nonparametric estimate of $Z_{p,t}$, (see McDonald, 1982). We may compare this to the more traditional periodogram approach, which amounts to fitting the two parameters of a sine wave (amplitude and phase) to this same scatterplot.

The periodogram approaches have had an infamous reputation for picking spurious periods, because—prior to the book of Blackman and Tukey (1959)—people often had not paid enough attention to the sampling properties of (19.5). The sampling properties of $\operatorname{ave}\{\hat{Z}^2_{p,t}\}$ clearly are in need of an equally careful scrutiny!

VIII. Finite sample implementations of PP methods.

20. Sample versions of PPR. We continue the discussion begun in Section 9. Assume that a response surface

$$f(x) = E(Y \mid X = x), \tag{20.1}$$

where X is d-dimensional and Y is one-dimensional, is to be estimated from a sample $\{(x_i, y_i)\}$ of size n, and is to be approximated by a finite sum of estimated ridge functions:

$$f(x) \sim \sum_1^m \hat{g}_j(\hat{a}_j^T x). \tag{20.2}$$

The "engineering" aspects of constructing a good sample version of PPR are very delicate, even more so than what transpires from the published account (Friedman and Stuetzle, 1981), and they deserve a careful discussion.

The situation is analogous to that in numerical spectrum analysis. There the real progress did not come through mathematical statistics in the usual sense, that is, through consistency and asymptotic normality proofs, but through a mathematically much more primitive, qualitative and quantitative understanding

(see Blackman and Tukey, 1959). This understanding involved recommendations for balancing bias against variability; one realized that one was not estimating the "true" spectrum, but a smoothed version thereof, using estimates that had an approximate χ^2-distribution with so-and-so many degrees of freedom. It was even more important to sort out the pitfalls due to aliasing and leakage, and to learn how to avoid them; some pitfalls were discovered and remedied only recently, e.g. the masking effect due to (small) gross errors (Kleiner, Martin and Thomson, 1979).

In PPR, we are only at the beginning of this process. The main problem is that we are trying to estimate a response surface in a setup where there are not enough observations to do it through a direct, d-dimensional nonparametric approach. Unless we are very careful, the PPR estimate may get trapped by (local) overfitting in one of the low-order $\hat{g}_j$, thereby invalidating subsequent approximations. It may also go astray by including too many terms.

The PPR fitting procedure begins by simultaneously determining a direction $\hat{a}$ and a smooth function $\hat{g}$, such that the square average of the residuals

$$r_i = y_i - \hat{g}(\hat{a}^T x_i) \tag{20.3}$$

becomes least possible (in a sense to be made precise). Then the process is repeated iteratively, with the residuals r_i in place of the y_i. It suffices to describe the first step of the algorithm.

According to (9.7), the "ideal" function g, for a given direction a, is the conditional expectation

$$g(x) = E(Y \mid a^T X = z). \tag{20.4}$$

For the following discussion it may help to decompose y_i and write it as

$$y_i = g(a^T x_i) + [f(x_i) - g(a^T x_i)] + u_i. \tag{20.5}$$

Even if f and g are smooth and the random error u_i is small, the y_i may show a seemingly erratic behavior when plotted against $z_i = a^T x_i$, because of the variability of f in directions other than a. Overfitting at this stage would have catastrophic consequences with regard to subsequent iterative steps.

For each fixed choice of a, the smoothing algorithm proposed by Friedman and Stuetzle (1981) makes several passes over the data:

0. Sort the data in ascending order of the $z_i = a^T x_i$.
1. Smooth the scatterplot of y_i against z_i by running medians of three.
2. Estimate response variability at each z_i by the average squared residual of a locally linear fit with constant bandwidth.
3. Smooth these variance estimates by a fixed bandwidth moving average.
4. Smooth the sequence obtained by pass (1) by locally linear fits with variable bandwidths determined by the smoothed local variance estimates obtained in (3).

A few comments on the different passes follow.

Pass (1) is suggested by robustness; it intends to safeguard against isolated gross errors in the u_i. On the other hand, note that long tails in the distribution

of $f(X) - g(a^T X)$ may indicate that the choice of a can be improved. The two requirements conflict with each other; attempts at being robust may cost us dearly in terms of our ability to find good projection directions.

Furthermore, we remark that for any d-tuple of points x_i, we may find a direction a projecting them to the same $z = a^T x_i$. Thus, whenever there are r or more large positive outliers anywhere among the y-observations, any median smoother with span $< 2r + 1$ will break down in some direction a. By the way, it is not at all evident without a detailed analysis of the algorithm whether breakdown of the smoother here implies that the direction a is unjustly preferred or spurned later on. Either alternative may have unpleasant consequences.

In the passes (2) to (4), observation i is omitted from the local averaging process determining the smoothed value $\hat{g}(z_i)$. The main purpose of this cross-validation approach is to protect against overfitting. Locally linear, rather than locally constant, fitting helps to reduce the bias near the ends of the sequence; note that a large bias in some of the fitted values $\hat{g}(a^T x_i)$ may foul up the search for the best $\hat{a}$.

The entire curve-fitting process (passes (0) to (4)) occurs within a minimization loop; it is vitally important that it be done in a fast fashion. It appears that locally linear fits with constant weights over a fixed number of neighboring points are an excellent compromise between quality and speed; the smoothed value at z_{i+1} can be obtained by a simple (numerically unstable, but adequate) updating procedure from that at z_i. The main drawback of the constant weights is that the smoothed curve remains locally wiggly (cf. Figure 9.1). Actually, one runs several (say three) concurrent smoothers with different, but constant bandwidths in (2) and (3), and then for (4) chooses the one which gives the smallest local variability.

For the minimization, a simple and crude Rosenbrock algorithm is used. Note that—except for purposes of interpretation—it does not matter very much if a particular direction a_j is determined inaccurately, later terms in the sum (20.2) will correct it.

Since, especially in the earlier stages of the procedure, the as yet unexplained part of the variability of f can be quite large, and the smoothing is correspondingly unreliable, backfitting is much more important than in the abstract version: readjust the earlier summands $\hat{g}_j$ (and possibly also the $\hat{a}_j$) in turn, keeping the other $m - 1$ contributions to (20.2) fixed.

The fine-tuning of the PPR algorithms so far has been based on the intuition of the originators and on uncontrolled experimentation. For further progress, we would need some crude theories explaining the quantitative after-effects of particular choices for the bandwidth of the smoothers and some theoretical insight into stopping rules.

Consistency results may be mathematically interesting, but will be rather irrelevant. The point (already made earlier in this section) is that PPR is designed to work in the transient region where the sample size n is not yet large enough for direct d-dimensional nonparametric regression. The only way a consistency result can become useful is when it is accompanied by a realistic estimate of the approximation error of $\sum_1^m \hat{g}_j(\hat{a}_j^T x)$ relative to the best approximation to f of the form $\sum_1^m g_j(a_j^T x)$, and which is valid for sample sizes smaller than those needed for direct d-dimensional approaches.

21. How many points? If the sample is too small, PP methods are likely to find spurious features. The paper by Day (1969) gives a graphical demonstration of this fact (with 50 points in 10 dimensions).

We begin with the one-dimensional case. There, the Kolmogorov distance between the true and the empirical cumulative satisfies the asymptotic bound

$$P(\sup_x | F_n(x) - F(x) | \geq \varepsilon) \leq 2e^{-2n\varepsilon^2}. \tag{21.1}$$

For the sample sizes and probability values of interest, the bound practically is an equality. In particular, we shall mark down that for $n = \varepsilon^{-2}$,

$$P(\sup_x | F_n(x) - F(x) | \geq \varepsilon) \approx 0.27. \tag{21.2}$$

Note, for example, that the two distributions $\mathcal{N}(0, 1)$ and $\frac{1}{2}\mathcal{N}(-0.8, 0.36) + \frac{1}{2}\mathcal{N}(0.8, 0.36)$, whose densities are shown in Figure 21.1, have Kolmogorov distance 0.046. This example would seem to suggest that we should aim for values of $\varepsilon = 0.05$ and smaller, that is, for sample sizes in the range $n = \varepsilon^{-2} = 400$ and larger.

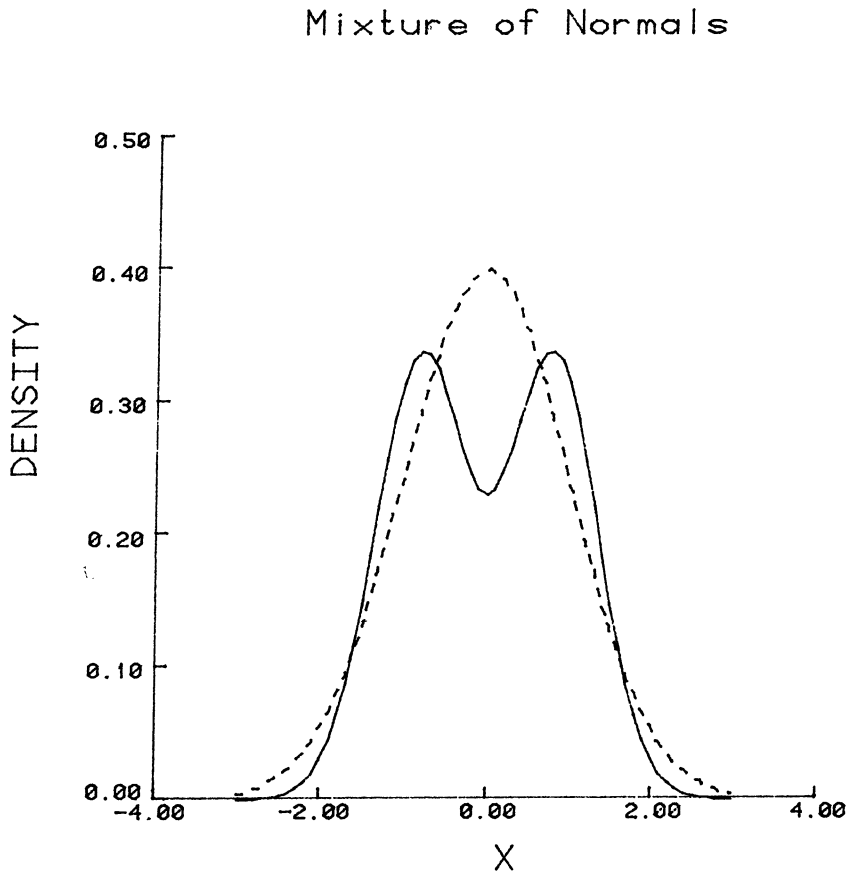

FIG. 21.1.

In higher dimensions, a theorem of Vapnik and Červonenkis (1971) gives the following upper bound for the Kolmogorov distance in the worst projection:

$$\eta = P\{\sup_a \sup_t |F_{a,n}(t) - F_a(t)| \geq \varepsilon\} \leq 4\Phi(d, 2n)e^{-n\varepsilon^2/8}. \tag{21.3}$$

Here, F_a and $F_{a,n}$, respectively, are the one-dimensional cumulatives of the true and of the empirical measure in $\mathbb{R}^d$, projected in direction a, and

$$\Phi(d, n) = \sum_{r=0}^{d} \binom{n}{r} \tag{21.4}$$

is the maximal number of regions into which d-space can be divided by n hyperplanes.

For $d \leq n/2$ we have

$$\Phi(d, n) \leq \binom{n}{d} \frac{n}{n-d} \leq 2\binom{n}{d}, \tag{21.5}$$

(this is shown by majorizing the sum by a geometric series), and

$$\binom{n}{d} \leq \frac{n^d}{d!} \leq n^d d^{-d} e^d (2\pi d)^{-1/2} \tag{21.6}$$

by Stirling's formula. Hence

$$\eta \leq 8\left(2e\,\frac{n}{d}\right)^d (2\pi d)^{-1/2} e^{-n\varepsilon^2/8}, \tag{21.7}$$

and thus

$$\log\left[\frac{(2\pi d)^{1/2}}{8}\,\eta\right] \leq d\left[\log\left(2e\,\frac{n}{d}\right) - \frac{\varepsilon^2}{8}\,\frac{n}{d}\right]. \tag{21.8}$$

This inequality improves the ones given by Vapnik and Červonenkis (1971), who had used the bound $\Phi(n, d) \leq n^d + 1$. In particular, it implies that for each $\varepsilon > 0$, the probability η of large deviations can be made arbitrarily small, uniformly in d, by choosing n/d sufficiently large. (This result is due to Ken Alexander.) This is about the weakest sufficient condition for consistency we could reasonably have hoped for.

The bad news is that the values of n/d turn out to be very large. For example, with $\eta = 0.27$ and $\varepsilon = 0.05$ we obtain from (21.8) that $n/d \approx 40000$. Even if this value should turn out to be too pessimistic by two orders of magnitude (as a comparison with the value $n/d \approx 400$ appropriate for $d = 1$ perhaps might suggest), the sample sizes required still would be much larger than the ones one usually encounters with multivariate data sets.

Perhaps the practical conclusion to be drawn is that we shall have to acquiesce to the fact that PP will in practice uncover not only true but also spurious structure, and that we must weed out the latter by other methods, for example by validating the results on different data sets.

Acknowledgements. In 1981–82, Persi Diaconis and myself conducted a seminar on PP at Harvard; this paper owes much inspiration to participants of

that seminar, in particular to Doug Critchlow and David Donoho. Most of the writing was done during a leave at the Mathematical Sciences Research Institute in Berkeley in the fall of 1982. Part of the material was presented at the Kiefer-Wolfowitz Memorial Conference at Cornell University, July 6–9, 1983. Finally, I thank the editors and referees for several constructive suggestions.

REFERENCES

BILLINGSLEY, P. (1968). *Convergence of Probability Measures.* Wiley, New York.

BLACHMAN, N. M. (1965). The convolution inequality for entropy powers. *IEEE Trans. Inform. Theory* **11** 267–271.

BLACKMAN, R. B. and TUKEY, J. W. (1959). *The Measurement of Power Spectra.* Dover, New York.

CHEN, Z. and LI, G. (1981). Robust principal components and dispersion matrices via projection pursuit. Research Report, Dept. of Statistics, Harvard University.

CLEVELAND, W. S., DIACONIS, P. and MCGILL, R. (1982). Variables on scatterplots look more highly correlated when the scales are increased. *Science* **216** 1138–1141.

CLEVELAND, W. S. and MCGILL, R. (1984). The many faces of a scatterplot. To appear in *J. Amer. Statist. Assoc.*

CRITCHLOW, D. (1981). Unpublished notes.

CZISZAR, I. (1967). Information-type measures of difference of probability distributions and indirect observations. *Studia Math. Sci. Hungar.* **2** 299–318.

CZISZAR, I. (1975). *I*-divergence geometry of probability distribution and minimization problems. *Ann. Probab.* **3** 146–158.

DAY, N. E. (1969). Estimating the components of a mixture of normal distributions. *Biometrika* **56** 463–474.

DEVLIN, S. J., GNANADESIKAN, R. and KETTENRING, J. R. (1981). Robust estimation of dispersion matrices and principal components. *J. Amer. Statist. Assoc.* **76** 354–362.

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

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

DONOHO, D. L. (1981). Minimum entropy deconvolution. In *Applied Time Series II.* (D. Findley, ed.)

DONOHO, D. L. (1982). Breakdown properties of multivariate location estimators. Ph.D. Qualifying Paper, Dept. of Statist., Harvard University.

ENRIGHT, J. T. (1981). Data Analysis. In *Handbook of Behavioral Neurobiology,* 4, *Biological Rhythms.* (J. Aschoff, ed.) Plenum Press, New York and London.

FERGUSON, T. S. (1961). On the rejection of outliers. *Proc. Fourth Berkeley Symp. on Math. Statist. and Probab.* **1** 253–288, (J. Neyman, ed.) Univ. California Press.

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

FRIEDMAN, J. H. and STUETZLE, W. (1980). Projection pursuit classification. Unpublished manuscript.

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

FRIEDMAN, J. H., STUETZLE, W. and SCHROEDER, A. (1984). Projection pursuit density estimation. *J. Amer. Statist. Assoc.* **79** 599–608.

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

GORDON, R., BENDER, R. and HERMAN, G. T. (1970). Algebraic reconstruction techniques (ART) for three-dimensional electron microscopy and x-ray photography. *J. Theoret. Biol.* **29** 471–481.

HAMAKER, C. and SOLMON, D. C. (1978). The angles between the null spaces of x-rays. *J. Math. Anal. Appl.* **62** 1–23.

HARMAN, H. H. (1967). *Modern Factor Analysis.* Univ. Chicago Press.

KAGAN, A. M., LINNIK, Y. V. and RAO, C. R. (1973). *Characterization Problems in Mathematical Statistics.* Wiley, New York.

KEMPERMAN, J. H. B. (1969). On the optimum rate of transmitting information. *Lecture Notes in Math.* **89** pp. 126–169, Springer-Verlag, Berlin.

KLEINER, B., MARTIN, R. D. and THOMSON, D. J. (1979). Robust estimation of power spectra. *J. Roy. Statist. Soc. Ser. B* **41** 313–351.

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

KRUSKAL, J. B. (1972). Linear transformation of multivariate data to reveal clustering. In *Multidimensional Scaling: Theory and Application in the Behavioral Sciences*, I, *Theory.* Seminar Press, New York and London.

LI, G. and CHEN, Z. (1981). Robust projection pursuit estimation for dispersion matrices and principal components. Research Report, Dept. of Statist., Harvard University.

LOGAN, B. F. (1975). The uncertainty principle in reconstructing functions from projections. *Duke Math. J.* **42** 661–706.

LOGAN, B. F. and SHEPP, L. A. (1975). Optimal reconstruction of a function from its projections. *Duke Math. J.* **42** 645–659.

MCDONALD, J. (1982). Unpublished manuscript.

SHEPP, L. A. and KRUSKAL, J. B. (1978). Computerized tomography: the new medical x-ray technology. *Amer. Math. Monthly* **85** 420–439.

STAHEL, W. A. (1981). Robuste schätzungen: infinitesimale optimalität und schätzungen von kovarianzmatrizen. Ph.D. Thesis, Swiss Federal Institute of Technology, Zurich.

STONE, C. J. (1977). Nonparametric regression and its applications. *Ann. Statist.* **5** 595–645.

SWITZER, P. (1970). Numerical Classification. In *Geostatistics.* Plenum, New York.

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

TUKEY, J. W. (1982). Control and stash philosophy for two-handed, flexible, and immediate control of a graphic display. *Bell Labs. Tech. Memo.*

TUKEY, P. A. and TUKEY, J. W. (1981). Graphical display of data in three and higher dimensions. In *Interpreting Multivariate Data.* (V. Barnett, ed.) Wiley, New York.

VAPNIK, V. N. and ČERVONENKIS, A. Y. (1971). On the uniform convergence of relative frequencies of events to their probabilities. *Theory Probab. Appl.* **16** 264–280.

VITUSHKIN, A. G. (1978). On representation of functions by means of superpositions and related topics. *Enseign. Math.* Monogr. no. 25.

WEGMAN, E. J. (1972). Nonparametric probability density estimation I: A summary of available methods. *Technometrics* **14** 533–547.

WIGGINS, R. A. (1978). Minimum entropy deconvolution. *Geoexploration* **16** 21–35.

DEPARTMENT OF STATISTICS
HARVARD UNIVERSITY
CAMBRIDGE, MASSACHUSETTS 02138

DISCUSSION

JEROME H. FRIEDMAN

Stanford University

I congratulate Professor Huber for an excellent survey of Projection Pursuit methods. Putting together the diverse research in this area into a coherent prospective is a difficult and challenging task. This paper represents an important