J. R. Statist. Soc. A (1987)
150, *Part* 1, *pp.* 1–36

What is Projection Pursuit?

By M.C. JONES and ROBIN SIBSON
University of Bath, UK

[*Read before the* Royal Statistical Society *on Wednesday, November 19th, 1986, the President* Professor J. Durbin *in the Chair*]

SUMMARY

Friedman and Tukey (1974) introduced the term "projection pursuit" for a technique for the exploratory analysis of multivariate data sets; the method seeks out "interesting" linear projections of the multivariate data onto a line or a plane. In this paper, we show how to set Friedman and Tukey's idea in a more structured context than they offered. This makes it possible to offer some suggestions for the reformulation of the method, and thence to identify a computationally efficient approach to its implementation. We illustrate its application to empirical data, and discuss its practical attractions and limitations. Extensions by other workers to problems such as non-linear multiple regression and multivariate density estimation are discussed briefly within the same framework.

Keywords: PROJECTION PURSUIT, EXPLORATORY MULTIVARIATE DATA ANALYSIS, CLUSTERING, NON-NORMALITY, ENTROPY, KERNEL DENSITY ESTIMATION, SKEWNESS, KURTOSIS, PRINCIPAL COMPONENTS, SPHERING

1. INTRODUCTION

1.1. *Background*

The term "projection pursuit" was first used by Friedman and Tukey (1974) to name a technique for the exploratory analysis of reasonably large and reasonably multivariate data sets; projection pursuit reveals structure in the original data by offering selected low-dimensional orthogonal projections of it for inspection. For projection from, say, two dimensions to one dimension, it is possible to examine essentially all such projections to select those of interest: the appearance of the projected data set does not change abruptly as the projection direction changes, and the space of projection directions, although forming a continuum, is of low dimensionality. For projection from higher dimensions it is still true that the appearance of the projected data changes smoothly, but (as pointed out by Tukey and Tukey, 1981b) it becomes increasingly impractical to explore possible projections exhaustively because of the high dimension of the space of projection directions. Friedman and Stuetzle (1982) describe interactive procedures for such exploration, but this does not extend the applicability of the approach as far as is needed. An automatic procedure for selecting potentially interesting projections is required. Projection pursuit is the process of making such selections by the local optimisation over projection directions of some index of "interestingness".

We cannot directly appreciate patterns of variation in more than three dimensions. Even three-dimensional point-clouds are not easy to display without the use of exotic computer graphics hardware of the kind employed in molecular modelling. In practice, then, projection will be onto one- or two-dimensional space; it would considerably limit the impact of the approach if only one-dimensional projections could be inspected, so it is important that both

Address for correspondence: Professor Robin Sibson, School of Mathematics, University of Bath, Claverton Down, Bath, Avon BA2 7AY, UK.

 0035–9238/87/150001 $2.00

the theory and the computational practicability of the method should extend at least to projection onto a plane.

Multivariate data sets in which several numerical variates are recorded for each individual (case) can easily contain observations on thousands of individuals; the number of variates is not usually more than about twenty, and very commonly between four and ten. A classic example is given by the "Fisher *Iris* data" (Anderson, 1935; Fisher, 1936) with 150 cases and 4 variates. An example studied later in this paper (Lubischew, 1962) has 74 cases and 6 variates. Such examples are typical of the smaller data sets which may be encountered.

Exploratory multivariate analysis is usually based on the hope that much of the data is redundant, and the main features can be described in terms of a tendency for the point cloud to concentrate into clusters, or about a curve or (generally non-flat) surface. Even where more subtle levels of structure are sought, it is usually necessary, as a first step, to be aware of these systematic variations in the density of the point cloud. Principal components analysis is a familiar exploratory technique of this kind; it is in fact a projection pursuit method in which the index of interestingness is the proportion of total variance accounted for by the projected data. Useful though PCA can be, particularly as a means of removing uninteresting directions of variation and reducing high-dimensional data to moderate dimensionality, it is something of a blunt instrument. It relies for its success on the tendency for large variation also to be interestingly structured variation, a connexion which is not logically necessary and often fails to hold in practice, at least, as precisely as we might wish. Friedman and Tukey (1974) built on work by J.B. Kruskal (1969, 1972) to construct an index of interestingness purposely designed to reveal clustering. Like other plausible indices apart from total variance, this has to be optimised numerically. They were able to do this and to demonstrate in examples that a potentially powerful exploratory technique resulted. This work apparently received little attention at the time, and Tukey and Tukey (1981b) were unable to report any real development of the original idea. A number of parallel developments used some of the ideas of projection pursuit in other contexts during this period, such as projection pursuit regression (Friedman and Stuetzle, 1981) and projection pursuit density estimation (Friedman, Stuetzle, and Schroeder, 1984); these techniques, which are of considerable promise but are clearly distinct from the original idea, are discussed briefly in Section 5.5. We are aware of only two *direct* lines of development from Friedman and Tukey's work, apart from those initiated very recently (Friedman, 1985; Jee, 1985). One is our own work, reported in Jones (1983, 1985) and now in the present paper. The other is the work of Huber (1981, 1985). These two lines developed independently, but have much in common. They both recognise the need to establish a framework within which projection pursuit can be understood as something more than the purely *ad hoc*, if effective, technique proposed by Friedman and Tukey (1974); we conjecture that its slow acceptance may stem partly from this *ad hoc* character. But whereas Huber's papers explore this framework in detail and relate projection pursuit derivatives such as those mentioned above to it, we concentrate more heavily on the practical implementation and application of the ideas involved. There are numerous specific ideas in common, such as the use of entropy as the basis for an index function.

The object of the present paper is partly expository; we hope that by providing a review of the background, implementation, and application of the basic method of projection pursuit as reformulated in Jones (1983), we shall promote the use of a potentially valuable exploratory method hitherto neglected by applied statisticians. We hope also that the development, here and in Huber (1985), of a structure within which to view projection pursuit systematically may stimulate further development of the basic technique and its derivatives.

1.2. *Notation, Terminology and Basic Requirements*

In what follows, N will always denote the number of individuals, or cases, in a data set, with m,n running through the cases. J,K will always denote dimensionality with i,j,k

running through dimensions. The observations are assumed to lie on a numerical scale, and form a $K \times N$ matrix X; thus coordinate vectors are column vectors, and univariate samples are row vectors. The transpose of a matrix or vector will be indicated by superscript T.

If X is such a data matrix, and a is a column K-vector, then $a^T X$ is a row N-vector; this vector is the orthogonal projection of the sample onto direction a, scaled by the magnitude of a. If H is a function measuring the interestingness of a one-dimensional sample, then $H(a^T X)$ is, for fixed X, a function $I(a)$ of the projection direction a; such a function I is a *projection index*. Projection pursuit attempts to find, by numerical optimisation, projection directions a which are local optima of $I(a)$. The case $a=0$ is uninteresting, and is excluded; usually a is constrained to be of unit length.

For orthogonal projection onto a space of dimension J greater than one, a is replaced by a $K \times J$ matrix A, so $A^T X$ is a $J \times N$ matrix. In this case, the columns of A are not merely required to be of unit length, but mutually orthogonal. H is a measure of interestingness for J-variate samples, and $I(A) = H(A^T X)$. The choice of orthogonal axes within J-dimensional space is not of interest, and H must respect this by an invariance property: if Y is $J \times N$, and P is $J \times J$ and orthogonal, then $H(PY) = H(Y)$. It is noteworthy that Friedman and Tukey's (1974) suggestion for a two-dimensional index did not enjoy this property, although a possible modification which would achieve it is fairly obvious. In one dimension, the invariance property is simply $H(y) = H(-y)$.

Commonly, the origin of coordinates is arbitrary, and it is desirable to reflect this in the behaviour of H. The appropriate condition is $H(Y + c1_N^T) = H(Y)$ for all column J-vectors c, where 1_N denotes the column N-vector of 1's; in other words, H is invariant under translation.

These conditions of course imply that from the point of view of projection pursuit, $K \times N$ data matrices which differ only by a euclidean transformation in K-dimensional space are equivalent. Projection pursuit will accordingly not be capable of extracting information from a data set which relies on there being something special about the choice of axes or origin in K-space. The question of how changes of scale should affect the situation is discussed in Section 2.1.

The practical consideration that it must be possible to optimise I numerically has significant implications. There is little prospect of being able to do this unless I is continuous; usually it is appropriate to obtain this property by forcing H to be continuous. The set on which I is defined is closed and bounded, and hence compact, so I is bounded and attains its bounds. This remark gives some reassurance that pathological behaviour will not arise. In practice we shall usually wish to employ value-and-gradient-based optimisation methods, so we require more: continuous differentiability of I, usually derived from that of H.

Throughout this section, definitions and remarks have been expressed in terms of finite samples. A completely parallel treatment can be given for distributions; the details are obvious, and do not bear repetition. The distributional viewpoint is crucial for our treatment of projection pursuit.

2. THE PROJECTION INDEX

Friedman and Tukey (1974) designed their projection index specifically to reveal clustering in the data; but in one of their most impressive examples the structure revealed by the method is not a cluster structure in the usual sense. This suggests that it may be very difficult to avoid confounding together different concepts of "interestingness" in the design of an index. We shall prefer to work in the opposite direction, by considering what constitutes an *un*-interesting projection and attempting to maximise divergence from it. It is of course not logically necessary that this will always and only lead to interesting projections; sometimes, as the example in Section 4 illustrates, some of the projections are of little help in interpreting the data. However, the indices which result from this approach are similar in

general character to those designed to capture specific aspects of the data such as clustering. Thus, clustering of the data may be expected to emerge automatically from such an approach as one interesting case; it will not be the only one. Our approach in this paper is to analyse and simplify Friedman and Tukey's construction until some insight into what it is doing emerges; then to vary it in directions we believe to be desirable. The parallels with Huber's (1985) independent treatment will be apparent. It is easiest to begin by considering projection to one dimension; unless otherwise stated, this is the case being considered.

2.1. *Centring and Sphering*

Questions of invariance have already been mentioned briefly in Section 1.2. Translational invariance will be assumed, and advantage taken of this to take the centroid of the data as the origin of coordinates, a standardisation which is preserved under projection. If X denotes the original $K\times N$ data matrix, then the centred data matrix is $X(I_N - 1_N 1_N^{\mathrm{T}}/N)$, where I_N is the $N\times N$ unit matrix; the second factor is a projection matrix. To keep the notation simple, the centred data matrix will now be denoted by X.

The method of principal components analysis already extracts scale effects, and there is no point in duplicating that information from projection pursuit. We shall adopt this requirement in its strongest form, by insisting that the one-dimensional projection index be scale-invariant, and we shall again choose to impose it by a transformation of the original data set. The appropriate transformation is called sphering and is discussed by Tukey and Tukey (1981a), who comment "sphering the data ... forces us to examine only aspects of our collection of data vectors other than its general ellipsoidal nature". Sphering is the process of transforming X by premultiplication by a $K\times K$ matrix Q such that $(QX)(QX)^{\mathrm{T}}/N = I_N$, in other words so that the transformed data has unit variance matrix (where N rather than $N-1$ is used as the denominator). This property is then inherited by every projection. Q is not characterised uniquely by this requirement, but there is a canonical transformation given by $Q = S^{-\frac{1}{2}}$, where $S = XX^{\mathrm{T}}/N$ is the sample variance matrix, which is assumed to be nonsingular, and where $S^{-\frac{1}{2}}$ is defined in the usual way by replacing each eigenvalue of the positive-definite matrix S by the inverse of its positive square root. Applying this transformation to the centred data is analogous to carrying out principal components analysis on the centred data; as with PCA, it is commonly preferable to work with the correlation matrix rather than the variance matrix. This corresponds to applying a preliminary transformation to the centred data by rescaling each axis to have unit variance; it is then the scaled centred data which is actually sphered as above. Again, it will be convenient to redefine X to denote the matrix currently being operated on: thus it becomes in turn the scaled centred data matrix and then the sphered scaled centred data matrix. The sphering transformation, like PCA, is a procedure equivariant under orthogonal transformations but not under general affine transformations. Projection pursuit is unaffected (see Section 1.2) by choice of orthonormal axes in which to express the scaled centred data. It is possible to take advantage of this to make a sphering transformation which is equivalent to the canonical one as far as projection pursuit is concerned, but a little simpler to calculate. Tukey and Tukey (1981a) discuss robust/resistant methods of sphering: see Section 5.1 below.

2.2. *The Friedman-Tukey Index*

Friedman and Tukey (1974) reported that researchers using their interactive projection display system commonly sought "projections that tended to produce many very small interpoint distances while, at the same time, maintaining the overall spread of the data". Their index, like earlier attempts by J.B. Kruskal (1969, 1972), tried to formalise this; they took a product of a "spread" term $s(a)$ and a "local density" term $d(a)$. The spread term partly normalised the index against scale effects, a task we have accomplished by sphering. It also incorporated outlier protection, an aspect we discuss further in Section 5.1.

It is the local density term which actually serves to capture structure in the data. Write $r_{mn}(a) = |Z_m(a) - Z_n(a)|$ where $Z(a) = a^{\mathrm{T}}X$ is the projection (of the sphered scaled centred

data, with a of unit length). Then

$$d(a) = \sum_{m=1}^{N} \sum_{n=1}^{N} g_h\{r_{mn}(a)\} I_{[0,\infty)}\{h - r_{mn}(a)\} \tag{1}$$

where h is a positive window-width parameter, I is the usual indicator function inserted to guarantee cutoff whenever $r_{mn}(a)$ exceeds h, and g_h is a score function. $d(a)$ is designed to score small separations highly, and one possible function suggested by Friedman and Tukey is a linearly decreasing score clipped to zero at $r=h$ by the indicator function factor. There is clearly redundancy in expression (1): provided that appropriate properties for g_h are specified, the indicator function factor can be dropped.

Expression (1) is plausible but apparently *sui generis*. To relate it to established ideas we introduce the concept of a kernel density estimate (see Rosenblatt, 1956; Fryer, 1977; Wertz and Schneider, 1979; Silverman, 1986) into the discussion. Such an estimate is given by

$$\hat{f}(z) = N^{-1} w^{-1} \sum_{n=1}^{N} \varphi\{(z - Z_n)/w\} \tag{2}$$

where φ is the kernel function and $w>0$ is the window-width or smoothing parameter. φ is typically an even nonnegative function decreasing monotonically towards zero as its argument tends to $\pm\infty$, and having unit integral: for example, the standard normal density function. The Z_n are any data points; in our case they will be the projected multivariate data points $Z_n(a)$. In the cases proposed by Friedman and Tukey, (1) can actually be written in the form

$$d(a) = C(h) \sum_{m=1}^{N} \sum_{n=1}^{N} \varphi\{h^{-1} r_{mn}(a)\},$$

the normalising constant $C(h)$ being chosen so as to make φ a density. Since it is only relative values of d that are of interest, with h fixed during such comparisons, d can be replaced by any

$$d(a) \propto \sum_{n=1}^{N} \hat{f}\{Z_n(a)\} \tag{3}$$

where $\hat{f}$ is the kernel density estimate based on the (even, or symmetric) kernel φ with window width h. For example, if $g_h(r) = h - r$ with cutoff at $r = h$, the corresponding kernel is triangular on $[-h,h]$.

The Friedman-Tukey index now emerges as the result of constructing a kernel density estimate from the projected data points, and then summing its values at those projected data points. Rewriting (3) in a different notation gives

$$d(a) = \int \hat{f}(z) dF_N(z) \tag{4}$$

where F_N is the empirical distribution function of the projected data, and a definite scale for d has now been chosen. If the data are regarded as a sample from a multivariate distribution, then the projected data are a sample from its projection, and (4) is an estimate of $\int f(z) dF(z) = \int f^2(z) dz$ where F is the marginal distribution function, and f its density, for projection onto a.

Hodges and Lehmann (1956) showed that, for given mean 0 and variance 1, the density which minimises $\int f^2$ is given uniquely by $f(z) = \max\{0, c(b^2 - z^2)\}$ where $c = 3/(20\sqrt{5})$, $b = \sqrt{5}$. This is a parabolic density function, zero outside the interval $(-\sqrt{5}, \sqrt{5})$. A high value of a quantity approximately equivalent to $\int f^2$ accordingly corresponds to a large departure from parabolic form, and Friedman and Tukey's procedure is seeking any such departures rather than specifically looking for clustering.

2.3. *An Entropy Index*

Rényi (1961) introduced the general concept of order-α information or entropy for discrete distributions, and Rényi (1970, p. 592) gives the analogous definitions for continuous distributions; this is the relevant case here. The index $\int f^2$ is a simple and explicit monotone function of order-2 entropy. The usual (order-1) entropy measure $\int -f\log f$ immediately suggests itself as the basis for an alternative index $\int f\log f$. The density of mean 0 and variance 1 which minimises this is uniquely the standard normal density, a far more plausible candidate than the parabolic density as a norm from which departure is to be regarded as "interesting". Thus in proposing $\int f\log f$ as a projection index we are implementing the change of viewpoint anticipated at the start of Section 2, equating "uninteresting" with "normal" and choosing a particular way of measuring deviation from it. Note that if f is a standard normal density, $\int f\log f$ takes the value $-\frac{1}{2}\log 2\pi e$; this is an appropriate origin for the scale if one is needed. The use of order-1 entropy in this way was also suggested by Huber (1985), who pointed out that it was by no means the only functional uniquely optimised by the normal density: the Fisher information $\int (f')^2/f$ is another possibility; see also Jee (1985). To apply an entropy index in the sample case, it is necessary to use some form of density estimate, such as a kernel density estimate, in place of the true density f. This involves the choice of a window width parameter. The resultant density estimate will commonly have an inflated variance, and this will alter the numerical range for the index, an effect apparent in the values obtained in the example in Section 4. Once the window width has been chosen this inflation is fixed independently of the projection being inspected, and thus the operation of the projection pursuit algorithm is unaffected by any correction which may be made for it; such correction would be necessary only if index values obtained in different contexts were to be compared, an unusual procedure. Vasicek (1976) showed how to construct a test of normality based on entropy (via a nearest-neighbour density estimate rather than a kernel estimate), and demonstrated by simulation that its properties compare well with those of other tests of normality. Huber (1985) discusses the requirements for a projection index at length using the theme of departure from normality as the guideline; see also Diaconis and Freedman (1984).

2.4. *A Moment Index*

To optimise the entropy index, it is necessary to recalculate it at each step of the numerical procedure. We know of no method of obtaining the index via summary statistics of the multivariate data set, so the workload of the calculation at each iteration is determined by the number of points in the data set. For projection onto a plane, or to higher dimensions, the theory will be seen to carry over without difficulty, but the efficient one-dimensional density estimation techniques are lost. These considerations motivate us to obtain an approximation to the entropy index which can be calculated in terms of summary statistics of the multivariate data set and which will serve in extended form for projection to two or more dimensions at no more than a proportionate increase in workload.

The index we propose is based on third and fourth sample moments of the projected data set, along the lines of a suggestion made but not followed up by Huber (1985). These can be obtained from the third and fourth sample moment tensors of the multivariate data set, which can be calculated once-for-all from the data and then used without further reference back to the original data. We are of course very conscious of the reservations one must have about the use of higher multivariate moments, so it is important to emphasise that they are used here purely as a means of calculating the third and fourth moments of the projected data. Our data sets are large enough to make the projected moments relatively stable. It is convenient to think in terms of cumulants $\kappa_3=\mu_3$, $\kappa_4=\mu_4-3$ rather than moments μ_3, μ_4 (with $\mu_1=0$, $\mu_2=1$) since these cumulants are zero for the normal case; an index with any hope of tracking the entropy index must at least incorporate information up to the level of

symmetric departures from normality. The simplest such index is a positive-definite quadratic form in κ_3, κ_4. It must be invariant under sign-reversal of the data, and since κ_3 is odd and κ_4 even under such reversal, there can be no $\kappa_3\kappa_4$ term. Note that any index of this kind will not involve a choice of window width when used in the sample case; in a sense, the analogous choice is made by the decision to use only third and fourth moments. Suppose a density f is of the form $f(x) = \varphi(x)\{1+\varepsilon(x)\}$ where φ is the standard normal density function and ε satisfies

$$\int \varphi(x)\varepsilon(x)x^r dx = 0 \text{ for } r = 0,1,2. \tag{5}$$

Then for small ε, suitably well-behaved at $\pm\infty$,

$$\int f(x)\log f(x)dx \approx \tfrac{1}{2}\int \varphi(x)\varepsilon^2(x)dx \tag{6}$$

by using (5). The RHS of (6) is in itself a reasonable measure of departure from normality. Now if f is expressed as a Gram-Charlier expansion

$$f(x) = \varphi(x)\{1+\kappa_3 H_3(x)/6+\kappa_4 H_4(x)/24\ldots\} \tag{7}$$

(see Kendall and Stuart, 1977, p. 169) where H_r is the rth Hermite polynomial, then truncation of this expansion and use of orthogonality and normalisation properties of Hermite polynomials with respect to φ yields

$$\tfrac{1}{2}\int \varphi(x)\varepsilon^2(x)dx \approx (\kappa_3^2+\tfrac{1}{4}\kappa_4^2)/12. \tag{8}$$

As presented here this is a largely formal rather than analytical argument; the "approximation" will be used in cases as non-normal as possible, so the real question is whether the balance between the κ_3^2 and κ_4^2 terms which it suggests is effective in practice. This indeed appears to be the case.

Obviously an index based only on third and fourth moments is optimised by other distributions as well as the normal distribution. This is not a problem in practice: empirical data do not usually display very small third and fourth cumulants but large higher cumulants.

Huber (1985) cites results of Ferguson (1961) showing that tests based on skewness and kurtosis are most powerful for testing normality against the presence of outliers; this suggests that the moment index may be more liable than the entropy index to be driven by outliers, but there is no reason to suppose that other interesting projections will be lost sight of provided that local optima of the index are well explored.

2.5. *Projection to a Plane*

Generally speaking, it is obvious how to extend the one-dimensional index definitions to two or more dimensions; actual implementation of a projection pursuit method in more than one dimension is another matter. We deal explicitly with the practically useful planar case; projection to three or more dimensions raises no new questions of principle.

Entropy indices extend readily to the planar case, simply by taking integrals of the same functions of the density, where both the density itself and the integral are relative to Lebesgue (*i.e.* uniform) measure in the plane. This density is obtained by a bivariate kernel density estimation method, for example using a spherical bivariate normal kernel. Such indices are automatically invariant under the Euclidean group.

It is also straightforward to obtain a bivariate moment index. It is given by

$$\{(\kappa_{30}^2+3\kappa_{21}^2+3\kappa_{12}^2+\kappa_{03}^2)+\tfrac{1}{4}(\kappa_{40}^2+4\kappa_{31}^2+6\kappa_{22}^2+4\kappa_{13}^2+\kappa_{04}^2)\}/12 \tag{9}$$

where κ_{rs} is the bivariate cumulant of order (r,s) which (in the case of unit variance) is given by $\kappa_{rs}=\mu_{rs}$ for $r+s=3$, $\kappa_{40}=\mu_{40}-3$, $\kappa_{04}=\mu_{04}-3$, $\kappa_{31}=\mu_{31}$, $\kappa_{13}=\mu_{13}$, $\kappa_{22}=\mu_{22}-1$, where the μ's are bivariate central moments. The proof of the rotational invariance of this expression requires a little care. Further details may be found in Jones (1983).

3. IMPLEMENTATION

Projection pursuit is a multivariate constrained optimisation problem. For projection onto a line, the problem is that of selecting a K-vector, the projection direction, subject to one constraint, that it should be of unit length. For projection onto a plane, two K-vectors have to be selected, subject to three constraints to normalise them to unit length and make them orthogonal. For K typically ranging up to about 20, such problems are not likely to be difficult if the objective function is reasonably well-behaved. An effective optimisation method will however require the value and the gradient of the objective function to be calculated many times over, once for each iterative step, so efficiency is important.

3.1. *Computing Entropy Indices*

We consider first the one-dimensional case. The technique we propose for computing entropy indices is based on the use of a kernel density estimate $\hat{f}$ of f, given by (2). From this, two different estimates of the entropy index can be obtained:

$$\int \hat{f}(z) \log \hat{f}(z) dz \tag{10}$$

and

$$N^{-1} \sum_{n=1}^{N} \log \hat{f}\{Z_n(a)\} = \int \log \hat{f}(z) dF_N(z). \tag{11}$$

(10) is obtained by integrating $\log \hat{f}$ (numerically) with respect to the distribution with density $\hat{f}$, (11) by integrating it with respect to the empirical distribution function F_N of the projected data. We have found no guidance in the literature to establish any strong preference between (10) and (11), and accordingly propose to use the most convenient form in practice in any particular situation.

It is well-known (see, for example, Epanechnikov, 1969) that the properties of a kernel density estimate depend on the window width but that the exact form of the kernel is not important. Choice of a normal kernel, in particular, is quite inoffensive from the point of view of density estimation; it is a useful choice for applications in projection pursuit. One reason for this is that it guarantees smoothness as a function of projection direction; Jones (1983) gives details of the calculation of the derivative of the index with respect to projection direction for this case. A further advantage of the normal kernel is that it allows the use of an efficient algorithm for computing density estimates for large data sets, due to Silverman (1982): the convolution of regularly gridded approximations to the data and the kernel is achieved by using a Fast Fourier Transform algorithm. A discretisation procedure for the data which offers enhanced accuracy and smoothness by comparison with the crudest such procedure has been discussed by Jones and Lotwick (1983, 1984). Silverman's algorithm has a workload dependent linearly on the number N of data points at the discretisation stage, and is thereafter dependent only on the number of grid points, thus giving a reasonably efficient treatment of large data sets. The gridded output from it can be used very conveniently for the numerical evaluation of (10). Although efficient, this procedure still requires an order N calculation for each projection direction explored in the course of the optimisation procedure. Direct calculation of (11) is order N^2 in general; by using a kernel of bounded support, rather than a normal kernel, a preliminary sorting operation (order $N \log N$) can be used to leave an order N calculation.

The sample versions of the index based on order-2 entropy corresponding to (10) and (11) are

$$\int \hat{f}^2(z) dz \tag{12}$$

and (4) respectively; numerical integration is again required to evaluate (12). If the window width in (4) is $\sqrt{2}$ times that in (12), and the kernel is normal, then (4) and (12) coincide (Bhattacharyya and Roussas, 1969). This does not happen with (10) and (11), but it may still

be appropriate to rescale the window width when moving between them (Jones, 1983). Choice of window width is discussed in Section 3.4.

The direct computation of the two-dimensional analogue of (11) is only a little more laborious than in one dimension, but this only covers the case of comparatively small data sets. Unfortunately, the grid-based technique for calculating (10) does not extend usefully to the planar case, because of the large number of grid points required and the lack of a bivariate Fourier transform procedure comparable in efficiency to the univariate Fast Fourier Transform.

3.2. *Computing Moment Indices*

This computation is very similar for projection to a line or to a plane. Once the data have been centred, scaled, and sphered, the third and fourth outer product tensors are calculated. They are given by

$$\left.\begin{aligned} T_{ijk} &= \sum_{n=1}^{N} X_{ni}X_{nj}X_{nk} \\ U_{ijkl} &= \sum_{n=1}^{N} X_{ni}X_{nj}X_{nk}X_{nl} \quad . \end{aligned}\right\} \tag{13}$$

Any further calculations do not depend on N; T_{ijk} and U_{ijkl} are the appropriate summary statistics of the data. However, they have many distinct components, despite being totally symmetric: $K(K+1)(K+2)/6$ components for T_{ijk}, and $K(K+1)(K+2)(K+3)/24$ for U_{ijkl}. So they only constitute a "summary" in a computational sense if

$$K(K+1)(K+2)/6 + K(K+1)(K+2)(K+3)/24 < KN,$$

that is,

$$(K+1)(K+2)(K+7)/24 < N. \tag{14}$$

For example, with $K=10$ the method is likely to display computational economies from about $N=100$; clearly any attempt to incorporate moments of higher orders will quickly lose any computational benefits, and these in any case diminish rapidly as K increases. In the distributional case the cumulants are defined unequivocally; in the sample case they are being estimated, just as is the density for the entropy indices. The usual k-statistics (Kendall and Stuart, 1977) will be used as estimators, since attempts to estimate the squared cumulants by unbiased estimators would involve access to higher order moments — up to order 8 for third and fourth order squared cumulants (Wishart, 1952). Third and fourth order cumulants of the projected data are obtained straightforwardly by way of contraction of the outer product tensors with respect to the line or plane of projection. Derivatives with respect to the line or plane of projection can also be obtained without substantial further computation; details are given by Jones (1983).

3.3. *Optimisation*

In most optimisation problems, the global optimum is required and non-global local optima are a nuisance. In projection pursuit the aim is to explore all local optima which are not too severely sub-optimal globally, so what is usually a problem with "hill-climbing" methods of optimisation actually works advantageously here. It is also the case that optima are not required to very high accuracy, because the appearance of the projected data changes smoothly and not too rapidly with projection direction.

The optimisation problems are actually constrained:

$$\left.\begin{array}{l}\text{minimise } I(a) \text{ subject to } a^{\mathrm{T}}a = 1; \\ \text{or,} \\ \text{minimise } I(a,b) \text{ subject to } a^{\mathrm{T}}a = 1,\ b^{\mathrm{T}}b = 1,\ a^{\mathrm{T}}b = 0.\end{array}\right\} \tag{15}$$

When projection indices were defined in Section 2, the constraints were assumed to be satisfied. In practice it is convenient to extend the definition of I to arbitrary arguments by setting $I(a) = I(\tilde{a})$ where $\tilde{a}^{\mathrm{T}}\tilde{a} = 1$ and $\tilde{a}$ defines the same line as a, and $I(a,b) = I(\tilde{a},\tilde{b})$ where $\tilde{a}^{\mathrm{T}}\tilde{a} = 1$, $\tilde{b}^{\mathrm{T}}\tilde{b} = 1$, and $\tilde{a}^{\mathrm{T}}\tilde{b} = 0$, and where $\tilde{a},\tilde{b}$ define the same plane as a,b. The only requirements on the arguments are then $a \neq 0$ for projection to a line, and $a \neq 0$, $b \neq 0$, a and b not parallel for projection to a plane. This ensures that the gradient of I is tangential to the feasible region, and hence that drift away from this region is slow during optimisation and can be corrected by an occasional re-orthonormalisation.

Adby and Dempster (1974) give an account of a wide variety of multivariate optimisation methods, and emphasise the importance of using gradient information if at all possible. The simplest such methods are those which iterate in the direction of the steepest slope. One such method was suggested by J.B. Kruskal (1964) for use in nonmetric multidimensional scaling, and we have found that a slight adaptation of that technique, designed to overcome its tendency to cycle near the optimum, is quite well-suited to use in projection pursuit; details are given in Jones (1983). Steepest-slope methods are generally appropriate for the approximate location of the local optima of a function which is not particularly well-approximated by a quadratic. This is precisely what is needed here, and it is doubtful if there is much benefit from using the more sophisticated conjugate-gradient methods, which are well-suited to the high-accuracy location of an optimum near which quadratic approximation is good. However, some further experiments with such methods might well be of interest.

For one-dimensional projection pursuit, Friedman and Tukey (1974) suggested the use of a complicated transformation to give an effectively unconstrained problem in $K-1$ variables. We can see no particular merit in this. For two-dimensional projection pursuit it is far from clear what their optimisation strategy actually is, but there is some indication that it is computationally much more laborious than in the one-dimensional case. By contrast, the method outlined here extends straightforwardly, the workload increasing in line with the greater number of variables which are involved.

3.4. *The Window Width*

The choice of a window width is a prerequisite for the construction of the kernel density estimate used to obtain the entropy index. Bowman (1985) and Silverman (1986) give references to the extensive literature on the choice of window width for the estimation of the density as a whole; there is little guidance available for the case where the density estimate is to be used to estimate some particular functional. We find empirically that the entropy index is comparatively insensitive to window width, although it is desirable to track the theoretically optimal dependence of window width on sample size. Thus we suggest the use of a window width $w=N^{-0.2}$ in the notation of Section 2.2 for (10) and (12), and $\sqrt{2}$ times this for (11) and (4). This choice works well in practical examples, but obviously in any implementation of projection pursuit the user should be able to override it if desired.

3.5. *Starting and Stopping the Search*

Useful starting-points include the principal axes of the original data (for projection to a line; pairs of them for projection to a plane), and, of course, random starts. Repeated runs are necessary to explore the local optima. Optima with small domains of attraction, if such actually occur, are inevitably likely to be missed, but we have found that in empirical

examples a rather limited number of distinct optima actually recur from repeated random and principal-axis starts, and offer a useful selection of informative views of the data.

3.6. *Computational Load*

The computational load of carrying out data exploration by projection pursuit is primarily governed by the number of separate runs made in the search for local optima; this is as much dependent on the structure of the data as on its scale. The number of iterations needed to locate a local optimum may vary considerably from run to run on a given set of data. Thus only the most general guidelines are meaningful. Our experience is that, provided the appropriate version of projection pursuit for the circumstances is chosen, typical runs take about 15 seconds on a processor capable of executing 1 million Whetstone instructions per second. Storage requirements are linear in the amount of data. The overall picture is that under current circumstances projection pursuit can comfortably be used on an interactive basis except perhaps on small microcomputers.

4. EXAMPLE

Tables 4, 5, and 6 of Lubischew (1962) give 6 measurements on individual male flea-beetles of three species (which we refer to as A, B, and C) of the genus *Chaetocnema*. There are 74 cases in total. We make no use of the species labelling in our analysis, but it is of course interesting to compare that labelling with any structure found in the data by projection pursuit or any other method. All projections are of the sphered scaled centred data.

The case of projection to a line is considered first, with form (10) of the entropy index used as the projection index and denoted by e, and a window width as suggested in Section 3.4 of $w=74^{-0.2}=0.4228$. For this particular data set, the principal components of the scaled centred data offer reasonably interesting views of the data, so it is natural to include these among the possible starting points for projection pursuit. Comparison between these starting points and the local optima found from them is given in Figures 1, 2, and 3 for the first three principal axis directions of the scaled centred data; species labels have been added afterwards to facilitate an assessment of the extent to which the species have emerged naturally as point clusters in the data, and of the extent to which these clusters emerge from the method. There seems little room to doubt that projection pursuit is being driven by the species separation, with the views in Figures 1(b), 2(b), and 3(b) respectively splitting off species B, species C, and, rather less successfully, species A, from the other two. Incidentally, view 1(b), although started from the direction of the principal axis corresponding to the largest principal component, is not the projection of highest index; view 2(b), started from the second principal axis, has higher index.

Lubischew's beetle data is obviously a well-structured data set, and it is not surprising that this should show up in the first few principal axis views, and, indeed, in one or two of the original variates, although it is not easy to see much (except in the projection onto the first principal axis) *until the species labels are added.* Projection pursuit gives a persuasive separation of the data points into clusters *without* the use of the labels, and they subsequently confirm that it is the distinct species that have been separated off.

Canonical variates analysis, which *does* make use of species information, identifies a projection to a line which separates all three species into distinct clusters. This projection is shown in Figure 4. The projection obtained by Ottestad (1975) by discriminant analysis, again using species information, is fairly similar. It is fortuitous for projection onto a line to be able to separate more than two clusters. The three-cluster projection shown in Figure 4 has comparatively low index and is not at or near a local optimum of the index. One reason for this is that an approximate 2:1 split of the data points is more non-normal than a 1:1:1 split. This extends to data with larger numbers of clusters: they are likely to be split off one at a time, an effect also reported by Friedman and Tukey (1974). Similarly, projection onto a plane is likely to show up three clusters at a time, or to split off two from the residue.

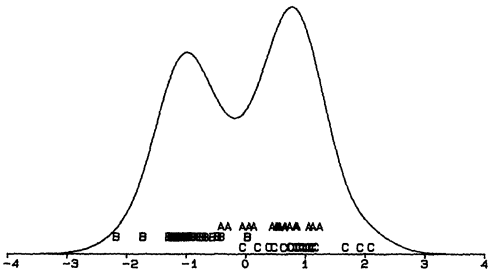

Fig. 1(a). View along first principal component, $e = -1.44$.

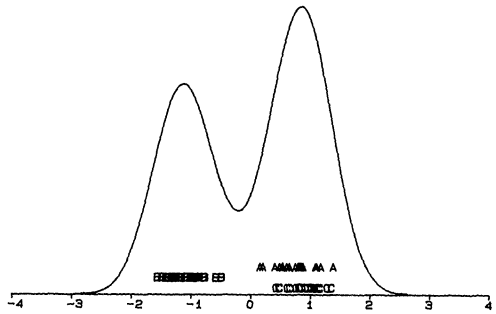

Fig. 1(b). Corresponding projection pursuit solution, $e = -1.35$.

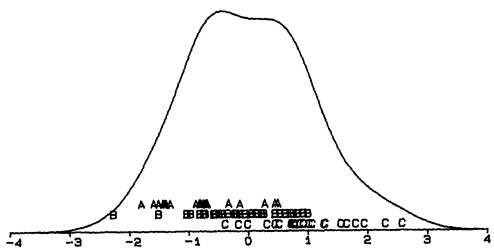

Fig. 2(a). View along second principal component, $e = -1.49$.

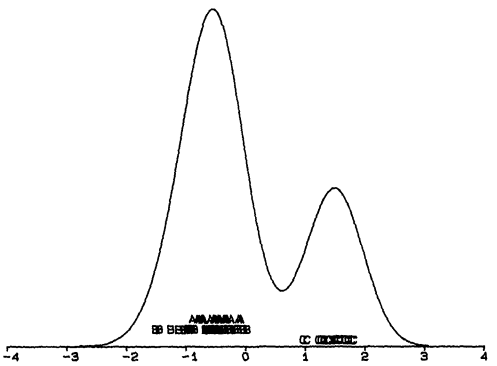

Fig. 2(b). Corresponding projection pursuit solution, $e = -1.32$.

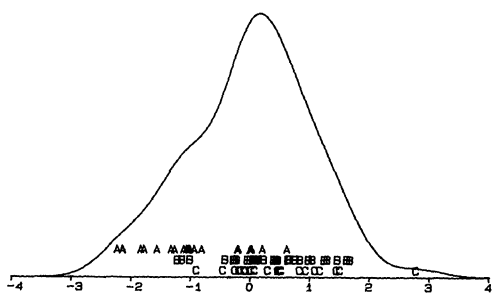

Fig. 3(a). View along third principal component, $e = -1.48$.

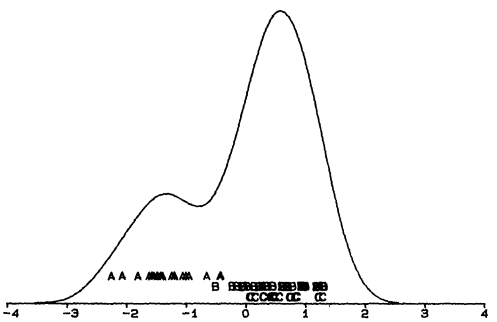

Fig. 3(b). Corresponding projection pursuit solution, $e = -1.41$.

Figures 5 and 6 shows views corresponding to other local optima of the index: Figure 5 suggests two points as possible outliers, and Figure 6 is simply not very interesting. The choice of window width affects the number of local optima of the index: too large a value causes loss of interesting views, and too small a value swamps the user with too many uninteresting views.

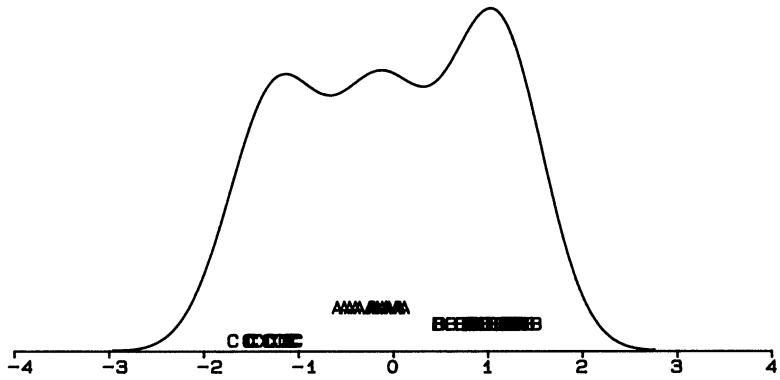

Fig. 4. View along the first canonical variate, all three species separated, $e = -1.43$.

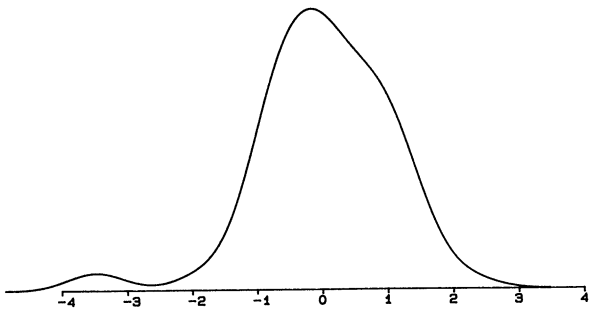

Fig. 5. A solution exhibiting possible outliers, $e = -1.45$.

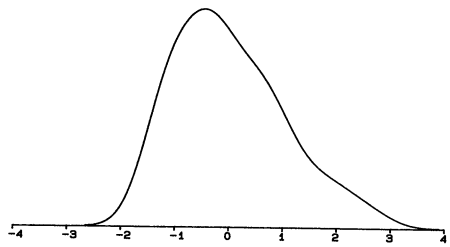

Fig. 6. A solution showing little of interest, $e = -1.43$.

We now illustrate projections to a plane, presented as scatter plots with the origin and unit circle shown. The moment index, denoted by m, has been used in this case for projection pursuit. Figure 7(a) shows projection to the plane of the first two principal axes; there is little more obvious indication of structure than is given separately by the first two principal axis projections, although if species labels were added it would be clear that there is some structure there. Figure 7(b) is the projection obtained by optimising the moment index from this starting-point, and even without the species labels it leaves little room for doubt about the structure of the data, perhaps with some ambiguity over the three A-points near the origin. Figure 8 is the projection defined by the first two canonical variates; for convenience of comparison, it has been matched to Figure 7(b) by procrustean fitting (see Sibson, 1978; the minimised average squared positional difference $G_E/74$ is 0.046). Although there is some visible difference between Figures 7(b) and 8, for example in the improved location of the A-points near the origin, the closeness of the correspondence is very striking; again, it must be emphasised that the species labelling is not used in the construction of Figure 7(b) by projection pursuit, but is used in the construction of Figure 8 by canonical variates analysis. Because in two dimensions it is reasonable to expect three clusters to be resolvable, there need be no inhibitions over making direct comparisons of these two projections. Other planar projections, such as that shown in Figure 9, identify possible outliers, but there is no suggestion of any further structure of real interest.

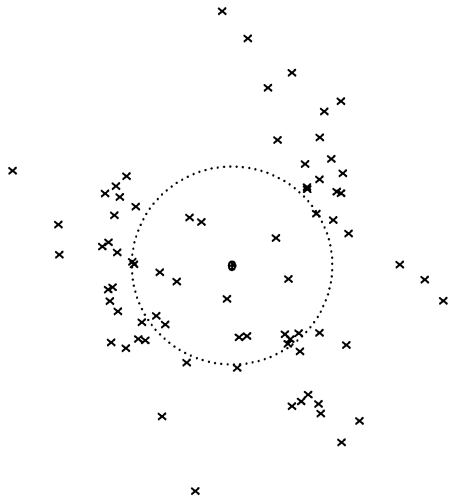

Fig. 7(a). View in the plane defined by the first two principal components, $m = 0.90$.

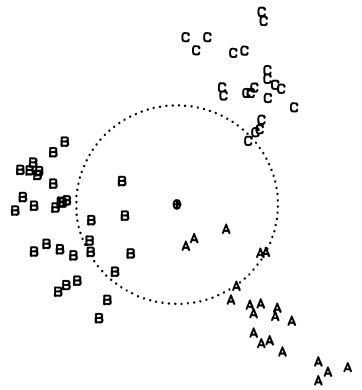

Fig. 7(b). Corresponding projection pursuit solution, $m = 2.53$.

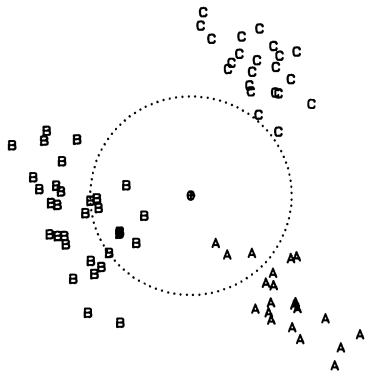

Fig. 8. View in the plane defined by the first two canonical variates, $m = 2.36$.

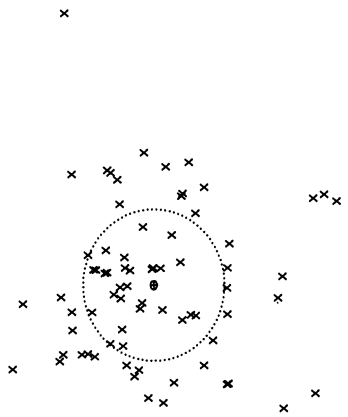

Fig. 9. Another two-dimensional solution, $m = 2.40$.

5. FURTHER DISCUSSION

5.1. *Outliers*

The presence of outliers in a sample gives it the appearance of non-normality, and the various projection indices we have proposed in Section 2 will reflect this, perhaps to varying extents. Thus projection pursuit in the form presented in this paper will tend to *identify* outliers, as illustrated in Section 4, and these may simply be *noted*, or, if they dominate to the extent of masking other structure in the data, they may be *discarded*. An alternative approach to outliers — see Barnett and Lewis (1984) for a general discussion — is to design procedures which will *accommodate* them. In the case of projection pursuit, this means that the effect of outliers has to be damped so that it does not obscure the cluster or other interesting structure which is sought. The robustification adopted by Friedman and Tukey

(1974) had this objective. Our own preference for identifying rather than accommodating outliers is a deliberate one, but there is nothing in our approach to projection pursuit which is inconsistent with the introduction of robustification in a variety of ways. Accommodation of outliers should be incorporated in such a way that the projection index is still minimised to a fixed value at normality in an appropriate sense, with, as a minimal requirement, large-sample behaviour converging to that for distributions; and computational efficiency must not be compromised to the point where the method loses its attractions as an exploratory tool.

5.2. *Isolation*

The technique, called "isolation" by Friedman and Tukey (1974), of splitting off clusters for separate "internal" analysis has been used extensively in cluster analysis in general, and in other exploratory cluster-revealing methods such as multidimensional scaling. It is particularly important in projection pursuit because of the inability of low-dimensional projections to display more than a few clusters at once; the example of Section 4 shows how projection to a line tends to split off one cluster from a residue, although different clusters may be split off by different locally optimal projections. Although we do not report the details here, we have applied the isolation approach to the analysis of the Lubischew beetle data; the three species clusters emerge readily, and the cluster-by-cluster analysis reveals nothing further apart from a few possible outliers.

5.3. *Scope of the Method*

Jones (1983) reports results on a data set of 3000 cases in 10 dimensions; this is not by any means the limit of what is practicable, and indicates that there is a wide range of multivariate data sets whose size makes them amenable to investigation by projection pursuit, at least if the moment index is used. It is the value of the dimensionality K which is the limitation rather than the number N of cases, and for large K it may be necessary to effect a preliminary reduction of dimensionality by, for example, using principal components analysis and discarding small components. Large data sets make it most apparent that the basic simplicity of projection pursuit is both a strength and a weakness: a strength in that it is very easy to explain to non-expert "clients" what the method is doing, at least in general terms; and a weakness in that it is really good only at displaying structure of a comparatively simple and clear-cut kind, and can do this only through the limitations of orthogonal projection. Also, it is in its present form confined to cases where each variate is on a numerical scale; Diaconis (1983) has investigated the possibility of applying projection pursuit to wholly discrete data, but it is not as yet clear how to handle the common case of mixed continuous and discrete data.

5.4. *Sampling Fluctuation*

Day (1969) exhibits apparent clustering structure in one-dimensional projections of data generated from a multivariate normal distribution; obviously it is important in projection pursuit to avoid ascribing importance to apparent structure which could easily be the result of sampling fluctuation, even though the method is being used for exploratory purposes rather than in a formal statistical context. Unlike most exploratory procedures, projection pursuit has at its heart a perfectly conventional simple null hypothesis, that of the K-dimensional spherical normal distribution, and this does make it possible to simulate data against which the observed data may be compared. It is misguided to try to put an exploratory procedure on the same footing as a significance test, but it is certainly possible, and potentially useful, to calibrate a projection index from such simulations.

5.5. *Density Estimation and Regression*

Table 1 presents a comparison between the basic technique of projection pursuit as developed in the present paper, and two related but distinct techniques mentioned earlier: projection pursuit regression (Friedman and Stuetzle, 1981); and projection pursuit density

TABLE 1

A comparison of projection pursuit procedures

PROJECTION PURSUIT EXPLORATORY DATA ANALYSIS	PROJECTION PURSUIT DENSITY ESTIMATION (Friedman, Stuetzle, and Schroeder, 1984)	PROJECTION PURSUIT REGRESSION (Friedman and Stuetzle, 1981)
Data: X_n, $n = 1,\ldots,N$	Data: X_n, $n = 1,\ldots,N$ and initial density estimate $f_0(X)$	Data: X_n, Y_n, $n = 1,\ldots,N$ and initial model a constant
Project data onto one dimension $Z_n = a^{\mathrm{T}}X_n$	Project data onto one dimension $Z_n = a^{\mathrm{T}}X_n$	Project independent variable onto one dimension $Z_n = a^{\mathrm{T}}X_n$
Calculate (univariate) non-parametric density estimate, $\hat{f}$, of projected points	Calculate (univariate) non-parametric density estimate, $\hat{f}$, and (estimate of) marginal density, $\hat{g}$, of current multivariate density	Calculate (univariate) non-parametric regression of current residuals on Z's
Compute index* *e.g.* sample entropy $N^{-1}\sum\log\hat{f}(Z_n)$	Compute index* *e.g.* sample relative entropy $N^{-1}\sum\log\{\hat{f}(Z_n)/\hat{g}(Z_n)\}$	Compute index* *e.g.* fraction of unexplained variance $1 - \sum\{r_n - S_a(Z_n)\}^2/\sum r_n^2$ r_n : residuals (=Y_n first time) S_a : non-parametric regression
Optimise over choice of a	Optimise over choice of a	Optimise over choice of a
Gives: (locally) "most interesting" view of data	Gives: univariate density "most different" from current model	Gives: univariate regression "most different" from current model
	Combine: multiply current model by optimal $\hat{f}/\hat{g}$	Combine: add optimal regression to current model
Iterate: start again from new initial projection	Iterate: take updated model and go round again	Iterate: recalculate residuals, take updated model and go round again

* In each case, this and the preceding step could be combined as "Compute index"; we note here, though, that each implementation involves an appropriate non-parametric curve estimate.

estimation (Friedman, Stuetzle, and Schroeder, 1984). We concentrate on the case of projection to a line for the purpose of making this comparison. There are undoubtedly numerous technical and computational considerations which will arise with the latter two techniques to a greater extent than with projection pursuit itself. We believe that the exact choice of smoothing parameter and the application of a "stopping rule" are likely to be among these considerations.

It is clear that the latter columns each follow the "projection pursuit paradigm" given by Friedman and Stuetzle (1982). This is as follows.

(i) choose an initial model

Repeat:

(ii) find a projection that shows deviation of the data from the current model (the projection pursuit step)

(iii) change the model to incorporate the structure found in (ii) (update the model)

Until:

(iv) the current model agrees with the data in all projections.

As indicated above and in Table 1, it is essentially only stage (ii) of this process that we perform when using projection pursuit in the sense of the present paper for exploratory data analysis.

6. CONCLUSIONS

The range of available exploratory multivariate techniques is still dominated, as pointed out by Sibson (1984), by essentially algebraic "second-order" methods based on the singular value decomposition, and nonparametric variants of such methods. The value of such techniques is undisputed; we regard it as equally indisputable that their coverage is insufficient. The attraction of projection pursuit is that despite numerous limitations it is completely distinct from other methods, and exploits a simple basic idea to surprisingly good effect. Such a method must surely merit a welcome from applied statisticians.

ACKNOWLEDGEMENTS

We thank Professors J.B. Copas and B.W. Silverman, and Drs A. Bowman, A.M. Herzberg, and H.W. Lotwick, for helpful conversations, suggestions, and comments. We are grateful to the referees for numerous constructive remarks, to which we have been glad to respond. M.C. Jones was supported by an SSRC Research Studentship and later by a University of Bath Research Fund Studentship whilst the work reported in this paper was being carried out.

REFERENCES

Adby, P.R., and Dempster, M.A.H. (1974) *Introduction to Optimisation Methods*. London: Chapman and Hall.

Anderson, E. (1935) The irises of the Gaspe peninsula. *Bull. Amer.* Iris *Soc.*, **59**, 2-5.

Barnett, V., and Lewis, T. (1984) *Outliers in Statistical Data*. 2nd. ed. Chichester: Wiley.

Bhattacharyya, G.K., and Roussas, G.G. (1969) Estimation of a certain functional of a probability density function. *Skand. Aktuarietidskr.*, **52**, 201-206.

Bowman, A.W. (1985) A comparative study of some kernel-based nonparametric density estimators. *J. Statist. Comput. Simulation*, **21**, 313-327.

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

Diaconis, P. (1983) Projection pursuit for discrete data. Technical Report 198, Stanford University Department of Statistics.

Diaconis, P. and Freedman, D. (1984) Asymptotics of graphical projection pursuit. *Ann. Statist.*, **12**, 793-815.

Epanechnikov, V.A. (1969) Non-parametric estimation of a multivariate probability density. *Theory Probab. Appl.*, **14**, 153-158.

Ferguson, T.S. (1961) On the rejection of outliers. In *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1* (J. Neyman, ed.) pp. 253-287. Berkeley: University of California Press.

Fisher, R.A. (1936) The use of multiple measurements in taxonomic problems. *Ann. Eugenics*, **7**, 179-188.

Friedman, J.H. (1985) Exploratory projection pursuit. Technical Report 18, Laboratory for Computational Statistics, Department of Statistics, Stanford University.

Friedman, J.H., and Stuetzle, W. (1981) Projection pursuit regression. *J. Amer. Statist. Assoc.*, **76**, 817-823.

—— (1982) Projection pursuit methods for data analysis. In *Modern Data Analysis* (R.L. Launer and A.F. Siegel, eds.) pp. 123-147. New York: Academic Press.

Friedman, J.H., Stuetzle, W., and Schroeder, A. (1984) Projection pursuit density estimation. *J. Amer. Statist. Ass.*, **79**, 599-608.

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

Fryer, M.J. (1977) A review of some nonparametric methods of density estimation. *J. Inst. Math. Appl.*, **20**, 335-354.

Hodges, J.L., and Lehmann, E.L. (1956) The efficiency of some non-parametric competitors of the t-test. *Ann. Math. Statist.*, **27**, 324-335.

Huber, P.J. (1981) Projection pursuit. Research Report PJH-6, Harvard University Department of Statistics.

—— (1985) Projection pursuit (with discussion). *Ann. Statist.*, **13**, 435-525.

Jee, J.R. (1985) A study of projection pursuit methods. Technical Report TR 776-311-4-85, Rice University.

Jones, M.C. (1983) *The Projection Pursuit Algorithm for Exploratory Data Analysis*. PhD Thesis, University of Bath.

—— (1985) Contribution to the discussion of Huber (1985). *Ann. Statist.*, **13**, 508-510.

Jones, M.C., and Lotwick, H.W. (1983) On the errors involved in computing the empirical characteristic function. *J. Statist. Comput. Simulation*, **17**, 133-149.

—— (1984) Remark ASR50. A remark on Algorithm AS176: Kernel density estimation using the fast Fourier transform. *Appl. Statist.*, **33**, 120-122.

Kendall, M.G., and Stuart, A. (1977) *The Advanced Theory of Statistics. Volume 1: Distribution Theory*. 4th ed. London: Griffin.

Kruskal, J.B. (1964) Nonmetric multidimensional scaling: a numerical method. *Psychometrika*, **29**, 115-129.

—— (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.) pp. 427-440. New York: Academic Press.

—— (1972) Linear transformation of multivariate data to reveal clustering. In *Multidimensional Scaling: Theory and Applications in the Behavioural Sciences, Volume 1* (R.N. Shepard, A.K. Romney, and S.B. Nerlove, eds.) pp. 179-191. London: Seminar Press.

Lubischew, A.A. (1962) On the use of discriminant functions in taxonomy. *Biometrics*, **18**, 455-477.

Ottestad, P. (1975) Discrimination analysis. *Internat. Statist. Rev.*, **43**, 301-315.

Rényi, A. (1961) On measures of entropy and information. In *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1* (J. Neyman, ed.) pp. 547-561. Berkeley: University of California Press.

—— (1970) *Probability Theory*. Amsterdam: North-Holland.

Rosenblatt, M. (1956) Remarks on some nonparametric estimates of a density function. *Ann. Math. Statist.*, **27**, 832-837.

Sibson, R. (1978) Studies in the robustness of multidimensional scaling: procrustes statistics. *J. R. Statist. Soc.* B, **40**, 234-238.

—— (1984) Present position and potential developments: some personal views. Multivariate analysis (with discussion). *J. R. Statist. Soc.* A, **147**, 198-207.

Silverman, B.W. (1982) Algorithm AS176. Kernel density estimation using the fast Fourier transform. *Appl. Statist.*, **31**, 93-99.

—— (1986) *Density Estimation for Statistics and Data Analysis*. London: Chapman and Hall.

Tukey, P.A., and Tukey, J.W. (1981a) Preparation; prechosen sequences of views. In: *Interpreting Multivariate Data* (V. Barnett, ed.) Chichester: Wiley. 189-213.

—— (1981b) Data-driven view selection; agglomeration and sharpening. In: *Interpreting Multivariate Data* (V. Barnett, ed.) Chichester: Wiley. 215-243.

Vasicek, O. (1976) A test for normality based on sample entropy. *J. R. Statist. Soc.* B, **38**, 54-59.

Wertz, W., and Schneider, B. (1979) Statistical density estimation: a bibliography. *Internat. Statist. Rev.*, **47**, 155-175.

Wishart, J. (1952) Moment coefficients of the k-statistics in samples from a finite population. *Biometrika*, **39**, 1-13.

DISCUSSION OF THE PAPER BY DR JONES AND PROFESSOR SIBSON

Mr J. C. Gower (Rothamsted Experimental Station): As the authors' references bear witness, projection pursuit (PP) has been discussed in the American literature since 1974 but, apart from the Tukeys' sallies to some of Professor Barnett's triennual conferences in Sheffield and occasional seminars, the topic has hitherto received very little discussion in this country. We are about to put that right and I welcome tonight's paper for the opportunity it gives the Society.

Dr Jones and Professor Sibson present PP as a three-stage process: (1) scale the data, (2) find local minima of some index of un-interestingness among projections onto a line or plane, (3) present and interpret the results. Thus we are in the general area of multidimensional scaling; that is exploring a cloud of points set in a multidimensional Euclidean space, in the hope of finding interesting structures such as clusters, outliers and smooth manifolds. I shall comment on each of the three stages but note in passing that PP proper is really concerned only with the second stage.

Scaling is a problem with most multivariate quantitative methods that combine information on disparate variables (e.g. height and weight). The popular method of normalising the centred data by dividing by standard errors seems to be influenced by notions of unimodal, preferably normal (and hence *uninteresting*), distributions. Yet when we are concerned with mixtures, as in the distributional formulation of clustering problems, or with outliers, then the normalisers are much influenced by sampling vagaries such as the proportions of samples in the different mixture components or the extent of the outlier(s). There are no hard and fast rules, but I often prefer to deal with incommensurable ratio-scales by taking logarithms of the uncentred data—this handles the scaling problem independently of sampling fluctuations and gives values that are invariant to the acquisition of information on additional samples; the role of outliers is diminished but so is it with normalisation.

These misgivings are increased with sphering. The variance–covariance matrix $XX' = S$ (say) is similarly sensitive to mixture proportions and outliers; the choice of a "Mahalanobis" distance based on $X'S^{-1}X$ is influenced by the multinormal case and additionally S may be ill-conditioned. $X'S^{-1}X$ is idempotent so that projections onto all directions have equal variance; rather than consider this as an advantage are we perhaps sacrificing too much of the structure of the sample? Some insight can be obtained by thinking of three clusters at the vertice of a triangle (which for convenience I have chosen to be isosceles in Fig. D1).

Projections onto the principal axis show only two clusters, while projections onto the orthogonal axis (with smallest sum-of-squares) show all three clusters. Thus the PP notion of considering projections onto other than principal axes is worth pursuing but interesting projections can be found without sphering and sub-principal axes should not be ignored. If we assume that the clusters contain the same number of samples, sphering of *any* triangle will effectively put the three clusters at the vertices of an equilateral triangle and some projections will show three clusters and others two clusters just as in PCA and as PP has done in Figs 1(b), 2(b) and 3(b) of the paper. That the really interesting projection of Fig. 4 gives a poor value of the PP criterion e calls into question its effectiveness. The criterion seems to pick out the more skew projections and rejects the more symmetric projections, presumably because they are more approximately normal and hence deemed uninteresting. Indeed with the set-up considered here, the fourth moment too is constant so e becomes a function of skewness alone. When the numbers of samples per cluster differ the spherised triangle will not be equilateral and U-shaped projection densities giving the three clusters, the one with smallest sample size in the middle, will surely be picked up by PP; a cluster-finding technique should not be sensitive to numbers of samples in the clusters.

With c equal-sized clusters, the spherised version will give clusters at the vertices of a regular simplex with c vertices (or a projection of a simplex onto $K < c$ dimensions), unless the clusters lie in essentially fewer than c–1 dimensions, in which case S becomes ill-conditioned (e.g. sphering will obscure the simple geometry of $c = 3$ collinear clusters). Provided S is not ill-conditioned, it seems that sphering has the effect of separating cluster-centres and, provided one is not interested in metric information, this could be useful with any k-group (but not hierarchical) clustering method. However, as the authors point out, the effects of sphering are unaltered by orthogonal transformations, so we may assume that X is referred to its principal axes with $S = \Lambda$ (diagonal, eigenvalues). Sphering is now seen to replace X by $\Lambda^{-1/2}X$, exaggerating the distances in previously unimportant dimensions. Thus if in the triangle example, the cluster had little variation outside the plane shown in the figure, after sphering they would scatter much more in the space orthogonal to the spherised equilateral triangle and this could have an adverse effect on detecting clusters, countering the good effects of separating the cluster centres. Though useful structure can survive sphering, I think it can be harmful.

Curiously the authors normalise before sphering. Thus X becomes DX where D is the diagonal matrix

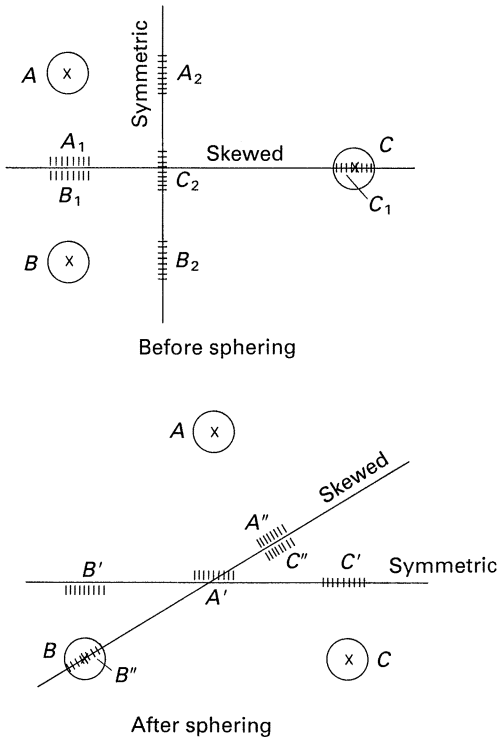

Fig. D1. The clusters labelled *A*, *B*, *C* are assumed to have some dispersion orthogonal to the plane of the paper.

of normalisers. Sphering now gives coordinates $(DXX'D)^{-1/2}DX$ but the sample interdistances and hence all projections, are derived from the same matrix $X'S^{-1}X$ as without normalisation, so why bother?

Often scaling is handled better by choosing a suitable distance-function, and sticking to it. This opens the way to using PP on all types of data (quantitative, categorical or mixtures of them etc, see Section 5.3) that permit the calculation of Euclidean dissimilarity coefficients (see Gower and Legendre, 1986) giving distances generated by a cloud of points which could, though I believe unwisely, be sphered. Fortunately sphering is not essential to PP, though if used, it has one important consequence to be discussed next.

Coming to stage 2, the actual examination of projections, and noting that the sphered projections have equal sums-of-squares in all directions, it is no surprise that a suitable criterion of uninterestingness then does not involve the second moment and is essentially a function of the third and fourth cumulants. Without sphering, the second moment could be considered too and perhaps it should be. I shall not comment in detail on the various forms of PP criteria discussed in the paper, interesting as they are, but note that the starting point in their derivations is that the yardstick of uninterestingness is the normal distribution. Presumably the multinormal distribution is the acme of uninterest—so much for classical multivariate analysis. I have very little quarrel with this point of view but would have thought that the uniform distribution is also pretty uninteresting and perhaps the same goes for any unimodal distribution. The various forms of the authors' criteria are certainly worthy of consideration but the assumptions underlying their mathematical justifications do not preclude consideration of other criteria; perhaps different criteria are needed for detecting different kinds of pattern.

The interpretation of projections is not straightforward. Striking projected patterns do not necessarily arise from interesting features of the data; interesting features of the data do not necessarily generate informative projections; real projected patterns may be hard to interpret. I recall asking John Tukey, after he had given a talk on *PRIM*-9, his program for exploring 9-dimensional space, how one would go about interpreting the projections from data which were the nine values for each of several thousand 3×3 orthogonal matrices. Because sums-of-squares of all rows and columns are unity and their

cross-products zero, then various two-dimensional cross-sections of the 9-dimensional space will show circles and hyperbolae whose sizes fluctuate depending on the cross-sectional planes used. These curious pyenomena will be picked up as projections but how to understand what is going on is a problem; what do projections of other simple manifolds look like?

In the examples the values of the criterion e are often surprisingly close. These are analogous to the eigenvalues of a components analysis and I wonder if the same kind of instability arises in PP as it does with near-equal eigenvalues in PCA. In what sense are all the local optima of interest? How many of them are there? In further analogy with PCA, might it be worth maximising uninterest in the space orthogonal to the line or plane of projection? So far as the detection of clusters goes, the obvious question to ask is what, if any, advantage PP has over standard methods of cluster analysis? No doubt the authors have views on such questions and have been constrained by lack of journal space; I look forward to their remarks in response to this discussion.

That I have reservations about using PP is not so much related to the potential usefulness of the method but more with the need for more work. PCA, 90 years after its first appearance, is still a much misused and misunderstood multivariate technique. I guess that work on PP will continue for many years and I suspect the authors would not disagree. Indeed, as they write very fairly in their closing remarks, "The atttraction of projection pursuit is that despite numerous limitations it is completely distinct from other methods, and exploits a simple basic idea to surprisingly good effect. Such a method must surely merit a welcome from applied statisticians." Surely it must. I have much pleasure in proposing the vote of thanks for this very discussable paper.

Dr F. H. C. Marriott (University of Oxford): One of the great attractions of this paper is that it emphasizes the importance of looking at data. Computer packages often offer the temptation simply to plunge into calculations without thinking about the assumptions made or even about the questions that should be answered. Recent work on exploratory data analysis, regression diagnostics and graphical methods (such as those described here) does a great deal to counteract that tendency.

Of course, there is a risk that projections of high-dimensional data may show apparent structure that is merely the result of random variation. I think it is much more important that any claim of natural structure based, for example, on cluster analysis should be supported by a visual demonstration. I should lay down the general principle in this context that if something cannot be seen, it is not there.

The examples given show very clearly that principal components are not effective at displaying the features of heterogeneous data. Of course, they are not designed to do so, but they are very often used for that purpose. As the authors have shown, different approaches can do very much better.

I was greatly impressed by the practical examples presented, but I have some doubts about the theory. First, the original Friedman-Tukey approach to projection pursuit is identified as maximising an entropy criterion on the smoothed data. Huber, in his *Annals* paper, goes further and discusses a whole range of projection-pursuit indices, based on test statistics for non-normality. That was not what Friedman and Tukey described themselves as doing in their earlier paper. Rather, they were looking for dense points, and the approach was much nearer to the work done by Wishart, in 1969, on modal analysis—or, at least the earlier description of it was. The important point here is that the kernel was designed, I believe, to pick out dense points and not to smooth the whole data set. If the whole data set is smoothed using a kernel appropriate for smoothing a normal distribution, a mutivariate normal distribution, or any simple unimodal distribution, we shall certainly over-smooth any clearly structured set of data. Some of the interesting features may even be smoothed out.

In their early papers, the kernel used by Friedman and Tukey was very much too narrow for smoothing a whole set of data, and it was much narrower than those used by tonight's authors.

Secondly, the assertion that empirical data do not usually display very small third and fourth cumulants but large higher cumulants is true of uninteresting data of the sort that can be represented by Pearson curves, but is most emphatically false for multimodal distribution mixtures.

To take a very simple example, if we have three multivariate normal distributions equally spaced on a line, giving the middle one four times the weight of the other two, the resulting distribution has all cumulants of order 3, 4 and 5 equal to zero. No moment-based criterion can spot it in any lower dimensional projection.

Skewness and kurtosis are not in themselves interesting, and a moment criterion, or any criterion dominated by third and fourth cumulants, will miss clustered projections that happen to be roughly symmetrical and nearly mesokurtic. Further, this criticism applies to nearly all criteria derived from tests of non-normality.

Fig. 4 of the paper (that is, the canonical projection) illustrates these points. First, it is fairly clear

that the projection is over-smoothed. If we look at the letters at the bottom of the figure, it is obvious that there are three non-overlapping normal distributions. That, of course, is cheating because we should not look at the letters. However, those heavy tails with no points in them are again an indication that the data have been over-smoothed. The distribution is nearly symmetrical, and slightly platykurtic. The smoothed curve is just trimodal, but the centre mode in the smoothed curve looks slightly dubious at least.

Suppose we increase the number of samples of species A—increase the middle group—that will make the distribution much more obviously trimodal, but will certainly reduce any moment criterion. I suspect that it may very well reduce the entropy criterion as well.

In conclusion, I found this a stimulating and valuable paper, and I very much hope that it will provoke more research into an important technique. I cannot, though, agree with the underlying philosophy that investigation of the effect of maximising a criterion should centre on what happens when it is minimised.

It gives me pleasure to second the vote of thanks to the authors.

The vote of thanks was passed by acclamation.

Professor K. V. Mardia (University of Leeds): The moment index given by Eqn. (9) as expected extends itself to higher dimensions. In fact, for the p-dimensional case the moment index given by Eqn. (9) becomes simply

$$\frac{1}{12}\left[\beta_{1,p}+\frac{1}{4}\left\{\beta^*_{2,p}-6\beta_{2p}+3p(p+2)\right\}\right],$$

where

$$\beta_{1,p}=E\{(\mathbf{X}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{Y}-\boldsymbol{\mu})\}^3,\quad \beta_{2,p}=E\{(\mathbf{X}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{X}-\boldsymbol{\mu})\}^2,\quad \beta^*_{2,p}=E\{(\mathbf{X}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{Y}-\boldsymbol{\mu})\}^4,$$

and $\mathbf{X}$, $\mathbf{Y}$ are i.i.d. with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. This index can be written easily in terms of cumulants.

It is interesting to note that this index involves previously proposed measures of multivariate skewness and kurtosis (Mardia, 1970). Now the affine invariance property is quite straightforward from this representation. Its sample counter-part is

$$\frac{n}{6}\left[b_{1,p}+\frac{1}{4}\{b^*_{2,p}-6b_{2,p}+3p(p+2)\}\right]=MI, \text{ say},$$

where

$$b_{1,p}=\frac{1}{n^2}\sum\sum D_{ij}^3,\quad b_{2,p}=\frac{1}{n}\sum D_{ii}^2,\quad b^*_{2,p}=\frac{1}{n^2}\sum\sum D_{ij}^4$$

with

$$D_{ij}=(\mathbf{X}_i-\bar{\mathbf{X}})'\,\mathbf{S}^{-1}(\mathbf{X}_j-\bar{\mathbf{X}}).$$

The moment index MI emerges after some tedious algebra as the score statistic for testing $H_0:\boldsymbol{\pi}_2=\mathbf{0}$ (normality) versus $\boldsymbol{\pi}_2\neq\mathbf{0}$ for the random samples of size n from the population with density of $\mathbf{X}$ as const $\exp\{\boldsymbol{\pi}_1'\mathbf{U}_1+\boldsymbol{\pi}_1'\mathbf{U}_2\}$ where $\boldsymbol{\pi}_1$ and $\boldsymbol{\pi}_2$ are parameters and $\mathbf{U}_1=\{X_1, X_2, \ldots, X_1^2, X_2^2, \ldots, X_1X_2, \ldots\}$, and $\mathbf{U}_2=\{X_1^3, X_2^3, \ldots, X_1^2X_2, \ldots, X_1^4, X_2^4, \ldots, X_1X_2X_3X_4\}$.

Under H_0, MI has an asymptotic χ^2 with degrees of freedom ${}^{p+2}C_3+{}^{p+3}C_4$. This formulation explains the relative weights 4 : 1 in Eqn. (9) of the skewness and kurtosis. For small samples, we can use the k-statistics as in the paper. The likelihood ratio test has an asymptotic distribution related to the above but now H_0 lies on the boundary of the parameter space. This extended moment index allows us to perform projection pursuit into three or more dimensions. What problems do the authors foresee in its implementation?

Since the above is an omnibus test of multinormality, it is analytically clear that non-normality is confounded with projection pursuit. If the moment index is small/not-significant for p-dimensional data ($p>2$) then there is no point in doing projection pursuit. For example, the test hypothesis of multinormality is rejected for the whole Fisher iris data as one would expect but for Rao's cork tree data the test is not rejected.

One of the important general points for discrimination of biological shapes such as irises, vertabrae, palm, etc is that the selection of variables has been underrated. With image processing developments, we should be able to optimize such selection, i.e. we might have some more varibles for the iris data to discriminate all the three classes. Then the projection pursuit might work as satisfactorily as for the

fleas–beetles in the paper! Have the authors given thought to such problems? I am sure much more work is needed in this new area.

Finally, let me say that I found the paper very stimulating.

Dr J. T. Kent (University of Leeds): Tonight's paper gives an elegant presentation of an interesting tool in exploratory data analysis. I will direct my comments to the entropy index used in the paper and will emphasize a quantitative interpretation (rather than just a comparison of relative values) for this index.

The one-dimensional entropy indices in equations (10) and (11), when suitably centered, become estimates of the Kullback-Leibler information gain. Let $f^{(w)}$ denote the convolution of the true density f with the normal smoothing kernel with standard deviation $w > 0$, and let f_w denote the normal density with mean 0 and variance $1 + w^2$. Then, allowing for the smoothing in $\hat{f}$, (twice) the appropriate centered indices in (10) and (11) become

$$2 \int \log\{\hat{f}(z)/f_w(z)\}\hat{f}(z)\, dz = \tilde{\Gamma},$$

$$2 \int log\{\hat{f}(z)/f_w(z)\}\, dF_n(z) = \hat{\Gamma}, \text{ say,}$$

If $w \to 0$ with increasing sample size, these are both natural estimates of the population information gain, Γ_0, where for $w \geqslant 0$ we set

$$2 \int \log\{f^{(w)}(z)/f_w(z)\}f^{(w)}(z)\, dz = \Gamma_w.$$

The statistics $\tilde{\Gamma}$ and $\hat{\Gamma}$ can be regarded as a "fitted information gain" and an "observed information gain", respectively. In a parametric setting, without smoothing, it can be shown that in general $\tilde{\Gamma}$ is more accurate than $\hat{\Gamma}$. (Kent, 1986). Perhaps a similar result also holds in the non-parametric context. Note that while kernel smoothing is necessary in the estimation of f, it depresses the population information gain, $\Gamma_w < \Gamma_0$.

It is also instructive to consider what happens in the two-group discrimination problem. Suppose f is a mixture of two normal densities with equal variances and suppose a random variable $u = 1$ or 2 is available to indicate group membership. The natural information gain to consider for this problem is the "conditional information gain",

$$2 \sum_{u=1}^{2} \int \log\{f(z \mid u)/f_0(z)\}\, f(z \mid u)dz\, P(u) = \Gamma_c, \text{ say,}$$

which is related to the product-moment correlation ρ between z and u by $\Gamma_c = -\log(1 - \rho^2)$ (Kent, 1983). Of course in the clustering problem, u is not observed, so $\Gamma_0 < \Gamma_c$. Thus $\tilde{\Gamma}$ and $\hat{\Gamma}$ estimate a lower bound for Γ_c. Numerical methods could be used to examine more precisely the difference between Γ_0 and Γ_c.

As a simple example of the quantitative interpretation of information gain, consider Figs 2(a) and 2(b) in the paper. We find $\tilde{\Gamma}$ takes the values 0.02 and 0.36 here. These values suggest a negligible and a substantial information gain, respectively, and a glance at the figures confirms this interpretation. Further, $\tilde{\Gamma} = 0.36$ corresponds to a correlation of $\rho > 0.55$ in the discriminant analysis framework.

Drs D. L. Banks and G. A. Young (University of Cambridge): The authors deserve applause for setting forth a new methodology so clearly; they spotlight strengths of the technique, and indicate issues that merit attention.

One issue is choice of the projection index. The Friedman-Tukey index, the entropy index, and the moment index do not exhaust the plausible candidates. Others include measures of roughness (Good and Gaskins, 1971) and measures of diversity (Rao, 1982). An index should give low values for dull projections, large values for interesting ones, and balance robustness against outlier detection. The authors identify the Gaussian distribution as dull, without specifying the interesting case. Heuristically, the most interesting projection is that putting mass 0.5 at ± 1. This suggests that good indices respond to bumps, rather than heavy tails or asymmetry. For ϕ the standard normal density, alternative indices include: $\int [f(z) - \phi(z)]^2 dz$, $\int [\gamma''(z)]^2 dz$ where $\gamma = \sqrt{f}$, or $\iint d(y, z)\, f(y)f(z)\, dy\, dz$ where $d(.,.)$ is a metric.

A second issue concerns estimation of the marginal corresponding to a particular projection. This

problem is not strictly equivalent to univariate density estimation; in PP, one wants to jointly estimate many marginals. Complications involve repeated use of the sample and the fact that ${}_nC_2$ projections superimpose two data points, which may cause spuriously high values of the index. Despite poor performance of kernel estimators in high dimensions, for PP purposes it may prove better to optimally estimate the K-variate density of the sample, and then examine the indices of its marginals. Wahba (1971) suggested that histograms with partitions at the order statistics yield density estimates that are asymptotically superior to kernel estimates.

If one insists on projection-wise kernel density estimation, then one should choose the bandwidth well. For entropy index (10), the authors use an h optimal for estimating the Gaussian density; curiously, they take $\sqrt{2}h$ for index (11). One ought to optimize for estimation of the chosen index, at a density intermediate between dull and interesting. The motivation is that one wants to classify projections into two categories, with most declared dull and a few found interesting. Usually, this is best accomplished if estimation of the index is most accurate at the dividing line. Thus a good h might minimize $\mathbb{E}_{f^*}[\{I(\hat{f}_h) - I(f^*)\}^2]$, where

$$f^*(z) = \frac{1}{2\sqrt{\{2\pi(1-\alpha^2)\}}}\left[\exp\left\{-\frac{(z-\alpha)^2}{2(1-\alpha^2)}\right\} + \exp\left\{-\frac{(z+\alpha)^2}{2(1-\alpha^2)}\right\}\right]$$

with α selected to match the practitioner's view of borderline interest. For $\alpha = .5$ and sample size 74, numerical integration shows the optimal bandwidths for (10) and (11) to be .55 and 1.04, respectively. The additional expense of this minimization is negligible compared to the cost of the PP.

Dr A. M. Herzberg (Imperial College of Science and Technology): In a forthcoming paper, Lindsey, Herzberg and Watts (1987) propose a method for cluster analysis based on projections and a minor variant of quantile-quantile plots. This method can be used for exploratory data analysis, outlier detection and classification.

Let $\mathbf{x} = (x_1, \ldots, x_k)$ represent measurements on k variables of an object. Andrews (1972) suggested that $\mathbf{x}$ could be represented by the function

$$f_{\mathbf{x}}(t) = x_1/\sqrt{2} + x_2 \sin t + x_3 \cos t + x_4 \sin 2t + x_5 \cos 2t + \ldots \; (-\pi < t < \pi).$$

The projection method proposed corresponds to selected values of t. As t varies from $-\pi$ to π, all possible projections or *views* are obtained. Therefore, to simplify presentation, a non-redundant set of values of t is selected, namely $0, \frac{1}{4}\pi, \frac{1}{2}\pi, \frac{3}{4}\pi$. In the general case, this can be considered as a set of views which correspond to points on a k-dimensional hyper-cube.

Let $\mathbf{X}$ be the $n \times k$ matrix of the data, where the uth row $\mathbf{x}_u = (x_{1u}, \ldots, x_{ku})$, is the measurement on k variables of the uth object.

In a k-dimensional analysis, there are $\frac{1}{2}(3^k - 1) = m$ non-redundant set of k weights. Let $\mathbf{W} = \{w_{ij}\}$ $(i = 1, \ldots, m; j = 1, \ldots, k)$ be the matrix of weights, where $w_{ij} = 1, 0, -1$, *i.e.* w_{ij} is the ith weight given to the jth measurement of an object. Let $\mathbf{V} = \mathbf{W}\mathbf{X}'$ be the view matrix, each row $\mathbf{v}$ giving a view of the data. The values in each row of $\mathbf{V}$ are standardized to lie between 0 and 1 by substracting the minimum of the row from each value and dividing by the range. The resulting values are plotted on a contrast plot, a minor variant of a quantile-quantile plot. In this case, the quantiles of each row of $\mathbf{v}$ are plotted against the order statistics of the uniform distribution.

In the paper by Lindsey, Herzberg and Watts (1987), a subset of the data in Lubischew is used, i.e. for each of the three species only the first ten insects as listed are used and only the first three measurements on each insect. Using our method, we found that we obtained the correct clustering of the data except for the fourth insect of Species B which was an outlier and the third insect of species B which was classified as species A. The method also correctly categorized two further insects from each of the three species. Would the authors identify the outliers they found?

Professor J. B. Copas (University of Birmingham): May I compliment the authors on an interesting paper. I have a comment and a question.

My comment is to emphasize the importance of sampling fluctuations mentioned in Section 5.4. Are the authors being too dismissive of the problem of spurious results? How large has N to be—I imagine extremely large if K is large? There is a risk that the capability of computers to do ever more extensive searches of a finite set of data will outstrip our ability to interpret the results. A better known example of this danger is optimum subset selection in multiple regression, where, unless N is large, the coefficients for a selected subset can be grossly inflated. The problem of overfitting is an important one, and progress

on the authors' suggested simulation method will be welcome. If a computer package based on tonight's paper is envisaged, I hope the user will be warned when an apparently interesting feature of a selected projection is of the kind to be expected of sample fluctuations.

My question is to ask how the authors' method extends to discrete data. I have a practical problem in mind. In psychiatry, one can be faced with a long list of signs and symptoms, thus defining K dichotomous variables for N subjects. In a search for a possible bipolar dimension, the folklore has it that we do a factor analysis and then examine the fitted factor scores. Whilst the logic of this escapes me, I wonder if there is a projection pursuit approach. A point scoring rule associates a weight with each presenting symptom and adds over symptoms, and so is equivalent to a projection of the variables when coded as 0 or 1. Each projection will give a discrete distribution with at most 2^K points of support. The entropy index (in its discrete version) will try to coalesce the points of support to produce as "rough" a distribution as possible. Each of the K projections which puts unit weight on just one symptom and zero weight on the others, will give a local maximum, and one of them (i.e. one of the original variables) will often give the global maximum. Clearly any sensible measure of multimodality has to recognize the discrete nature of the data, and in this the entropy index fails. Can a suitable index be found for a projection pursuit approach to this problem?

Professor G. A. Barnard (University of Essex): The Society is to be congratulated on inducing Professor Sibson and Dr Jones to present us with such a brilliant account of this method, which is one of the first crop (of which many more can be expected) resulting from the enormous extensions in computing which were not available to the founding fathers of our subject.

I have just two cautionary points to add to what has already been said: First, that the first sentence of the second paragraph of p. 1 in which it is stated that "We cannot directly appreciate patterns of variation in more than three dimensions" rather suggests to me that the authors may have forgotten about the multidimensional patterns in a ballet executed with coloured lights and music. We could well start to explore what coloured moving computer outputs could do for us.

My second cautionary remark concerns the discussion of invariance requirements on p. 3, where the caution that applied mathematicians should always observe may have been neglected, in that we should always check whether our statements could remain true if "approximately" were put in front of any adjectival phrase.

For example, the statement that the method should be invariant under location and scale shifts would, if attended to by Karl Pearson in the early days, have meant that we went without the histogram. We have had to wait for smoothed density estimates and Jack Kiefer's modification of the Kolmogorov-Smirnov test before we could properly assess densities from empirical data.

The histogram analogy leads me to enquire whether anyone has explored the idea of putting a grid on the plane projection of our set of points, then using the Poisson index of dispersion to assess whether there is a serious departure from a uniform distribution. As others have indicated, such a distribution could serve as a standard "uninteresting distribution" especially when data have been gathered by looking through a "window" at a wider set of data.

Professor B. W. Silverman (University of Bath): One of the interesting facets of projection pursuit is that it illustrates the use of probability density estimates, not as an object in their own right, but as a component part of a more complicated procedure. As the authors say in their Section 3.4, not very much is known about the appropriate choice of window width when estimating a functional of an unknown density. Functionals of densities also occur in the study of the smoothed bootstrap, and some work in that context is reported in Silverman and Young (1987). Unfortunately the optimal choice of window width depends, as one might expect, on the particular functional being estimated and it is hard to say definitively whether or not one should always adopt the authors' (implicit) approach of aiming at a good estimate of the density itself. Some interesting results on the estimation of functionals of densities have also been obtained by Joe (1986).

The authors stress that their approach is "completely distinct" from other methods. While it is obviously neat and elegant that this should be so, and also very natural that the proponents of new techniques should wish to stress the differences from other methods, I hope that an aim of future research in this field will be the investigation of hybrid approaches that combine projection pursuit ideas with both "classical" and modern dimension-reduction methods, such as principal components analysis and the tree classification methodology of Breiman *et al.* (1984). The authors of tonight's paper are to be congratulated on their excellent presentation which will, I hope, be an important stimulus in this direction.

Dr A. W. Bowman and Professor D. M. Titterington (University of Glasgow): We enjoyed reading tonight's paper and regret being unable to contribute to the discussion in person. We would firstly like to draw attention to the importance of sampling fluctuations which the authors mention briefly in Section 5.4. The authors describe their technique as exploratory and are reluctant to employ the machinery of a formal test. However, if the proposed techniques are to be used in a general way then it seems essential to have some detailed knowledge of the behaviour of the projected patterns when the data are from a multivariate Normal distribution. Without this, there must be considerable danger of over-interpreting projected shapes, which the authors briefly mention, if projection pursuit is used in a routine way.

A preliminary test of non-Normal behaviour is particularly appealing in view of the use of the entropy index itself as a powerful test statistic for Normality, as discussed by Vasicek, and in its kernel density form in some as yet unpublished work of one of us (AWB). Mr Peter Foster, a research student at the University of Glasgow, is currently studying the performance of estimates of entropy as test statistics for multivariate Normality. However, a test would also be possible at the projection level rather than the global level. It would be useful to tabulate percentage points of the maximised entropy index of projections of multivariate Normal data for a variety of dimensions and sample sizes. We do not propose a rigid use of significance levels but would simply point out that a natural means of assessing the entropy index is readily at hand.

On a somewhat different point, the Example is analysed on the basis of form (10) of the order-1 entropy index. Does the analysis greatly change if (11), or even (12) or (4), is used instead? In this context, we wonder if the order-2 entropy versions might be less "volatile", particularly if any attempt is made to choose an optimal window width. Our concern originates from the parallel with density estimation, where methods and theory related to loss-based choice of smoothing parameters are more awkward in the case of Kullback-Leibler "loss" than in that of mean integrated squared error.

The following contributions were received in writing, after the meeting.

Professor H. Caussinus (Université Paul Sabatier, France): In my discussion of this interesting and important paper, I will confine my remarks to the complementary use of principal components analysis (PCA) and projection pursuit. As stated by the authors (p. 15), it is sometimes necessary to perform a preliminary PCA to reduce dimensionality for computational purposes. I do agree with them, but I would further suggest that such a preliminary projection could be useful, even when there are no technical constraints. On one hand indeed projection pursuit is essentially a descriptive tool which may be too sensitive to random variations. On the other hand PCA is adapted to lessen the effects of possible irrelevant noises. The latter property is more apparent when PCA is presented as follows. The columns X_n of the matrix of (raw) data X are independent random K-vectors, $E(X_n) = x_n$, $\text{var}(X_n) = \sigma^2\Gamma$ (Γ is assumed to be known, often $\Gamma = I_K$), and there exists a linear manifold E_q of dimension q ($q < K$) such that x_n belongs to E_q for any $n = 1, \ldots, N$. Then PCA consists in estimating E_q and the x_n's by means of least squares, that is maximum likelihood if the errors $X_n - x_n$ are normal. Note that no hypotheses are required concerning the distribution of the x_n's in E_q. In Caussinus (1986, Section 3.3), I have discussed the model and stressed the fact that one can separate two questions which are closely related in most presentations of PCA: (i) the estimation of E_q and the x_n's, and (ii) the graphical representation of (the estimates of) the vectors x_n's. Projection pursuit can work efficiently to achieve (ii), while PCA would be used in the step (i). The consistency of the model above mainly depends on the choice of q, but here, of course, q can be taken larger than in classical applications of PCA. Incidentally, this gives rise to a new interest in the problem of dimensionality in PCA.

Dr C. Chatfield (University of Bath): Methods for exploring data vary widely from the simple techniques of EDA (exploratory data analysis, e.g. Tukey, 1977) and IDA (the initial examination of data, e.g. Chatfield, 1985), through "second-order" multivariate techniques such as principal component analysis, to projection pursuit. While simple in concept, the latter looks far from trivial from a computational point of view and I wonder if the authors can give any guidance on the availability of software. The latter will be needed if more people are to try out these interesting techniques.

Professor Jerome H. Friedman (Stanford Linear Accelerator Center, Stanford University): I congratulate the authors on a splendid exposition. My comments concern only three technical issues for which my experience has run counter to conjectures made in the paper. I relate some problems associated with each of these issues. Solutions are explored in Friedman (1987).

The authors were the first ones to appreciate the great computational advantage of a projection index based on polynomial moments. They observe that such a projection index will be heavily driven by projected outliers, but conjecture that this should not interfere with the discovery of projections with other types of interesting structure. Besides, the discovery of outliers should be one of goals of an exploratory projection pursuit. The problem lies in the fact that even quite well behaved multivariate distributions, with few or even no outliers, will produce many projections with outlying observations.

Consider a p-dimensional standard normal distribution. For $p = 5$, approximately 5% of the observations will have a distance from the origin greater than 3.3. For $p = 10$ and $p = 15$, the corresponding 5% distances are 4.3 and 5.0 respectively. Vectors from the origin to each of these points define projections for which the corresponding points will appear to be outliers. (I call these "pseudo-outliers.") An index that achieves high value for such projections will tend to be distracted from finding other types of structures that are less likely to be realized from multivariate normal data. In my experience this represents a serious problem in higher dimension.

It is not clear how to robustify a moment based index in the necessary way when it is calculated from the outer product tensors (13). Trimming based on (multivariate) distance from the origin will not work. Also, the robustification must be done in a way that preserves the existence and calculability of the derivatives of the index. This rules out simple trimming or other methods based on projected ranks.

It has been my experience that reliably finding the substantive maxima of the projection index is a difficult problem, and that simple gradient-directed methods (such as steepest ascent) are generally inadequate. The power of a projection pursuit procedure depends crucially on the reliability and thoroughness of the numerical optimizer. Even though the projection index is strictly continuous, it is difficult to optimize owing to the very large number of decidedly suboptimal maxima caused by sampling fluctuations. These "pseudo-maxima" can, and often do, trap a gradient directed optimizer unless one has the good fortune to start near a substantive maximum. Unless one wishes to rely on good fortune, a very thorough optimizer is required for projection pursuit.

My final comment concerns the search for more than one informative view of the data. I have found that simply trying to find the multiple maxima of the projection index to be fairly ineffective. Also interpretation is difficult since one does not immediately know the extent to which a new view reflects additional structure (not seen in the other views) or whether it is a direct reflection of already discovered structure.

Dr P. J. Green (University of Durham): It may be worth mentioning briefly that there are other projection indices, plausible for certain problems, that can be optimised algebraically. I am thinking of multispectral signal- and image-processing applications where an additive signal-plus-noise model is appropriate. Several promising noise removal methods operate spatially but non-linearly on univariate data, and before applying them it is sensible to rotate the data space at each pixel in order to obtain orthogonal components that are ranked in order of *signal-to-noise ratio*. This is the basis of the Maximum Noise Fraction transformation proposed and analysed in yet unpublished work by Green, Berman, Switzer and Craig of CSIRO and Stanford. The resulting projection directions are generalised eigenvectors of the covariance matrix of the signal with respect to that of the noise, and of course the whole exercise is equivalent to the principle components transformation in the (unusual) case of noise that is uncorrelated with equal variance in each band.

This work is quite different in spirit from that of tonight's excellent paper, but involves a use of projection pursuit intermediate in formality between the first and second columns of the Table.

Drs Trevor Hastie (AT&T Bell Labs) and **Robert Tibshirani** (University of Toronto): We congratulate the authors for their refreshing look at projection pursuit methods. Their reformulation of the original Friedman-Tukey measure, leading to the entropy $\int f \log f$, seems very natural. Despite their encouraging results, however, we have a concern about their implementation of this idea. They decide to sphere the data at the outset so that there is no need to standardize each projection for scale. This sphering seems to be contrary to the spirit of projection pursuit. After all, sphering is something one does to *normal* (or close to normal) data, and projection pursuit, in their reformulation, seeks *non-normal* projections. More concretely, outliers and clusters could greatly affect the sphering operation, and not always in a predictable way. The authors are aware of this problem (section 5.1) but perhaps don't feel it's important.

The use of a robust sphering operation might help, but the most straightforward approach would be

to omit the sphering and instead incorporate scale into the entropy measure. The affine invariant version of entropy is

$$\int f \log f + \log s(f)$$

where $s(f)$ is an equivalent measure of scale (the median absolute deviation would be a good choice). This is (modulo a constant) the standardized Shannon entropy given by Huber (1985) in his equation 5.12, an example of a "Type III functional". Now if the standard deviation is used for $s(f)$, then maximizing this measure over all projections can be seen to be equivalent to the authors' procedure in the sense that the optimal projections, expressed in terms of the original varibles, would be the same. However, if a robust $s(f)$ is used then they are different and we believe that our suggestion will work better because it requires no sphering, relying only on a one-dimensional measure of scale.

There is a computational price for this modification, namely that derivative-based optimization methods can no longer be used and a derivative-free technique like the Nelder-Mead simplex algorithm must be used. However in today's era of exponentially increasing computer capablities this should not be a concern. In addition, Friedman (1985) expresses concern with blind use of derivative based optimization procedures since they tend to get trapped in local maxima. He suggests first using derivative-free methods to get in range of "substantive maxima", and then using derivative based methods for fine tuning.

This paper and the recent work of Huber and others make it clear that projection pursuit is a mathematically interesting and potentially powerful tool for data analysis. Now we statisticians must work with other scientists in applying this tool where it might produce some benefits.

We thank David Andrews for helpful discussion and the Natural Sciences and Engineering Research Council of Canada for its support of the second author.

Dr J. Rodney Jee (Jet Propulsion Laboratory, California Institute of Technology): The requirement for projection pursuit to maximize departures from normality is a key idea; but, an issue which surfaces is whether different normality indices yield significantly different results in this context. This is one issue pursued in my Ph.D. thesis, Jee (1985), from which my comments are drawn.

In designing a study to affect comparisons of various indices, a decision was made to avoid the possibly confounding issue of appropriate window width for density estimates by applying projection pursuit to

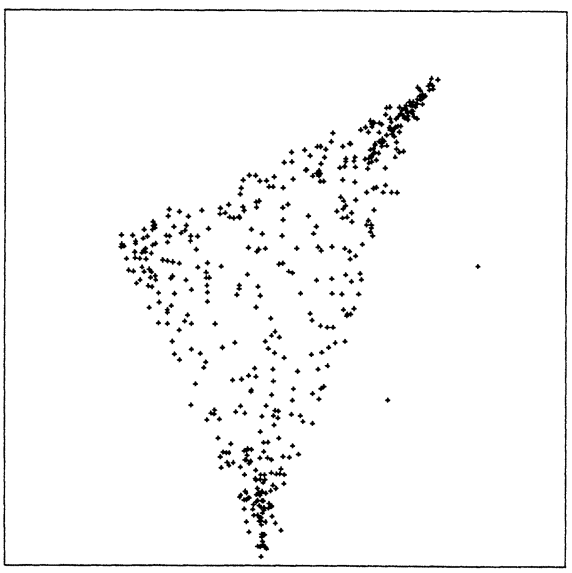

Fig. D2. Projection maximizing Fisher information.

population models instead of their finite sample realizations. Thus, a comparative analysis was performed by optimizing functionals of lower dimensional marginals of a population density in the original superspace. The measures of normality studied included Fisher information and order-1 entropy. Reduction from two dimensions to one and from three to two was studied in this fashion.

Of the two dimensional population models to which projection pursuit was applied, the one which revealed the most differences among the various indices was a mixture density composed of three equally proportioned bivariate normals all with covariance $\frac{1}{2}I$ and means (2, 2), (−2, −2), and (2, −2). At least two modes are present in every marginal, and three well separated modes are present in the marginal projected on the line connecting (2, 2) and (−2, −2). Among the indices, only Fisher information achieved a global maximum at this trimodal projection. Unexpectedly, entropy was minimized by this projection at a value well below its global maximum. Application of projection pursuit for reduction to two dimensions to an analogous three dimensional model produced similar results with Fisher information favouring a projection with more modes than that favoured by entropy.

Also in my thesis, versions of projection pursuit using histogram density estimates, from which Fisher information and entropy were calculated, were implemented for use on synthetic and real data sets. One real data set which served to differentiate between the two indices was the seven dimensional particle physics data of Friedman and Tukey (1974). Entropy did not find projections very different from those reported in their paper, but Fisher information in addition to finding a similar projection also found the projection of Fig. D2 to be significantly higher in information content—a finding which coincided well with our own notions of a more interesting projection.

From these results, it is apparent that there is work to be done in determining which measure(s) of normality may serve best in a projection pursuit algorithm.

Dr Harry Joe (University of British Columbia): I would like to congratulate the authors on an excellent paper.

I will restrict my comments to Section 3 as I have recently done some related work on kernel density estimates. In my paper (Joe, 1986) I study $N^{-1}\sum_{n=1}^{N} J(\hat{f}(X_n))$ as an estimator of $\int J(f)fd\mu$ where $\hat{f}$ is a kernel density estimate of the p-variate density f based on the random sample $X_1, \ldots, X_N$, and from mean squared error expressions, I obtain data-dependent methods for choosing the window width. Special cases of $J(f)$ lead to (4) and (11), the integral of the square of the density and the negative of entropy respectively. I have found both (4) and (11) to be decreasing as the window width increases; this is because both $\int f^2$ and $\int f \log f$ decrease as f becomes flatter and less dispersed. Although the authors find empirically that the entropy index is comparatively insensitive to window width, I have found situations, such as for heavy-tailed densities, where there is sensitivity to the window width, especially in the multivariate situation. From the mean squared error expressions for the multivariate versions of (4) and (11), I obtain the optimal window width to be of order $N^{-1/(p+2)}$. For estimating $\int f^2$, an estimator leading to a mean squared error of smaller order is

$$N^{-1}\sum_{n}\tilde{f}(X_n) = [N(N-1)h^p]^{-1}\sum_{n\neq m}\phi((X_n - X_m)/h),$$

where $\tilde{f}(X_n)$ is a cross-validatory estimate of $f(X_n)$ and ϕ is the kernel. For this estimator, the optimal window width is of order $N^{-2/(p+4)}$. This estimator and my method for choosing the window width provide an alternative to using (4) and $\sqrt{2}N^{-0.2}$ as the window width. Also, I have provided an alternative window width for (11).

Professor Iain M. Johnstone (Stanford University): The authors are to be congratulated on an excellent paper. I should like only to report a couple of points that arose from discussion in the course of Friedman's recent work on moment indices in exploratory projection pursuit.

In experiments using multivariate Gaussian data with moment indices of the form (8), (but including terms through order 6), Friedman found repeatedly that the projection optimizing the index consisted of a nice Gaussian cluster with a single distant outlier. This situation persisted, even when points furthest from the mean were removed by a trimming procedure. This problem is endemic in high dimensions, and reflects the fact that "each point is an outlier in its own projection". For definiteness, imagine 100 standard Gaussian points x_i in ten dimensions. In a projection in the direction determined by the origin and x_1, the distance of x_1 from 0 is distributed as χ_{10}, with median about $\sqrt{10}$ and 90th percentile at 4. The projections of $x_2, \ldots, x_{100}$ onto x_1 will form a standard univariate Gaussian sample of size 99,

with largest value typically around 2.3. Even trimming ten percent of the points will leave the 90th observation almost two standard deviations away from the pack in its own projection.

The second point relates to the deleterious effect on the 'pseudo-outlier' phenomenon when the data arise from distributions with tails that are heavier than Gaussian. We made one attempt to quantify this by considering a sample of size n from the (spherically symmetric) multivariate t-distribution in d dimensions with s degrees of freedom. The upper qth percentile of the distribution of the largest distance from the origin turns out to be $\{s(1-x)/x\}^{1/2}$ where x is the $1-(1-q)^{1/n}$th quantile of the beta distribution on $s/2$ and $d/2$ degrees of freedom. When divided by $\{s/(s-2)\}^{1/2}$, this bounds below the number of (univariate) standard deviations away from the origin that the furthest point will appear (on q percent of trials) when plotted in its own projection. Fig. D3 shows how the quantile varies as a function of tail heaviness of the multivariate t for three representative cases.

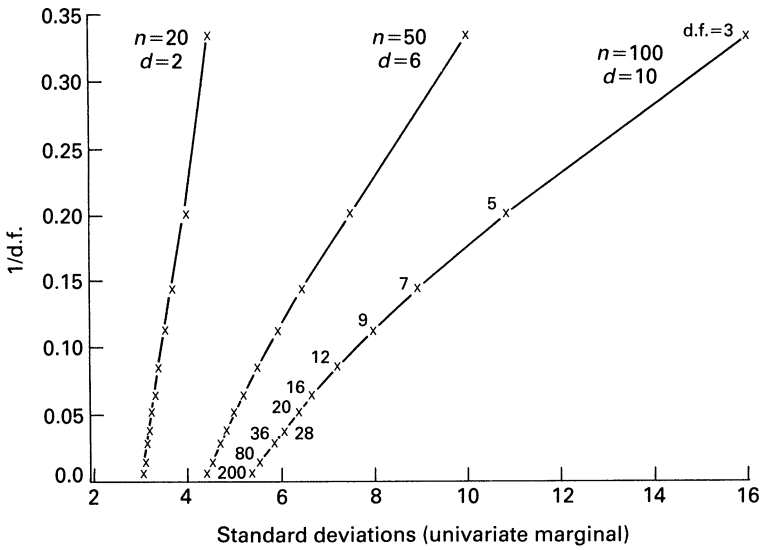

Fig. D3.

Dr Tormod Næs (Norwegian Food Research Institute): Since the first papers on projection pursuit appeared, I have found this use of smoothing, projections and graphics very appealing. In my opinion, this and similar approaches heavily based on computer graphics will have a great future in statistics and data analysis. After reading the excellent paper by Jones and Sibson I am even more convinced about this.

In this comment I will focus on two different aspects of projection pursuit regression.

(i) From my experience with multiple regression, interpretation of a solution is not very easy, and perhaps the most important goal of a regression analysis from a practical point of view is prediction. A problem then arises in projection pursuit regression because the method gives no closed-form predictor in which we can plug in our predictor variables. One can of course use interpolation between smoothed values, but what is the best way of doing this? It would be very convenient to have some general techniques and guidelines for making closed- form representations for projection pursuit solutions.

(ii) The second point I would like to touch is the collinearity problem. From linear regression theory it is known that collinearity of predictor variables can destroy the regression solution completely. This is serious both for interpretation and prediction studies. I am worried about the same phenomenon in projection pursuit regression and think this is an important field for research since collinear data are very common in practice. In my field which is chemometrics I would say that nearly all datasets are of this type. In Friedman and Stuetzle (1981) is indicated how a "stepwise" projection pursuit could be performed, but from by background with such data I prefer methods using the whole set of variables. Perhaps it could be possible to develop a principal component version of projection pursuit regression

or a ridge regression analogue. The most simple suggestion would of course be to compute the principal components of the covariance matrix and do a projection pursuit regression on the r first of them. Perhaps the solution then would be more stable and be better suited for prediction.

Dr M. Pawlak (University of Manitoba): I would like to point out some of the properties of two kernel estimates of the Friedman-Tukey projection index, that is, the estimates of a functional $I = \int f^2(x)\, dx$, where f is a density of the projected data.

The possible sample versions of I, based on the kernel density estimate with the symmetric kernel K and window width h, are:

$$I_N = \frac{1}{N^2 h} \sum_{i=1}^{N} \sum_{j=1}^{N} K\left(\frac{Z_i - Z_j}{h}\right)$$

and

$$\bar{I}_N = \frac{1}{N(N-1)h} \sum_{\substack{i=1 \\ i \neq j}}^{N} \sum_{j=1}^{N} K\left(\frac{Z_i - Z_j}{h}\right).$$

The estimate I_N has been employed by Friedman-Tukey and adapted by the authors as well, while the estimate $\bar{I}_N$ is obtained by removing the diagonal terms from I_N.

Let $$\int (f'(x))^2 dx < \infty \quad \text{and} \quad \text{let} \int f^3(x)\, dx < \infty.$$

Using arguments similar as in Pawlak (1986), see also Aubuchon and Hettmansperger (1984), we can show, under some selection of h, that

$$NE(I_N - I)^2 \overset{N}{\to} \alpha \quad \text{and} \quad NE(\bar{I}_N - I)^2 \overset{N}{\to} \alpha,$$

where $\alpha = 4[\int f^3(x)\, dx - (\int f^2(x)\, dx)^2]$.

Thus, a sharper result is necessary in order to distinguish the performance of these estimation methods.

Further analysis leads to the following second order expansions:

$$E(I_N - I)^2 \simeq \frac{\alpha}{N} + \frac{K^2(0)}{N^2 h^2} + h^4 \beta^2 b^2 \tag{1}$$

and

$$E(\bar{I}_N - I)^2 \simeq \frac{\alpha}{N} + \frac{a}{N^2 h} + h^4 \beta^2 b^2 \tag{2}$$

where

$$a = 2 \int K^2(x)\, dx \int f^2(x)\, dx, \quad b = \int x^2 K(x)\, dx/2 \quad \text{and}$$

$$\beta = -\int (f'(x))^2\, dx.$$

From (1) and (2), it appears that the best values for h are given by

$$h^* = (K^2(0)/2\beta^2 b^2)^{1/6} N^{-1/3} \tag{3}$$

and

$$\bar{h}^* = (a/4\beta^2 b^2)^{1/5} N^{-2/5}, \tag{4}$$

for the estimates I_N and $\bar{I}_N$, respectively.

This, in turn, determines the asymptotic formulae for the mean squared error

$$E(I_N - I)^2 \simeq \frac{\alpha}{N} + \theta_1 N^{-4/3}$$

and

$$E(\bar{I}_N - I)^2 \simeq \frac{\alpha}{N} + \theta_2 N^{-8/5},$$

where θ_1, θ_2 are positive constants dependent on f and K. Apparently, the estimate $\bar{I}_N$ is more efficient than I_N.

Jones and Sibson suggest the use of a window width $\sqrt{2}N^{-1/5}$ for the estimate I_N. In light of (1) and (3) it seems to be a suboptimal choice.

The only problem which remains is that h^* and $\bar{h}^*$ depend on the unknown density f.

If, however, we assume that f can be roughly estimated by a normal (μ, σ^2) parametric family, then

$$h^* = (32\pi\ C_1)^{1/6}\sigma N^{-1/3}$$

and

$$\bar{h}^* = (C_2/6\sqrt{\pi})^{1/5}\sigma N^{-2/5},$$

where

$$C_1 = K^2(0)\Big/\left(\int x^2K(x)dx\right)^2 \quad \text{and} \quad C_2 = \int K^2(x)dx\Big/\left(\int x^2K(x)dx\right)^2.$$

Hence, our data dependent estimates of (3) and (4) for nearly-normal densities f, become $(32\pi C_1)^{1/6}\sigma_N N^{-1/3}$ and $(C_2/6\sqrt{\pi})^{1/5}\sigma_N N^{-2/5}$, respectively, where σ_N^2 is the sample variance.

For the Gaussian kernel, the latter resolves to $2^{2/3}\sigma_N N^{-1/3}$ and $(12\pi)^{-1/5}\sigma_N N^{-2/5}$.

Dr Werner Stuetzle (University of Washington): The authors have obviously spent a lot of effort thinking about projection pursuit; the fact that others (Huber, 1985; Friedman, Stuetzle, and Schroeder, 1984) independently arrived at similar ideas speaks in favour of the ideas as well as the authors.

There are two points I wish to comment upon.

(i) *Projection pursuit exploration versus projection pursuit density estimation*
I do not agree with the statement that projection pursuit density estimation (PPDE) is "clearly distinct from the original idea". Multivariate density estimation *per se* is not an important problem. The motivation behind PPDE always was to detect structure in the multivariate data distribution by looking at marginals along the directions found by the algorithm. Iteratively constructing a density estimate was merely a means for deflating optima in the projection index and thus allowing the optimizer to find structure not previously uncovered. It is worthwhile to note that if a multivariate Gaussian with sample mean and covariance matrix is chosen as the initial model in PPDE, the projection index used in PPDE for finding the first projection direction is the entropy index suggested by Huber and the authors.

(ii) *The role of sphering*
It is clear that sphering, as described by the authors in section 2.1, makes projection pursuit into an affine equivariant procedure, no matter what projection index is used. If affine equivariance is the only goal to be achieved by sphering, there would be no reason to go to (affine equivariant) robust estimates of the covariance. The choice of sphering method will, however, influence the numerical optimization. Numerical optimizers are not necessarily equivariant, and certainly the choice of random starting directions is not.

Mr P. C. Taylor (University of Bath): Firstly, I would like to congratulate the authors on their paper, and say how much I enjoyed their presentation. I would also like to comment on the results of the Canonical Variates analysis of the beetle data. This analysis generated a one dimensional projection of the data, in which (when the species labels are introduced) the three species can be discriminated. Recently, I have applied the technique called C.A.R.T. (Breiman *et al*, 1984) to this data set. C.A.R.T. is a procedure for forming a decision tree given a training set. The resultant tree has the property that in order to classify an object, you ask a sequence of questions of the form, "Is x_i greater than K?", where x_i is the value of the ith (untransformed) variable for the object to be classified. This results in a tree which is easily used and understood by non-statisticians, and is computationally cheap to construct. For the beetle data it is possible to classify every beetle correctly by asking (at most) two questions

about individual variables. I feel that this ease of use is more of an advantage than the mathematically pleasing one dimensional projection produced by Canonical Variates analysis. Of course, like Canonical Variates analysis, C.A.R.T. uses the species labelling, and hence cannot be compared with Projection Pursuit.

John W. Tukey (Princeton University, Fine Hall, Washington Road, Princeton, NJ 08544): Much of Jones and Sibson's contribution falls under two heads: (i) new and perhaps improved projection indices and (ii) moving the technique closer to more classical multivariate approaches.

The first of these may be only a matter of fine tuning, which could indeed be very important, or it may be a matter of essential distortion. Evidence on this point has to come from comparisons of different techniques on a variety of different, hopefully distinctively different, sets of *real* data. Lacking this, we can only speculate about comparisons.

Probable difficulties in using moment indices, due to apparent or real outliers, are discussed by Friedman in his comments on this paper. One way to reduce, or even eliminate such problems would be to: first, sphere the data; second, displace radically inward (toward the data's center of gravity) points of larger radius (we might, for instance, after sphering to ave $r^2 = 2$, replace each r greater than 1 by $1 + \log_e r$ or $3 - 2r^{-1/2}$; third, resphere the modified data; finally, apply Jones and Sibson's moment algorithm.

Another possible modification would be to use the 3rd cumulants alone, or the fourth ones alone, rather than using a particular linear combination. (Since we expect to look for multiple extrema in any event, trying two or more criteria will often be reasonable.)

To the second point, I must express my doubts that all—or even, I fear, most—multivariate data sets will prove best handled by techniques smoothly related to techniques apt for ellipsoidal Gaussian blurs. To "If it's not broken, don't fix it!" we need to add "If its badly broken, don't try to fix it with only string and bandages!". While there are certainly a fair number of near-Gaussian data sets, there also seem to be many that show little or no sign of Gaussianity, and for which Gaussian-based methods appear "badly broken". (Most particle-physics data sets I have seen, and the Rubinstein data set discussed in Section 4.4 of Belsley, Kuh and Welsch (1980) are examples.)

I fail to understand the author's feeling at the end of Section 3.1, that we "lack a bivariate [Fast] Fourier Transform procedure". One way to explain the speed with which the ordinary FFT deals with 2^{10} points is to think of the equivalent 10-dimensional problem. A double FFT on $2^{10} \times 2^{10}$ points is just as fast as a single FFT on the same total number of points, namely 2^{20}.

At the start of Section 2, the authors assert that the original projection index was "designed ... to reveal clustering". Not so. Friedman and I were initially concerned with a particle-physics example where points were concentrated near three skew "rods" joined together in pairs, and our intent was to match the manually interactive tendency to seek views where *one* of the rods projected into a compact pile. The basic idea was to focus on "clottedness" or "local concentration", not on "clustering". Our concern with Gaussian clusters only arose in using previously discussed examples to compare the technique with earlier approaches.

The authors then go on to say "the Friedman-Tukey index now emerges as the result of constructing a kernel density estimate" (below formula (13)), "departure from parabolic form, and Friedman and Tukey's procedure is seeking any such departures" (below formula (4)), and, finally, "choice of a normal kernel is quite inoffensive from the point of view of density estimation" (below formula (11)). I venture to disbelieve each of these statements as relevant to the original projection-pursuit procedure.

Equally, of course, I do not accept the normative sentences at the close of the authors' section 5.1 as binding anyone except the authors themselves.

We should all be grateful to Jones and Sibson for careful work, new ideas, and interesting deformations, but there remains a need to gather much experience about the practical performance of a variety of procedures.

The Authors replied later, in writing, as follows.

We thank all the discussants for their constructive and thought-provoking contributions. It is impossible to do justice to all the ideas raised in the space and time allowed for composing this reply.

Mr Gower gives a detailed discussion of the effects of sphering. The comment of Drs Hastie and Tibshirani that "sphering seems to be contrary to the spirit of projection pursuit" is arguable: we assert that one of the ideas behind sphering is to get rid of normal aspects of the data in order to clarify further features. Mr Gower points out that we run the risk of affecting this further structure adversely in the

process. His comment on our preliminary scaling is, of course, correct; both choice of scaling and of sphering matrix are related to interpretation of projection directions rather than truly affecting projection pursuit itself. It is useful to retain the separate steps in conjunction with the auxiliary use of principal components.

Sphering is also confounded with the important question of outliers. Professors Friedman and Johnstone amply illustrate the problem which arises since in small multivariate datasets most "inlying" datapoints look like univariate (or bivariate) outliers, "pseudo-outliers", in some projections as do those that are truly multivariate outliers. One small point is that if this leaves a "truly" interesting view at a local maximum amidst pseudo-outlier maxima, some of which are somewhat higher, this may not matter: the investigator could ascribe greater importance to the good view on the basis of the pictures. However, many such views will indeed distract the method from finding truly interesting projections and there is a risk that interesting views might not correspond to local maxima at all. How best to robustify the algorithm is another matter. Professor Tukey suggests a method incorporating robustification into the sphering step: this looks guaranteed to reduce values of projection indices at "pseudo-outlier" projections, but is it likely to make (enough of) them locally sub-maximal? Many other robustification methods are possible, but no one seems ideal given our wish not to compromise too much the computational advantages of the basic algorithm. Drs Hastie and Tibshirani note that it might be preferable "to omit the sphering and instead incorporate [a robust measure of] scale into the entropy measure"; this recalls Friedman and Tukey's (1974) original formula. Using both sphering and a robustified index is also possible.

We have always recognised the susceptibility of the moment index to the outlier problem. Dr Marriott points out another situation where it may not work well. We regard these examples as illustrating its limitations rather than destroying its foundations. Much the same can be said of the lack of interest shown by the entropy index in multimodal views by comparison with asymmetrically bimodal views; the important thing, in our opinion, is to find *enough* views to give an understanding of the structure. We are grateful to Professor Mardia for the interesting interpretation of the moment index and equally to Dr Kent for useful comments on the entropy index.

The entropy and moment indices are just two of the many possible indices even within the framework of seeking deviation from normality or (Gower) from something equally uninteresting. We conjecture that the results obtained by using each index will be broadly comparable, but there may occasionally be important differences. Dr Jee makes the case for Fisher information; we wonder if any advantages it might have in the population context may fail to carry over to finite samples because of the difficulty of estimating the derivative of the (log) density. Drs Banks and Young suggest "that good indices respond to bumps" and colleagues of ours are considering explicitly using nonparametric bump-hunting methods (e.g. Silverman, 1981) in a related context; such a method equates uninteresting with unimodal, as mentioned by Mr Gower. Professor Barnard makes a valuable suggestion which does not conform to our invariance requirements. Invariance should not be a rigid barrier to use of such an index, although substantial compensating advantages would need to be offered.

Dr Marriott makes the valid point that "investigation of the effect of maximising a criterion [based] on what happens when it is minimised" is a dubious procedure. There is, however, a further justification for such an approach in the current context. This is that "for most high-dimensional point clouds most low-dimensional projections are approximately normal". The quote is from Huber (1985) who makes this point emphatically; theoretical quantification is made by Diaconis and Freedman (1984). Thus, most projections should have index values close to the minimum and high index views are the exception rather than the rule.

When using a functional of a nonparametric density estimate as projection index, an important question is that of choice of window width. Professor Silverman, for one, notes that a good w for estimation of the density as a whole may well differ from a good choice for estimation of a given functional; this is an important observation with which we concur. The contributions of Drs Joe and Pawlak contain theoretical work which indicates the sort of differences that might exist. Drs Banks and Young put forward an interesting argument leading to more smoothing than we used; Dr Marriott calls for less smoothing. In an earlier report of the beetle example (Jones, 1983), the structure was displayed more convincingly with a narrower window, as Dr Marriott suggests. However, it is not necessary to use the same w in the algorithm as that used for presentational purposes, for the very reason given above. Dr Bowman and Professor Titterington make the point that ease of choice of window width may be a factor in choosing the form of index used.

The above concerns choice of w for estimating a functional of a density of a single projection of the dataset. Drs Banks and Young make what looks an odd comment (surely kernel density estimates of

marginals are marginals of multivariate kernel estimates) but perhaps they are driving at the following point: for projection pursuit, the window width needs to be chosen to estimate appropriately the chosen index for all projections simultaneously. Furthermore, the important attribute is the shape of the index $I(a)$ thought of as a function of a, in the sense at least of appropriate numbers and positions of local maxima and, possibly, of their relative heights, too. This is the sense in which we made the claim in Section 3.4 that "the entropy index is comparatively insensitive to window width"; simulation experience in Jones (1983) suggests that the general shape of $I(a)$ remains much the same for a wide range of values of w and that virtually the same results would thus obtain from use of projection pursuit in each case. We intend pursuing further this line of inquiry, perhaps by choosing an appropriate loss function for theoretical work on what is a slightly unusual variant of the usual choice-of-smoothing-parameter situation.

Choice of window width has repercussions for the numerical optimisation step. It is tempting to choose w as large as possible within the broad range of reasonable choices conjectured above, to make optimisation easy. Professor Friedman's experience with the optimisation stage is interesting; his views and ours can be reconciled as follows. Professor Friedman supposes that there will always be a large number of uninteresting local optima and he proposes using a sophisticated procedure capable of stepping over them to get to the more substantive maxima; we aim to control the number of uninteresting maxima so that a simpler optimisation method finds the remaining maxima, virtually all of which are of potential interest.

Professor Copas and Dr Bowman with Professor Titterington stress the importance of not ascribing structure to the results of sampling fluctuation. The latter are joined by Professor Mardia in making the valid point that projection pursuit is worthwhile only if we believe that our data are not normally distributed; a preliminary test of multivariate normality could be used, but only on very large datasets. We recognise that our approach to choosing a projection index is closely tied in with testing univariate (or bivariate) normality of projections. While standing by our remarks in Section 5.4, we note that "testing the reality" of apparent structure is not all that simple a problem. Were it the case that an index had a single (global) maximum for any dataset, a comparison of its value with those obtained for comparable multivariate normal samples would be straightforward and meaningful, in fact as a test for multivariate normality (Bowman and Titterington). However, it is not clear even how properly to formulate the problem of assessing reality of the lower local maxima inevitably obtained—there is no simple correspondence with the several local maxima found for multivariate normal data. This problem is compounded in practice by the failure to guarantee finding all relevant optima. Friedman (1985) compares each (local) maximum he obtains with replications of a single (hopefully global?) maximum from multivariate normal simulations; this is an obvious conservative way to proceed. Any such simulation approach is very computationally demanding; a shortcut to simulating sphered multivariate normal data is to simulate data from a uniform distribution on a Stiefel manifold (Heiberger, 1978; Cheng, 1985), but the savings would be minimal in the context of the entire exercise. As Dr Bowman and Professor Titterington note, there is a real need for understanding the behaviour of projected data and, we add, projection indices, in the multivariate normal situation: some very limited results on the asymptotic behaviour of the Friedman-Tukey index (its tending to a Gaussian process when bivariate data is projected onto one dimension) are given in Jones (1983), but much more work is needed.

It is clear that potential modifications to the projection pursuit algorithm are manifold. Alternative improvements might be affected by changing the method more radically. Two such ideas have occurred to us. A first is the possibility of a sequential projection pursuit procedure wherein interesting views are sought one at a time and in decreasing order of importance; this idea brings the mechanism closer to that of principal components analysis. A second proposal is a converse of the current method: seek out and discard the least interesting views of the data in order to leave only the interesting structure. Such a method has the conceptual advantage of seeking the minima of projection indices, the theoretical basis of which is well understood. Professor Silverman rightly points out that there is considerable scope for combining projection pursuit with other multivariate techniques. The method does lose its effectiveness as K increases and a preliminary dimension-reduction, for example by using principal components as discussed by Professor Caussinus, is appealing and may be essential.

Professor Copas is referred to Diaconis (1983) who considers exactly the kind of discrete data that he mentions. It seems that the sensible approach is to change the notion of projection rather than to modify projection indices based on ordinary projection in euclidean space; see also Diaconis (1985) and Hastie and Tibshirani (1985). Diaconis (1985) mentions a suggestion of Professor Tukey that discrete data be mapped into euclidean space in some appropriate way. Although the former method is appealing for purely discrete data, perhaps the latter might be preferable for mixed discrete and continuous data.

Mr Gower proposes transforming from discrete to continuous too, but via suitable distance/dissimilarity functions. The opportunity provided for an interplay of projection pursuit and multidimensional scaling is attractive.

Professor Barnard mentions representation of datasets in three or more dimensions, Mr Gower discusses the interpretation of projections and we add the prospect of interactively exploring multivariate point-clouds. Each of these is of real import to exploratory multivariate data analysis. Such problems are currently the area of much research both computationally and in the field of human perception. Projection pursuit can be thought of as a static version of much more computationally demanding dynamic searches through projections. With computational advances, the role and nature of projection pursuit might change, but there is likely to remain a major part for it to play in helping guide the researcher to interesting areas in which to look.

Comparison of projection pursuit with other relevant techniques is also needed. In answer to Dr Herzberg's question, in Figure 3(b) it is the eighth insect of species B which overlaps with species A. None of the (pseudo?) outliers in Figures 5 and 9 correspond to the outlier Dr Herzberg mentions, nor did we find any *strong* suggestion of the fourth species B beetle being wayward elsewhere. The different subset of the beetle data used may well account for much of this difference. We have nothing to add to the comments of Drs Green or Næs or Mr Taylor. Professor Stuetzle shows that there is a close link between projection pursuit density estimation and the exploratory method. Our comment on the bivariate fast Fourier transform (Tukey) means that a double FFT on $2^{10} \times 2^{10}$ points is (essentially) 2^{10} times as burdensome as a single FFT on 2^{10} points. Professor Tukey's objection to our assertion about the basis of the Friedman-Tukey index can be dismissed as a matter of semantics rather than of any real misunderstanding.

Dr Chatfield asks about the availability of programs for projection pursuit: FORTRAN code implementing the method of the paper can be obtained, in strictly "pre-release" form, from the first author.

Two final points pervading much of the discussion bear repetition. The first is that what is especially needed is considerable practical experience with projection pursuit to highlight both its advantages and deficiencies. Secondly, it is clear that there are many questions requiring further research in this area. We hope that others besides ourselves will be stimulated to pursue further the ideas raised.

REFERENCES IN THE DISCUSSION

Andrews, D. F. (1972) Plots of high dimensional data. *Biometrics*, **28**, 125–136.

Aubuchon, J. C. and Hettmansperger, T. P. (1984) A note on the estimation of the integral of $f^2(x)$. *J. Statist. Planning & Inf.*, **9**, 321–331.

Belsley, D. A., Kuh, E. and Welsch, R. E. (1980) *Regression Diagnostics*. New York: Wiley.

Breiman, L. *et al.* (1984) *Classification and Regression Trees*. Belmont, Calif.: Wadsworth.

Caussinus, H. (1986) Models and uses of principal components analysis. In *Proceedings of the Workshop on Multidimensional Data Analysis, Cambridge, 1985*. Leiden: DSWO Press.

Chatfield, C. (1985) The initial examination of data (with Discussion). *J. R. Statist. Soc.* A, **148**, 214–253.

Cheng, R. C. H. (1985) Generation of multivariate normal samples with given sample mean and covariance matrix. *J. Statist. Comput. Simulation*, **21**, 39–49.

Diaconis, P. (1985) Contribution to the discussion of Huber (1985). *Ann. Statist.*, **13**, 494–496.

Friedman, J. H. (1987) Exploratory Projection Pursuit. *J. Amer. Statist. Ass.*, **82**, in the press.

Good, I. J. and Gaskins, R. A. (1971) Nonparametric roughness penalties for probability densities. *Biometrika*, **58**, 255–277.

Gower, J. C. and Legendre, P. (1986) Metric and Euclidean properties of dissimilarity coefficients. *J. Classification*, **3**, 5–48.

Hastie, T. and Tibshirani, R. (1985) Contribution to the discussion of Huber (1985). *Ann. Statist.*, **13**, 502–508.

Heiberger, R. M. (1978) Algorithm AS 127. Generation of random orthogonal matrices. *Appl. Statist.*, **27**, 199–206.

Joe, H. (1986) Estimation of a functional of a multivariate density. Technical Report No. 44, Dept of Statistics, University of British Columbia.

Kent, J. T. (1983) Information gain and a general measure of correlation. *Biometrika*, **70**, 163–173.

———(1986) The underlying structure of non-nested hypothesis tests. *Biometrika*, **73**, 333–343.

Lindsey, J. C., Herzberg, A. M. and Watts, D. G. (1987) A method for cluster analysis based on projections and quantile–quantile plots. *Biometrics*, in the press.

Mardia, K. V. (1970) Measures of multivariate skewness and kurtosis with applications. *Biometrika*, **57**, 519–530.

Pawlak, M. (1986) On nonparametric estimation of a functional of a probability density. *IEEE Trans. Information Theory*, **IT32**, 79–84.

Rao, C. R. (1982) Convexity properties of entropy functions and analysis of diversity. In *Inequalities in Statistics and Probability* (Y. L. Tong, ed.) Institute of Mathematical Statistics Lecture Notes Series, Vol. 5.

Silverman, B. W. (1981) Using kernel density estimates to investigate multimodality. *J. R. Statist. Soc.* B, **43**, 97–99.

Silverman, B. W. and Young, G. A. (1987) The bootstrap: to smooth or not to smooth? *Biometrika*, **74**, in the press.

Tukey, J. W. (1977) *Exploratory Data Analysis*. Reading, Mass.: Addison-Wesley.

Wahba, G. (1971) A polynomial algorithm for density estimation. *Ann. Math. Statist.*, **42**, 1870–1886.

Wishart, D. (1969) Numerical classification method for deriving natural classes. *Nature*, **221**, 97–98.

As a result of the ballot held during the meeting, the following were elected Fellows of the Society.

Cope, Frank
Francis, Michael
Hands, Stephen D.
Hayes, Susan J.
Kempthorne, Peter J.
Martin, Ann Marie
Matcham, James
Mathieson, Claire Ann
McColl, Elaine Margaret Mary
Overton, Richard Stanley
Papaioannou, Anna
Richards, Tim David
Ridout, Martin Spencer
Scott, James Richard
Seymour, David Gordon
Sharples, Linda
Waterfield, Malcolm Roy
Wickham, Carol Anne Charlotte
Wolf, Frederic Marc
Chan, Kwong Shing
Lim, Chuan Chai
Parkin, Stanley Frederick
Turnbull, Maurice

Brock, Andrew William
Chen, Siew Pheng
Donnan, Peter Thomas
Edwards, Sarah
Gould, Ann
Hart, Robert Michael
Jones, Margaret
Kelly, Gabrielle Elizabeth
Lawton, Andrew Magnus
Little, Sarah Louise
McShane, Philip
Oller, Lars-Eric
Pearson, Jeffery Edward
Peters, Timothy James
Prentice, Michael James
Salomaa, Hely Tuulikki
Small, Maria Karen
Vint, Hazel Millar
Wass, Ann Caroline
Wilson, Adrian Michael
Withey, Robin Paul
Wood, Jean
Worrall, Jeffrey Edward