SLAC-PUB-3841
STAN-LCS 18
December 1985
M

EXPLORATORY PROJECTION PURSUIT*

JEROME H. FRIEDMAN

Stanford Linear Accelerator Center
and Department of Statistics
Stanford University, Stanford, California 94305

ABSTRACT

Exploratory projection pursuit is concerned with finding relatively highly revealing lower dimensional projections of high dimensional data. The intent is to discover views of the multivariate data set that exhibit nonlinear effects - clustering, concentrations near nonlinear manifolds - that are not captured by the linear correlation structure. This paper presents a new algorithm for this purpose that has both statistical and computational advantages over previous methods. A connection to density estimation is established. Examples are presented and issues related to practical application are discussed.

Submitted to *Journal of the American Statistical Association*

* Work supported by the Department of Energy under contract DE-AC03-76SF00515, by the Office of Naval Research under contract N00014-83-K-0472.

1 Introduction

Often - especially during the initial stages - the analysis of a data set is exploratory. One wishes to gain insight and understanding about the nature of the phenomena or system that produced the data without imposing preconceived notions or models. For multivariate data a first set of useful summary statistics is based on the locations and scales of the measurement variables as well as their correlational structure. Classical multivariate analysis has provided powerful tools for their estimation (see Anderson, 1958). If the data closely follow and elliptically symmetric distribution (such as the normal) in the p–dimensional variable space then these summary statistics usually provide nearly all of the relevant information.

Sometimes, however, there is important structure in the data that is not adequately captured by the linear associations (correlations) among the variables. Such effects include clustering of the observations into distinct groups and/or concentrations near nonlinear manifolds. Knowledge of the existence and nature of such effects can often help in understanding the underlying phenomena. In contrast to linear effects, the variety of shapes and other attributes of nonlinearity are immense. It is thus impossible to prespecify all possibilities in advance and attempt to estimate their corresponding parameters. Powerful approaches can be based on making informative pictorial representations of the data upon which the human gift for pattern recognition can be applied. By viewing (lower dimensional) representations of the data density, the analyst can often detect striking as well as subtle structure that was impossible to anticipate.

Unfortunately, the human gift for pattern recognition is limited to low dimension. In addition the technology available to the analyst may place further restrictions on the viewing dimension. Ordinary plotting is limited to two dimensions. Sophisticated uses of motion and color with computer graphics displays can increase the dimensionality for viewing the data to three or perhaps a little more. If one is to graphically explore multivariate data, it is necessary to find highly revealing lower (one, two or three) dimensional representations.

The most commonly used dimension reducing transformations are linear projections. This is because they are among the simplest and most interpretable. Also, projections are smoothing operations in that structure can be obscured by projection but never enhanced. Any structure seen in a projection is a shadow of an actual (possibly sharper) structure in the full dimensionality. In this sense those projections that are the most revealing of the high dimensional data distribution are those containing the sharpest structure. It is of interest then to pursue such projections.

Friedman and Tukey (1974) presented an algorithm for attempting this goal. The basic idea was to assign a numerical index to every (one or two dimensional) projection, that characterized the

amount of structure present (data density variation) in the projection. This index was then maximized (via numerical optimization) with respect to the parameters defining the projections. They termed this method "projection pursuit." Since all (density) estimation is performed in a univariate (or bivariate) setting this method has the potential of overcoming the "curse of dimensionality" (Bellman, 1961) that afflicts such nonlinear methods as parametric mapping (multidimensional scaling) and cluster analysis that are based on interpoint distances. Also, the projection index can (and should) be affine invariant (see Huber, 1985). Therefore, unlike (projections based on) principal components or factor analysis, it is unaffected - and thus not distracted - by the overall covariance structure of the data which often has little to do with clustering and other nonlinear effects.

Projection pursuit solutions are seldom unique. Usually data structuring in the full dimensionality will be observable in several lower dimensional projections and the viewing of each can provide additional insight. It is thus important that a projection pursuit algorithm find as many of these views as possible. Each of these views will (hopefully) give rise to a substantive local maximum in the projection index. One way to encourage the discovery of several important views is to repeatedly invoke the optimization procedure, each time removing from consideration the solutions previously found. Structure removal is therefore an important part of any successful projection pursuit procedure.

This paper presents a new projection pursuit algorithm. Its projection index has superior sensitivity and similar robustness properties to the Friedman-Tukey (1974) index and it is much more rapidly computable. The optimization procedure is faster and far more thorough. Finally, a systematic solution to the structure removal problem is presented.

2 The Projection Index

The projection index forms the heart of a projection pursuit method. It defines the intent of the procedure. Our intent is to discover interesting "structured" projections of a multivariate data set. This rather vague goal must be translated into a numerical index that is a functional of the projected data distribution. This functional must vary continuously with the parameters defining the projection and have a large value when the (projected) distribution is defined to be "interesting" and a small value otherwise. The notion of interesting will obviously vary with the application (see Huber, 1985). As stated in the introduction, our goal is to discover additional structure not captured by the correlational structure of the data. A way to insure this is to make the projection index invariant to all nonsingular affine transformations in the p–variable data space

(see Huber, 1985 and Jones, 1983).

Although the notion of interesting may be difficult to quantify, the converse notion of uninteresting seems more straightforward. Huber (1985) and Jones (1983) give strong heuristic arguments to the effect that the (projected) normal distribution ought to be considered the least interesting:

(1) All projections of a multivariate normal distribution are normal. Therefore, evidence for non-normality in any projection is evidence against multivariate joint normality. Conversely, if the least normal projection is - not significantly different from - normal then there is evidence for joint normality of the measurement variables. The multivariate normal density is elliptically symmetric and is totally specified by its linear structure (location and covariances).

(2) Even if there are several linear combinations of variables that are (possibly highly) structured (non normal), most linear combinations (views) will be distributed approximately normally. Roughly, this is a consequence of the central limit theorem (sums tend to be normally distributed). This notion is made precise by Diaconis and Freedman (1984).

(3) For fixed variance, the normal distribution has the least information (Fisher, negative entropy).

Following this view, any test statistic for testing normality could serve as the basis for a projection index. Different test statistics have the property of being more (or less) sensitive to different alternative distributions. It is the particular alternatives that are of interest here since (in the context of projection pursuit) they define the notion of an interesting distribution. We must choose a statistic that has preferential power against the (projected) distributions that we are seeking with our projection pursuit algorithm.

The most powerful tests for non normality emphasize alternative distributions with heavy tails. Our intent is to seek projected distributions that exhibit clustering (multimodality) or other kinds of concentrations near nonlinear manifolds. Such distributions differ from the normal mainly near the center of the distribution, rather than in the tails. We, therefore, seek a projection index (test statistic) that emphasizes departure from normality in the main body of the distribution and gives correspondingly less weight to he tails.

Since the projection index serves as the objective function for a multiparameter optimization, its computational properties are crucial. For a given set of parameter values, its value should be rapidly computable. It should be absolutely continuous so that at least its first derivatives (with respect to the parameters) exist everywhere. These derivatives should also be rapidly computable.

The most computationally attractive projection indices are based on polynomial moments. No sorting of the projected values is required, and the values of the polynomials as well as their derivatives are rapidly computed by means of recursion relations. Since we are interested in departure

from normality, it would be natural to base a projection index on the standardized (absolute) cumulants of the projected distribution (see Huber, 1985, p 445). These are the moments of the Hermite polynomials. Jones (1983) suggests a projection index based on sums of squares of standardized cumulants.

Despite their computational attractiveness projection indices based on standardized cumulants are (unfortunately) not useful for our application. This is because they very heavily emphasize departure from normality in the tails of the distribution. A (projected) distribution with only slightly heavier than normal tails receives a much higher index value than a highly clustered projection.

It is possible to base a projection index on moments having the required statistical properties. Such an index is developed below. The main idea is to change scale by transforming the projected data using the normal cumulative distribution function, and then comparing the transformed distribution to the uniform.

Following Huber (1985) the algorithm will be described first in its "abstract" version. That is we imagine it operating on a p–dimensional probability distribution. This makes some of the notation simpler. In our case the "practical" version, that is applied to data samples, is usually obtained by simply replacing the expected value operation by the corresponding data average. There are other minor differences that are pointed out at the appropriate places in the description. Random variable terminology will be used. Upper case letters will denote random variables and there lower case counterparts (usually with subscripts) will denote realized values in the sample.

As a first computational economy, we begin by "sphering" the data (Tukey and Tukey, 1981, Huber, 1981, Jones, 1983). The idea is to perform a linear transformation (rotation, location and scale change) that removes (incorporates) all the location, scale, and correlational structure. Let Y be a random variable in R^p. We perform an eigenvalue-eigenvector decomposition of the covariance matrix

$$\begin{aligned}\Sigma &= E[(Y - EY)(Y - EY)^T] \\ &= UDU^T\end{aligned}$$

with U an orthonormal and D a diagonal $p \times p$ matrix. We then define new variables

$$Z = D^{-\frac{1}{2}}U(Y - EY).$$

More specifically, let q be the rank of Σ. Then the q components of Z are given by

$$Z_j = \frac{1}{\sqrt{D_j}} \sum_{i=1}^{p} U_{ij}(Y_i - EY_i), \quad (1 \leq j \leq q). \tag{1}$$

The rows of U and D are assumed ordered in descending (nonnegative) values of D_j. By definition $E[Z] = 0$ and $E[ZZ^T] = I$, the identity matrix.

The computational advantage gained by sphering is reflected by the fact that data constraints in the N-dimensional observation space become geometrical constraints in the p-dimensional variable space. First, any linear combination

$$X = \alpha^T Z = \sum_{i=1}^{q} \alpha_i Z_i$$

has variance

$$var\ (X) = \alpha^T \alpha = \sum_{i=1}^{q} \alpha_i^2,$$

and thus enforcing the constraint

$$\alpha^T \alpha = 1 \tag{2}$$

insures that all linear combinations have unit variance. Second, two linear combinations based on orthogonal vectors are uncorrelated. That is $\alpha^T \beta = 0$ implies that $E[(\alpha^T Z)(\beta^T Z)] = 0$.

All operations are performed on the sphered variables Z. (Only at the end do we transform the solutions back to reference the original coordinates Y.) This frees us from having to compute variances in individual projections in order to standardize density estimates during the numerical optimization. Since sphering need only be done once at the beginning, this results in a substantial computational saving. By definition the Z variables are affine invariant, thus any projection index based solely on them will necessarily inherit this property.

Although in most exploratory applications two (or higher) dimensional projection pursuit will likely prove the more informative, we begin by describing our projection index for a one–dimensional projection pursuit. The concepts underlying the two algorithms are nearly identical and the notation is simpler for the one–dimensional case. The extension to two (and higher) dimensions is seen to be straightforward.

In a one–dimensional exploratory projection pursuit we seek a linear combination

$$X = \alpha^T Z \tag{3}$$

such that the probability density $p_\alpha(X)$ is relatively highly structured. As discussed above, we regard the (standard) normal as the least structured density and we are concerned with finding departures that manifest themselves in the main body of the distribution rather than in the tails. To this end we begin by performing a transformation

$$R = 2\Phi(X) - 1 \tag{4}$$

with $\Phi(X)$ being the standard normal cumulative distribution function

$$\Phi(X) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{X} e^{-\frac{1}{2}t^2} dt. \tag{5}$$

Clearly, R takes on values in the interval $-1 \leq R \leq 1$, and if X follows a standard normal distribution then R will be uniformly distributed in this interval. Specifically,

$$p_R(R) = \frac{1}{2} p_\alpha[\Phi^{-1}(\frac{R+1}{2})]/g[\Phi^{-1}(\frac{R+1}{2})],$$

with $g(X)$ being the standard normal density. Thus, a measure of non–uniformity of R corresponds to a measure of nonnormality of X. We take as a measure of nonuniformity the integral–squared–distance between the probability density of R, $p_R(R)$, and the uniform probability density, $p_U(R) = 1/2$, over the interval $-1 \leq R \leq 1$:

$$\int_{-1}^{1} [p_R(R) - 1/2]^2 dR = \int_{-1}^{1} p_R^2(R) dR - 1/2. \tag{6}$$

Our projection index $I(\alpha)$ is taken to be a moment approximation to (6).

Expanding $p_R(R)$ in Legendre polynomials we have

$$\int_{-1}^{1} p_R^2(R) dR - 1/2 = \int_{-1}^{1} [\sum_{j=0}^{\infty} a_j P_j(R)] p_R(R) dR - 1/2$$

where the Legendre polynomials are defined by

$$\begin{aligned} P_0(R) &= 1, \\ P_1(R) &= R, \text{ and} \\ P_j(R) &= [(2j-1)RP_{j-1}(R) - (j-1)P_{j-2}(R)]/j \end{aligned} \tag{7}$$

for $j \geq 2$. The coefficients are given by

$$a_j = \frac{2j+1}{2} \int_{-1}^{1} P_j(R) p_R(R) dR = \frac{2j+1}{2} E_R[P_j(R)]$$

so that

$$\int_{-1}^{1} p_R^2(R) dR - 1/2 = \sum_{j=1}^{\infty} (2j+1) E_R^2[P_j(R)]/2 \tag{8}$$

For a uniform distribution $U(-1,1)$, $E[P_j(R)] = 0$ for $j > 0$.

Our projection index is obtained by approximating the sum in (8) by its first J terms,

$$I(\alpha) = \sum_{j=1}^{J} (2j+1) E_R^2[P_j(R)]/2. \tag{9}$$

Note that this projection index measures departure from normality even if the J-term expansion is not an accurate approximation to $p_R(R)$, since it achieves its minimum value (zero) for X normally distributed (R uniform). Of course, for finite J the (projected) normal is not the unique minimizer of $I(\alpha)$. Any distribution of X that after the transformation (4) results in a distribution $p_R(R)$ with zero values for its first J Legendre polynomial moments will also be regarded as a least interesting distribution.

For a practical version of the algorithm operating on a data sample, the expected values in (9) are estimated by the corresponding sample averages. Substituting from (3) and (4) we have

$$\hat{I}(\alpha) = \frac{1}{2}\sum_{j=1}^{J}(2j+1)[\frac{1}{N}\sum_{i=1}^{N}P_j(2\Phi(\alpha^T z_i)-1)]^2 \tag{10}$$

as the sample version of our projection index. This is to be maximized with respect to the q–components of α under the constraint $\alpha^T\alpha = 1$. Details of the optimization procedure are given in Section 5.

The projection index (10) can be computed fairly rapidly. Fast approximations (to machine accuracy) for the normal integral (5) exist (see Kennedy and Gentle, 1980) and are provided as built–in intrinsic functions by many programming language compilers. The Legendre polynomials to order J are quickly obtained via the recursion relation (7).

For efficient optimization it is useful to have derivatives of the objective function. These are easily obtained for our projection index via the chain rule for differentiation. The result is

$$\frac{\partial I}{\partial \alpha_k} = \frac{2}{\sqrt{2\pi}}\sum_{j=1}^{J}(2j+1)E[P_j(R)]E[P_j'(R)e^{-X^2/2}(Z_k - \alpha_k X)]. \tag{11}$$

with X given by (3) and R given by (4). The derivatives of each Legendre polynomial with respect to its argument is rapidly obtained by the recursion relation

$$P_1'(R) = 1, \text{ and } P_j'(R) = RP_{j-1}'(R) + jP_{j-1}(R) \tag{12}$$

for $j > 1$. The derivative calculation (11) takes into account the contstraint $\alpha^T\alpha = 1$ by keeping the gradient vector $\nabla_\alpha I(\alpha)$ orthogonal to the gradient of the constraint function $\nabla_\alpha(\alpha^T\alpha) = 2\alpha$. This is the purpose of subtracting $\alpha_k X$ from Z_k in the second expectation (11). The derivatives of $\hat{I}(\alpha)$ (10) are obtained by applying sample averages in place of the expectation operators.

The projection index for a two–dimensional projection pursuit is developed in direct analogy with the one–dimensional index. We seek two linear combinations

$$X_1 = \alpha^T Z,\ X_2 = \beta^T Z \tag{13}$$

such that the joint distribution (probability density) $p_{\alpha\beta}(X_1, X_2)$ is highly structured. Since we are interested in nonlinear structure we require the two linear combinations to be uncorrelated, $\text{corr}(X_1, X_2) = 0$. As a consequence of our definition of Z (data sphering) this constraint is equivalent to requiring α and β to be orthogonal, $\alpha^T\beta = 0$. We must also require that the variances in all projections be equal. This is insured by the sphering and constraints $\alpha^T\alpha = \beta^T\beta = 1$.

We regard the bivariate standard normal to be the least structured joint distribution and are interested in departures that are manifest in the main body of the distribution rather than in the tails. We begin by transforming the X_1, X_2 plane to the square $(-1, 1) \times (-1, 1)$ via

$$\begin{aligned} R_1 &= 2\Phi(X_1) - 1 \\ R_2 &= 2\Phi(X_2) - 1 \end{aligned} \tag{14}$$

with $\Phi(X)$ defined by (5). If X_1, X_2 have a joint standard normal distribution then R_1, R_2 will be uniformally distributed on the square. We take as a measure of nonuniformity the integral–square–distance from the uniform

$$\begin{aligned} &\int_{-1}^{1}\int_{-1}^{1}[p_R(R_1, R_2) - \frac{1}{4}]^2 dR_1 dR_2 \\ &= \int_{-1}^{1}\int_{-1}^{1} p_R^2(R_1, R_2) dR_1 dR_2 - \frac{1}{4} \end{aligned} \tag{15}$$

Our projection index $I(\alpha, \beta)$ is taken as a product moment approximation to (15). Expanding $p_R(R_1, R_2)$ in a product Legendre expansion and proceeding in direct analogy with the development of the one–dimensional index we have

$$\begin{aligned} &\int_{-1}^{1}\int_{-1}^{1} p_R^2(R_1, R_2) dR - 1/4 \\ &= \frac{1}{4}\sum_{j=0}^{\infty}\sum_{k=0}^{\infty}(2j+1)(2k+1)E^2[P_j(R_1)P_k(R_2)] - \frac{1}{4} \end{aligned}$$

with the Legendre polynomials defined by (7). Our bivariate projection index is obtained by truncating the expansion at order J,

$$\begin{aligned} I(\alpha, \beta) = &\sum_{j=1}^{J}(2j+1)E^2[P_j(R_1)]/4 + \sum_{k=1}^{J}(2k+1)E^2[P_k(R_2)]/4 \\ &+ \sum_{j=1}^{J}\sum_{k=1}^{J-j}(2j+1)(2k+1)E^2[P_j(R_1)P_k(R_2)]/4. \end{aligned} \tag{16}$$

As for the univariate index, derivatives of the bivariate index are easily obtained:

$$\frac{\partial I}{\partial \alpha_m} = \frac{1}{\sqrt{2\pi}} \sum_{j=1}^{J} (2j+1) E[P_j(R_1)] E[P_j'(R_1) e^{-\frac{1}{2}X_1^2} (Z_m - \alpha_m X_1 - \beta_m X_2)]$$
$$+ \frac{1}{\sqrt{2\pi}} \sum_{j=1}^{J} \sum_{k=1}^{J-j} (2j+1)(2k+1) E[P_j(R_1) P_k(R_2)]$$
$$\cdot E[P_j'(R_1) P_k(R_2) e^{\frac{1}{2}X_1^2} (Z_m - \alpha_m X_1 - \beta_m X_2)]$$

$$\frac{\partial I}{\partial \beta_n} = \frac{1}{\sqrt{2\pi}} \sum_{k=1}^{J} (2k+1) E[P_k(R_2)] E[P_k'(R_2) e^{-\frac{1}{2}X_2^2} (Z_n - \alpha_n X_1 - \beta_n X_2)]$$
$$+ \frac{1}{\sqrt{2\pi}} \sum_{j=1}^{J} \sum_{k=1}^{J-j} (2j+1)(2k+1) E[P_j(R_1) P_k(R_2)]$$
$$\cdot E[P_j(R_1) P_k'(R_2) e^{-\frac{1}{2}X_2^2} (Z_n - \alpha_n X_1 - \beta_n X_2)]$$

The quantities X_1 and X_2 are given by (13) while R_1 and R_2 are given by (14). The data version of the index and its derivatives are obtained by substituting sample averages for the expectations. The derivatives account for the constraints ($\alpha^T\alpha = \beta^T\beta = 1, \alpha^T\beta = 0$) by keeping the gradient vector simultaneously orthogonal to the gradients of the three constraint functions.

The computation of the bivariate index is analogous to that of the univariate index. For a corresponding order J it is more expensive since it and its derivatives contain more terms. Also, the optimization is over twice as many parameters. On the other hand, the bivariate solutions often contain considerably more information concerning the multivariate density.

There is one (user defined) parameter associated with the (one- or two- dimensional) projection index. It is the order J (9) (16) of the polynomial approximation to the (transformed) density. It controls the amount of smoothness imposed on the approximation. Limited experience indicates that the results are insensitive to the value chosen for J over a fairly wide range ($4 \leq J \leq 8$) except for very small sample size. Intuition suggests that the value of J should increase with the sample size, but there are as yet no specific guidelines. The computation increases linearly with increasing J for one–dimensional projection pursuit and quadratically for two–dimensional projection pursuit.

3 Structure Removal

It is the purpose of the optimization algorithm (detailed below) to find a substantive maximum of the projection index. The corresponding (one– or two– dimensional) projection will (hopefully)

present an informative view of the p–dimensional data density. It is unlikely, however, that there is only one such informative view. Usually, the nonnormality of the full p–dimensional data distribution will be manifest in several one– or two– dimensional projections. Each of these projections can help in the identification and interpretation of the effects. Also, there is no reason to believe that the algorithm will find the most informative of these views first. For these reasons it is important that the projection pursuit procedure find as many of these informative views as possible.

A variety of approaches for accomplishing this have been suggested (see Huber, 1985, p 449). The most systematic of these (for one–dimensional projection pursuit) is the recursive approach associated with the projection pursuit density estimation procedure (Friedman, Stuetzle, and Schroeder, 1984). After an interesting projection has been found (solution maximizing the projection index), remove the structure that makes the projection interesting (deflate that maximum of the objective function) and then remaximize the projection index. This can be done repeatedly. In the projection pursuit density estimation approach this was implemented using a complex strategy for maintaining and updating an estimate for the p–dimensional probability density and involved Monte Carlo sampling. It is thus computationally quite expensive. We present a simple procedure for structure removal that is computationally much faster and in addition has a straightforward implementation for two–dimensional projections.

By definition of our projection index, a view (projected density) has zero interest if it is standard normal. Therefore, the structure can be removed by applying a transformation that takes the (projected) density to a standard normal distribution. We thus require a transformation of the q–variables, the result of which renders a standard normal distribution in the projected subspace, but leaves all orthogonal directions unchanged. We develop such a transformation below.

We first describe the procedure for a one dimensional projection. Let $X = \alpha^T Z$ be a one dimensional projection ($Var(X) = 1$) and $F_\alpha(X)$ be its cumulative distribution function. Then applying the transformation

$$X' = \Phi^{-1}(F_\alpha(X)) \tag{17}$$

to X results in a standard normal distribution for X'. Here Φ^{-1} is the inverse of the standard normal cumulative distribution function (5).

Let U be an orthonormal $(q \times q)$ matrix with α (the projection pursuit solution) as the first row. Then applying the linear transformation $T = UZ$ results in a rotation such that the new first coordinate is $T_1 = \alpha^T Z = X$. Let Θ be a (vector) transformation (with components $\theta_1 \cdots \theta_q$) that

takes T_1 to a standard normal distribution and is the identity transformation on $T_2 \cdots T_q$:

$$\theta_1(T_1) = \Phi^{-1}(F_\alpha(T_1))$$
$$\theta_j(T_j) = T_j \quad 2 \leq j \leq q. \tag{18}$$

Then the transformation

$$Z' = U^T \Theta(UZ) \tag{19}$$

transforms the projection $X = \alpha^T Z$ to a standard normal distribution leaving all orthogonal directions unchanged. We then reapply the maximization procedure with a projection index based on Z'.

By definition, the q–variate distribution of Z' exhibits no structure in the projection $X' = \alpha^T Z'$ (zero value of the projection index). The joint distribution of Z, $p(Z)$, and that of Z', $p'(Z')$, determine the same conditional density given $\alpha^T Z$:

$$p(\cdot|\alpha^T Z) = p'(\cdot|\alpha^T Z'). \tag{20}$$

In fact,

$$p'(Z') = p(Z') \frac{g(\alpha^T Z')}{p_\alpha(\alpha^T Z')} \tag{21}$$

with p_α the (univariate) density of $\alpha^T Z$ and g the standard normal density

$$g(X) = \frac{1}{\sqrt{2\pi}} e^{-X^2/2}. \tag{22}$$

In this sense the transformation (19) produces a new (vector valued random) variable Z' whose distribution is as close as possible to that of Z under the constraint that its marginal distribution along α be normal (zero interest). It also produces the closest distribution under this constraint in the sense of the relative entropy distance measure

$$\int \log(\frac{p}{p'}) p dZ = \int \log(\frac{p_\alpha}{g}) p_\alpha d(\alpha^T X) = \min \tag{23}$$

(see Huber, 1985).

The data sample version of (17) is easily implemented. One substitutes the empirical distribution $\hat{F}_{\alpha N}(X)$ $(X = \alpha^T Z)$ for the distribution function $F_\alpha(X)$:

$$x'_i = \Phi^{-1}(\hat{F}_{\alpha N}(x_i)) = \Phi^{-1}(\frac{r(x_i)}{N} - 1/2N) \tag{24}$$

with $r(x_i)$ being the rank of x_i among the N (projected) observations. This transformation simply replaces each observation by its corresponding normal score in the projection ("Gaussianization").

This process of repeatedly applying projection pursuit on the (structure removed) output of the previous pursuit can be continued until several applications result in funding no additional interesting structure. It should be noted that "Gaussianizing" a solution projection in this way at a particular stage, modifies the normality of (nonorthogonal) previous solution projections so that they no longer have exactly zero interest (unless backfitting is employed - see Friedman, Stuetzle and Schroeder, 1984). However, the stucture induced in previous solution projections, by structure removal at later stages, is small.

Structure removal in two dimensions is more difficult. We need a transformation that takes a general bivariate distribution $p_{\alpha\beta}(X_1, X_2)$ to the bivariate standard normal

$$g(X_1, X_2) = \frac{1}{2\pi} e^{(X_1^2 + X_2^2)/2}. \tag{25}$$

There is no difficulty in theory. One can transform one of the margins, say X_1, to normality via (17) and then transform each conditional orthogonal marginal $p(X_2|X_1)$ to normality (again via (17)). However, this prescription does not lead to a practical algorithm for application to (bivariate) data. A practical algorithm can be based on the observation that all projections of a normal distribution are normal. The idea is to repeatedly Gaussianize rotated (about the origin) projections of the solution plane until they stop becoming more normal.

Let

$$\begin{aligned} X_1' &= X_1 \cos\gamma + X_2 \sin\gamma \\ X_2' &= X_2 \cos\gamma - X_1 \sin\gamma \end{aligned} \tag{26}$$

be a rotation about the origin through angle γ. The distributions of X_1' and X_2' are then each transformed to normality via (17). This process is repeated (on the previously transformed distributions) for several values of γ $(0, \pi/4, \pi/8, 3\pi/8)$. This entire process is then repeated until the distributions stop becoming more normal. Any convenient index of (non) normality can be used.

During the first few iterations the non normality decreases rapidly in a monotonic fashion as the planar distribution approaches joint normality. After approximate normality has been achieved, the value of the normality index tends to oscillate with small amplitude on successive iterations, sometimes decreasing a small amount on the average. Convergence is defined when approximate stability has been achieved. Note that with finite samples absolute stability is impossible to achieve. Typically, the procedure takes 5 to 15 complete iterations to converge. It produces bivariate data distributions that are quite close to normal.

In analogy with the univariate case let U be an orthonormal $(q \times q)$ matrix with α and β (the linear combinations determining the solution plane) as the first two rows. The linear transformation $T = UZ$ performs a rotation aligning the first two new coordinates with α and β. Let Θ be a transformation that takes the joint distribution of T_1 and T_2 to standard normal (as described above) and is the identity transform on $T_3 \cdots T_q$. Then the transformation

$$Z' = U^T \Theta(UZ)$$

transforms the solution (α, β) plane to bivariate standard normal leaving all orthogonal directions unchanged. Thus, the joint distribution of Z and Z' determine the same conditional distribution given $\alpha^T Z$ and $\beta^T Z$,

$$p'(Z') = p(Z') \frac{g(\alpha^T Z') g(\beta^T Z')}{p_{\alpha\beta}(\alpha^T Z', \beta^T Z')} \tag{27}$$

with $g(X)$ the standard normal and the denominator the joint distribution of $\alpha^T Z$ and $\beta^T Z$.

Friedman and Tukey (1974) suggested two rudimentary forms of structure removal. One was to restrict later solutions to be orthogonal (with respect to the original coordinates and their scales) to previous solutions. This is clearly supplanted by the structure removal technique outlined here. There is no reason to expect good views of the data to be orthogonal with respect to any prespecified metric. The second suggested method was applicable when the structure in the solution projection took the form of clustering. The idea was to isolate the clusters into separate subsamples and apply projection pursuit to each such isolate individually. This could be iterated if clustering was found in subsequent solutions. This second structure removal technique (when applicable) can be viewed as complementary to the method outlined here. If there is clustering and it is largely hierarchical, then the isolation technique can provide a straightforward means for interpreting this kind of structure.

4 Density Estimation

Exploratory projection pursuit as described in the preceeding two sections can be incorporated into a multivariate density estimation procedure. Its properties (for one-dimensional projection pursuit) are similar to the projection pursuit density estimation procedure of Friedman, Stuetzle, and Schroeder (1984) and the projection pursuit density approximation techniques of Huber (1985). However, its computational aspects are considerably more attractive.

The projection pursuit strategy outlined in the preceeding two sections (distribution version) consists of finding the least normal projection $p_{\alpha_1}(\alpha_1^T Z)$ of the probability density $p(Z)$ by maximizing a measure of nonnormality (9). The procedure is then repeated on the density

$$p_1(Z) = p(Z) g(\alpha_1^T Z) / p_{\alpha_1}(\alpha_1^T Z)$$

(see (21)), obtaining a second solution $\alpha_2^T Z$. The distribution is again modified

$$\begin{aligned} p_2(Z) &= p_1(Z) g(\alpha_2^T Z) / p_{\alpha_2}^{(1)}(\alpha_2^T Z) \\ &= p(Z) \frac{g(\alpha_1^T Z) g(\alpha_2^T Z)}{p_{\alpha_1}(\alpha_1^T Z) p_{\alpha_2}^{(1)}(\alpha_2^T Z)} \end{aligned}$$

and so on. Here $p_{\alpha_2}^{(1)}(\alpha_2^T Z)$ is the univariate marginal density of $\alpha_2^T Z$ under the joint density $p_1(Z)$. At the Kth iteration one has

$$p_K(Z) = p(Z) \prod_{k=1}^{K} \frac{g(\alpha_k^T Z)}{p_{\alpha_k}^{(k-1)}(\alpha_k^T Z)}. \tag{28}$$

The quantity in the denominator, $p_{\alpha_k}^{(k-1)}(\alpha_k^T Z)$, is the marginal density of $\alpha_k^T Z$ under the joint distribution $p_{k-1}(Z)$, with $p_0(Z) = p(Z)$.

At some point in the iterative process the projection pursuit algorithm cannot find a projection that deviates substantially from normality. This indicates that $p_K(Z)$ is approximately multivariate standard normal. We then take as our density approximation

$$\bar{p}(Z) = g(Z) \prod_{k=1}^{K} \frac{p_{\alpha_k}^{(k-1)}(\alpha_k^T Z)}{g(\alpha_k^T Z)} \tag{29}$$

with

$$g(Z) = \frac{1}{(2\pi)^{q/2}} e^{Z^T Z/2} \tag{30}$$

The projected univariate densities $p_{\alpha_k}^{(k-1)}$ can be approximated by any appropriate method. One possibility is to use the Legendre polynomial expansion associated with the projection index

$$p_{\alpha_k}^{(k-1)}(\alpha_k^T Z) = g(\alpha_k^T Z) \sum_{j=0}^{J} (2j+1) E_{k-1}[P_j(R_k)] P_j(R_k)/2$$

with $R_k = 2\Phi(\alpha_k^T Z) - 1$ and $E_{k-1}(\cdot)$ the expected value under $p_{k-1}(Z)$. Truncation can be used to insure nonnegativity. Substituting this into (29) we have for this (multivariate) density approximation

$$\tilde{p}(Z) = g(Z) \prod_{k=1}^{K} [\sum_{j=0}^{J} (2j+1) E_{k-1}[P_{jk}] P_j(2\Phi(\alpha_k^T Z) - 1)]/2 \tag{31}$$

with $E_{k-1}[P_{jk}]$ being the expected value of the associated (adjacent) Legendre polynomial under $p_{k-1}(Z)$. A density estimate is obtained by estimating $E_{k-1}[P_{jk}]$ by sample averages over the transformed variables $Z_{k-1} = U_{k-1}^T \Theta_{k-1}(U_{k-1} Z_{k-2})$ [see (19)] obtained from the structure removal process during the projection pursuit. Thus,

$$\hat{E}_{k-1}[P_{jk}] = \frac{1}{N} \sum_{i=1}^{N} P_j[2\Phi(\alpha_k^T z_{(k-1)i}) - 1]. \tag{32}$$

Here $Z_0 = Z$, the original (sphered) data.

This density approximation/estimate is strongly influenced by the main body of the data and will give poor (usually under) estimates in the outlying tails. This is a result of the transformation (4) which compresses the tails into small intervals near the extremes of the interval $(-1, 1)$. Long tailed (compared to the normal) projected (univariate) distributions will result in very sharp spikes in the transformed density $p_R(R)$ at the ends of the interval. These cannot be captured by a low to moderate degree ($4 \leq J \leq 8$) Legendre polynomial expansion which will substantially underestimate them. This is how the projection index (9), (10), and (16) achieves its long tailed robustness. Also, of course, the projection index (by design) will tend not to produce solutions for which the projected density has long tails. As a result the density approximation/estimate provided by (31) and (32) will focus on capturing the density variation in the central part of the distribution and will approximate its tails by the tails of the normal. Of course, there is no other method that produces accurate density estimates in the tails of a multivariate distribution either. It is interesting to note the connection between this approach to projection pursuit density approximation and the "analytic" approach proposed by Huber (1985).

It is possible to base a density approximation/estimation procedure on the two–dimensional algorithm in direct analogy with the development of the one–dimensional procedure described above. However, it might not work as well as the one–dimensional algorithm (for this purpose) due to its increased complexity.

5 Optimization Strategy

Although an engineering detail, the technique used for maximizing the (one- and two- dimensional) projection index strongly influences both the statistical as well as the computational aspects of the procedure. The statistical power of the method is reflected in its ability (for a given sample size and data dimension) to find substantive maxima of the projection index. As observed by Switzer (1970) there are "an almost inevitable multiplicity of decidedly suboptimal local maxima" mostly caused by sampling fluctuations. This can distract a projection pursuit algorithm from finding important views (substantive maxima). These pseudo maxima can be visualized as a high frequency ripple superimposed on the main variational structure of the objective function (projection index). The amplitude of these ripples increases with increasing dimension and decreasing sample size. The extent to which the optimization procedure can ignore ("step over") these pseudo-maxima, and thus avoid being trapped by them, determines to a great extent its statistical power.

On very smooth objective functions the most powerful optimization methods (steepest descent, congugate gradients, quasi-Newton, see Gill, Murray, and Wright, 1981) involve the use of first derivatives. This is why the ability to rapidly compute derivatives was a design goal for our projection index. These methods very effectively (rapidly and accurately) find the first maximum of an objective function uphill from a starting point. Unfortunately, when applied to a projection index with the ripple phenomenon described above, this will very likely be a pseudo-maximum, unless the starting point is within range of a substantive maximum. The optimization strategy used by Friedman and Tukey (1974) did not employ (exact) derivatives and took fairly large steps in its search for a maximum. This gave it some robustness against pseudo-maxima at the expense of considerable computational effort.

We employ a hybrid optimization strategy. It begins with a simple (course stepping) optimizer that is designed to very rapidly get close (within range) of a substantive maximum. A gradient method (quasi-Newton) is then used to quickly converge to the solution.

We begin with the maximization of the one–dimensional index $I(\alpha)$ with respect to the q components of α $(\alpha_1 \cdots \alpha_q)$. As a first step $I(\alpha)$ is maximized over the coordinate axes $\alpha = e_i$ $(1 \leq i \leq q)$. Note that since we are working with the sphered data, Z, these axes are in fact the principal component directions when referenced to the original data, Y (1). Let α^* be the resulting maximizing axis. Starting with this direction the following optimization algorithm is performed:

Loop:

$I_0 = I(\alpha^*)$
For $i = 1$ to q do:
$f_+ = I[\frac{1}{\sqrt{2}}(\alpha^* + e_i)/(1 + \alpha_i^*)^{1/2}]$
$f_- = I[\frac{1}{\sqrt{2}}(\alpha^* - e_i)/(1 - \alpha_i^*)^{1/2}]$
If $f_+ > f_-$ then $f = f_+$; $s = +1$ (33)
else $f = f_-$; $s = -1$
end If
If $f > I(\alpha^*)$ then
$\alpha^* \leftarrow \frac{1}{\sqrt{2}}(\alpha^* + s \cdot e_i)/(1 + s \cdot \alpha_i^*)^{1/2}$
end If
end For
If $I(\alpha^*) = I_0$ then done

end Loop

This search algorithm takes large steps and thus it cannot be expected to converge with the value of α^* at a maximum of $I(\alpha)$. However, because of its course stepping, it is much less likely than a gradient method to be trapped on pseudo–maxima, thereby allowing it to converge in the vacinity of a substantive maximum. This algorithm typically requires two to four passes over the

coordinates (executions of the For loop) to converge.

Starting with the value of α^* obtained upon convergence of (33) a gradient directed optimization method is then employed to rapidly ascend to a maximum of the projection index $I(\alpha)$. We have employed both steepest–ascent and quasi–Newton methods with comparable results.

The maximization of the two–dimensional index $I(\alpha, \beta)$ is done in an analogous manner. First, it is maximized over the $q(q-1)/2$ pairs of coordinate axes, $\alpha = e_i \quad \beta = e_j$ $(2 \leq i \leq q, 1 \leq j < i)$. Then starting with the best coordinate pair an algorithm analogous to (33) is executed. In this algorithm the For loop is over $2q$ variables (the q components of α and the q components of β) and the constraint $\alpha^T \beta = 0$ must be maintained in addition to $\alpha^T \alpha = \beta^T \beta = 1$. Finally, after this procedure converges, a gradient directed optimization is employed to rapidly find the maximum.

6 Remarks

6.1 Robustness

The one– and two– dimensional projection indices (10) and (16) are (by design) quite robust in that they are largely unaffected by extreme outliers. As a consequence the pursuit procedure is thus similarly robust. Structure removal is also clearly unaffected by outliers. The only non–robust aspect of the procedure is the data sphering. It is based on the sample covariance matrix which is strongly influenced by extreme outliers. Experience indicates this projection pursuit procedure does not seem to be severely degraded when based on badly sphered data due to outliers. Nevertheless it seems sensible to use robust sphering when possible.

There are several methods for robust estimation of a covariance matrix (see Devlin, Gnanadesikan, and Kettering, 1981). In fact there are serveral attractive (but computationally expensive) projection pursuit approaches (Chen and Li, 1981). We have implemented a simple multivariate trimming method. It begins by sphering using all of the data. All observations for which $Z^T Z > D$ (see (1)) are deleted and the remaining data is resphered. Here D is some prespecified threshold conveniently taken to be a high ($\sim 1 - 0.01/N$) quantile of the chi-squared distribution on q–degrees–of–freedom. This procedure can be iterated until no observations are deleted. Often D is adjusted so that no more than a certain (small) fraction of the data are deleted.

6.2 Preliminary dimensionality reduction

The power of the projection pursuit algorithm to find important structure decreases with decreasing sample size and increasing dimension (see next section). As remarked earlier, the covariance structure (linear associations) often does not align with the nonlinear structure (clustering, concentrations near nonlinear manifolds) that we are seeking with our projection pursuit algorithm.

A typical exception to this, however, has to do with the existence of a subspace containing only a tiny fraction of the data variation.

Clearly, if a subspace contains no data variation, it cannot contain any structure. In this case the covariance matrix is singular and the dimension of the search space is reduced to $q < p$ (1), the rank of the covariance matrix. If there exists a $p - q$ dimensional subspace for which the data variation is very small compared to the complement subspace (covariance matrix nearly singular), then this subspace is usually dominated by the noise in the system and contains little data structuring. If this turns out to be the case, then the power of the projection pursuit procedure can be enhanced (and computation reduced) by restricting the pursuit search to the q dimensional complement space. If not, any structure represented in the $p - q$ dimensional subspace will be ignored. However, in cases where the data dimension is very high and the sample size small, there may be no choice but to restrict the projection pursuit search to the subspace spanned by the largest q principal component axes, where q is determined by the sample size. Also, if a high dimensional projection pursuit is unsuccessful in finding interesting structure, it might be worthwhile to restrict the search dimension (as described above) and try again.

6.3 Preliminary transformations

Sometimes marginal distributions on the original measurement coordinates exhibit considerable (nonnormal) structure. For example, substantial skewness is often associated with quantities that take on only positive values. Since inspection of the coordinate marginals should always be among the first parts of any data analysis, this structure is easily discovered. Often the data analyst would like to know if there is additional structure associated with combinations of the variables. In this situation it makes sense to perform a transformation on highly structured coordinates, to remove the obvious structure, and then apply projection pursuit to the data after these selected transformations. For example, taking logarithms often removes positive skewness (see Mosteller and Tukey, 1977, Ch. 5, for a catalog of such "first aid" transformations). Of course, the structure along any marginal can be completely removed (from the point–of–view of projection pursuit) by replacing the coordinate values by their corresponding normal scores. Such "Gaussianized" variables would then only contribute to data structuring through their (nonlinear) associations with other variables.

Sometimes measurement variables take on only a small number of distinct values. This can be due to the nature of the variables themselves or a consequence of the measuring process. If the number of such values is small (compared to the sample size), the marginal distribution will

exhibit many identical values or ties. If the number of distinct values is very small (less than five or so), the marginal distribution appears highly structured when compared to the normal. Gaussianization of such variables is a possible remedy; however, it is important that observations with the same value be ordered randomly so that associations between variables are not induced by the fact the original ordering of the observations may be associated with the values of some of the measurement variables. Categorical (nominal) variables can be accomodated by the introduction of corresponding (zero/one) dummy variables along with randomly breaking of the resulting ties and subsequent Gaussianization.

6.4 Significance

It is important to know whether a view is indicative of actual structure in the population or whether it is an artifact of sampling fluctuations. One way to access this is to compare the corresponding solution projection index to values obtained by applying the procedure to Gaussian data. One can repeatedly generate random samples from a Gaussian distribution of the same dimension and cardinality as the data sample. The identical procedure that was applied to the data can then be applied to these Monte Carlo multivariate normal samples. A comparison of the resulting (null) distribution of projection index values to the data sample value gives an indication of its significance.

6.5 Adjusted data plots

With the exception of the first, there are two projections of interest associated with each projection pursuit solution. One is the distribution of the data projected onto the solution line or plane. The other is the projection of the transformed data after removal of the structure associated with all previous solutions. In distributional terms, the former projection is the (joint) distribution of $\alpha^T Z$ (and $\beta^T Z$) under the original joint data density $p(Z)$. The latter projection is that distribution under $p_{K-1}(Z)$ (28), or the corresponding two–dimensional analog (see (27)), where K is the iteration number. At the Kth iteration projection pursuit is applied to $p_{K-1}(Z)$ in order to find additional structure. The Kth solution projection index and the resulting projection of the transformed data reflect the additional structure adjusted for (not directly associated with) all previous solutions. We refer to these as "adjusted" data plots.

6.6 Interpretation

The output of an exploratory projection pursuit is a collection of views of the multivariate data set. These views are selected to be those that independently best represent the nonlinear

aspects of the joint density of the measurement variables as reflected by the data. The nonlinear aspects are emphasized by maximizing a robust affine invariant measure of nonnormality, while the independence is induced by the structure removal process. The data analyst has at his disposal the values of the parameters (variable loadings) that define each solution (line or plane) as well as the projected data density. This information can be used to try to interpret any nonlinear effects that might be uncovered. A visual representation of the projected data density (histogram, smoothed density estimate, scatter or contour plot) can be inspected in order to ascertain the nature of the effect (clustering - concentration near nonlinear manifold). The (scaled) variable loadings that define the corresponding solution indicate the relative strength that each (corresponding) variable contributes to the observed effect.

In a two-dimensional projection pursuit the visual impact of the projected data density is insensitive to a particular orientation (within the plane) of the orthogonal axes used to define the solution plane. A rigid rotatin of the projected data about the origin of the solution plane provides the same picture of the nonlinear effects. It therefore makes sense to orient the defining axes so that the resulting variable loadings are most easily interpreted. This usually means maximizing the variance (or some other dispersion measure) of the (normed) variable loadings so as to give large loadings to as few variables as possible. Note that this "varimax" rotation is performed on the defining axes in the sphered variable representation Z, whereas the criterion to be maximized is the variance of the corresponding (normed) coefficients in the original data variables Y (see (1)). Experience indicates that a most useful varimax rotation is one that maximizes the variance of the loadings associated with one of the defining axes (e.g. the vertical axis).

Often projection pursuit solutions give rise to small loadings on several variables. If all (original) variables Y_j $(1 \leq j \leq p)$ have similar scales, then those with small loadings have correspondingly less effect in defining the solution. For interpretational purposes it is often important to know whether these variables have any importance to the observed structure. This is most easily determined by (manually) setting the small coefficients to zero (in perhaps a reverse stagewise manner) and then reprojecting the data.

6.7 Three- and higher-dimensional projection pursuit

In the preceeding sections a one– and two– dimensional exploratory projection pursuit algorithm were described. For data exploration the two–dimensional algorithm is likely to prove the most useful owing to the increased richness of structure that can be represented in two dimensions. In principal there is no upper limit to the dimensionality of the solution subspace. One

could envision a projection pursuit for finding informative three– and higher– dimensional views, although it is not clear that the richness of the representation would increase as much as in going from one to two dimensions. Three dimensional representations can be viewed using kinematic graphic techniques such as rigid rotation (see Fisherkeller, Friedman, and Tukey, 1975, McDonald, 1984, and Donoho, Donoho, and Gasko, 1986). There are techniques for (approximate) viewing of densities in four dimensions (see Tukey and Tukey, 1985). Among the more promising approaches are the Grand Tour methods (Buja and Asimov, 1985).

A projection index for higher–dimensional pursuit is easily developed in analogy with the two–dimensional index. The computational expense would be greater owing to the increased complexity of the product Legendre polynomial expansion and the increased number of optimization parameters. Structure removal in higher dimensions is, however, a much more serious problem. The difficulty lies in transforming the (projected) distribution to joint standard normality. A strategy analogous to that for the two–dimensional case would require a great many directions if they are chosen regularly. A good strategy would be to choose a carefully selected set of directions that depend on the actual (projected) data density. This is accomplished by running a one–dimensional projection pursuit algorithm in the higher dimensional projected (solution) subspace. As discussed in Section 4, this in fact constructs a transformation of the original data density to standard normality.

7 Examples

In this section we present the results of running the one– and two– dimensional projection pursuit algorithms on data. The first two examples are simulation studies in which we try to access the sample size requirements for detecting (known) structure as a function of increasing data dimensionality. The next three examples show the results of applying two–dimensional projection pursuit to real data sets of varying dimension and cardinality. In all examples no robustification was introduced into the sphering. To aid in interpretation all variables were standardized to zero mean and unit variance ("auto-scaled") before projection pursuit was applied. In the applications of two-dimensional projection pursuit the solutions were rotated to maximize the variance of the loadings on the second (vertical) defining axis (see Section 6.6). Except for the first real data example (states data - Section 7.3), the order of the Legendre expansion J (see (10) and (16)) was taken to be $J = 6$.

7.1 Single clustered projection in several dimensionalities

The purpose of this study is to get an idea of how the sample size requirements for finding a

single structured projection increase with the dimensionality of the data sample. The population for this study is a Gaussian mixture. Two–thirds of the data are generated from a joint standard normal distribution while the remaining one–third are normal with unit covariance matrix, but with location displaced six units in a random direction. The data are then scaled to have unit variance in this direction so that the structure is not reflected in the linear associations amongst the variables.

Three experiments were performed at dimensionalities five, ten and fifteen, respectively. Since (by design) the data structuring appears in only one view (the direction defined by the difference of the means), this example tests the projection pursuit algorithm's ability to find structure in the presence of an increasing number of pure noise variables. From the point of view of projection pursuit this represents a difficult example since the structure appears in only a single projection. Figure 1 shows a histogram of a random sample of size 200 from this population projected onto the solution direction.

At each dimensionality a series of one–dimensional projection pursuit runs were made to get a rough idea of the threshold sample size at which the algorithm could reliably find the (known) structured projection. Since (by design) the projection pursuit algorithm has some difficulty at these (threshold) sample sizes, a measure of that difficulty is the iteration number (projection pursuit followed by structure removal) at which the algorithm discovers the known structure as opposed to spurious structure (pseudo- maxima) induced by the small sample size and/or high dimensionality. If for a given sample size and dimensionality, the algorithm repeatedly finds the known structure at the first iteration, then it is having little difficulty. If, on the other hand, it finds several pseduo maxima (which are subsequently deflated by structure removal) before finding the real structured projection, this is an indication of some difficulty.

Figures 2-4 show the distribution of the iteration number at which the true (population) structured projection was found for ten random samples for each of six situations. Each situation consists of a specific dimensionality and sample size. Two sample sizes are shown for each dimensionality. The first (Figs. 2a, 3a, 4a) is a smaller sample size at which the algorithm seems to be having some difficulty and thus represents a minimal cardinality for finding the true structured projection at the corresponding dimensions. The second (larger) sample size (Figs. 2b, 3b, 4b) is seen to be large enough to find the true underlying structure fairly reliably. In all runs the determination as to whether the algotihm found the actual underlying structure was unambiguous. The projection index associated with this solution was typically four to five times that of the spurious solutions (pseudo-maxima) and the solution direction lined up very closely with the direction associated with

the underlying (population) optimal projection. It seems that once the optimizer gets close to the true solution (via the course stepping algorithm) the gradient directed search locks on to it very accurately (and rapidly).

As seen from the figures, the required sample size increases with dimensionality fairly rapidly, but still much slower than the exponential rate associated with the "curse of dimensionality." A qualitative explanation of why the increase is as rapid as it is has to do with the numerical optimization. In most statistical methods the size of the spurious structure associated with sampling fluctuations has to be comparable to that of the real underlying (population) structure to cause trouble. Here it need only be large enough (and numerous enough) to trap the numerical optimizer. Nevertheless the sample size requirements are seen to be fairly modest for a search dimension of $q \leq 10$. For large samples search dimensionalities of up to $q = 15$ or larger are possible (see section 6.2).

7.2 Needle in a hay stack

The population for this example is again a Gaussian mixture. In this case, however, the two components of the mixture have the same location but different covariance structure. A random sample of size 175 is drawn from a ten dimensional standard normal. Added to this is a sample of 25 observations which are standard normal in a four dimensional subspace (through the origin) and spherical normal with covariance matrix 0.0025 times the identity matrix in the six dimensional complement subspace. As in the previous example the data are scaled so that this structure is not reflected in the covariances of the combined data. The problem is to discover the presence of the small (25 observation) four–dimensional needle in a ten–dimensional haystack, in analogy with finding a one–dimensional needle in a three–dimensional haystack.

To this end the one–dimensional projection pursuit algorithm was applied to these data. This problem is difficult owing to high dimensionality (of the haystack) and the small cardinality of the needle. On the other hand the needle is visible in any projection that is orthogonal to its four dimensional subspace. Figure 5a shows such a projection of these data.

Figure 5b shows the distribution of the iteration number at which the projection pursuit algorithm found the needle in ten random samples from this Gaussian mixture. As in the previous example the determination of this was unambiguous. The results indicate that the algorithm was fairly well able to find a needle of this size. When the size of the needle is increased to 40 observations (out of 200), the algorithm always found it on the first iteration.

7.3 States data

The data for this example are seven summary statistics associated with each of the 50 United States (Becker and Chambers, 1984). Table 1 describes each of the seven variables. Two-dimensional projection pursuit was applied to this data set. The results of the simulation study above (Section 7.1) indicate that the sample size ($N = 50$) is small for a projection purusit in seven dimensions. The eigen-expansion of the correlation matrix shows that 92% of the (auto-scaled) data variance is captured by the first four eigenvalues. We therefore restrict the projection pursuit to the subspace spanned by the first four principal components ($q = 4$). Owing to the small sample size, it is unlikely that we will be able to detect very fine structure so the order of the Legendre expansion J (see 16) was set to $J = 2$.

In order to get an idea of the significance of the resulting solutions the identical procedure was applied to (different) random samples of size 50 drawn from a seven dimensional standard normal distribution. An ordered list of the solution projection indices for 20 such (null) runs is given in Table 2a.

When applied to the states data, four iterations of two-dimensional projection pursuit produced solution projection indices of 0.19, 0.08, 0.025, and 0.023 respectively. When referenced to the (estimated) null distribution represented in Table 2a, only the first value is seen to appear significant (at 5%). Table 3 presents the (varimax rotated) variable loadings (α and β) associated with the linear combinations defining this solution plane. Figure 6 shows the data projected onto the solution plane.

Viewing Figure 6 shows that the data appear to divide into two clusters mainly along the vertical direction. Also two outliers are seen in the upper right hand corner. A smaller cluster of 12 states seems fairly well separated from the larger group of 38 states. Tables 1 and 3 show that the dominant loadings comprising the vertical direction (β) involve income, high school graduation rate, and (negatively) illiteracy. This index seems to divide the states into two fairly distinct groups. The horizontal axis (α) is dominated by (negative) population, (negative) life expectancy, and (positive) homicide rate. Thus, increasing values along the horizontal direction involve generally lower population and life expectancy, and increasing homicide rate. The states in the upper cluster have a generally lower value of the horizontal index than those in the lower one with the dramatic exception of the two outliers in the upper right hand corner.

Table 4 lists the states comprising the smaller cluster in decreasing value of the vertical index. The extreme outlier in the upper right hand corner is Alaska while the other outlying point (closest

to it) corresponds to Nevada.

7.4 Automobile data

This data set consists of 10 characteristics associated with 392 automobile models sold in the United States and reported in Consumer Reports from 1972 to 1982 (Donoho and Ramos, 1982). Table 5 lists the ten variables. The last three are dummy variables for the manufacturing origin of the automobile. These dummy variables have only two distinct values while the second variable has ony five distinct values (the value three was encoded for rotary engine automobiles). Following the discussion in the last part of Section 6.3, we Gaussianize these variables after randomly ordering all observations corresponding to the same value. Table 2b lists in order of ascending value the (null) projection index values obtained by two-dimensional projection pursuit on 20 random samples of size $N = 392$ drawn from a $p = 10$ dimensional standard normal distribution.

Six iterations of two-dimensional projection pursuit on the automobile data produced projection index values of 0.31, 0.15, 0.13, 0.065, 0.11, and 0.088 respectively. All but the smallest value seem significant. The solutions corresponding to the largest two projection index values are presented in Table 6 and Figure 7. Table 6 shows the (varimax rotated) loadings (α and β) defining the two solution planes. Figure 7a shows the data projected onto the first solution plane, while Figures 7b and 7c show the adjusted and original data plots corresponding to the second solution (see Section 6.5).

The first solution exhibits a strong clustering along the vertical index (approximately twice engine size minus fuel inefficiency) especially for moderate values of the horizontal index (approximately engine size minus weight). The second solution displays a distinctly trimodal distribution. Note that this structure is not a direct reflection of the clustering shown in the first solution owing to the structure removal.

7.5 Boston Neighborhood Data

This compilation of census data (Harrison and Rubenfeld, 1978) is on its way to becoming a standard test bed for multivariate procedures. It is published in its entirety in Belsey, Kuh and Welsch (1980). Each observation is a neighborhood (standard metropolitan statistical area) in the Boston area. Associated with each of these 506 census tracts are 13 summary statistics that form the variables associated with each observation. Table 7 lists the quantities that comprise the variable set.

This data is well known to contain striking structure much of which is exhibited in the coordinate marginal distributions. Following the discussion in Section 6.3 we removed the most

obvious of this structure (extreme skewness in Y_1 and Y_{11}) by the transformations $Y_1' = \log(Y_1)$ and $Y_{11}' = \log(0.4 - Y_{11})$.

As with the previous examples we first obtain an estimate for the null distribution of projection index values by running two–dimensional projection pursuit on 20 random samples of size 506 from a 13 dimensional standard normal distribution. An ordered list of the values so obtained is shown in Table 2c. Note that none of the 20 values is greater than 0.07. Running ten iterations of two–dimensional projection pursuit on the Boston neighborhood data produced solution projection index values of 0.69, 0.51, 0.40, 0.25, 0.34, 0.26, 0.31, 0.20, 0.22, and 0.10 respectively. Clearly, all of these values but the last are highly significant when referenced to the null distribution (Table 2c). As the high projection index values indicate all of these views (but the last) exhibit striking structure. In the interest of brevity only the first five are shown. Table 8 lists the solution linear combinations, α and β, defining the solution planes for each of the five solutions.

Figure 8 shows the data projected onto the first solution plane. Figures 9a, b through 12a, b show both the adjusted and original data projections for each of the subsequent solutions. The first solution shows the data dividing into two groups on the second ("big lots") variable. Even after this structure is removed, the adjusted plots of the subsequent solutions show that there is considerable (additional) clustering and other nonlinear effects. In this case, projection pursuit has provided a great many views with which to begin exploring and trying to understand the data.

8 Discussion

The examples of the previous section are intended to illustrate that the exploratory projection pursuit procedures developed in the earlier sections can effectively discover nonlinear data structuring in fairly high dimensionality with practical sample sizes. The aspects contributing to this are a projection index that measures non-normality in the main body of the distribution rather than in the tails, an optimization algorithm that combines course stepping followed by gradient directed optimization, and an effective technique for structure removal. This algorithm is also much faster than the Friedman and Tukey (1974) implementation owing to the superior optimization procedure, but much more importantly to the rapidly computable form of the projection index (and its derivatives) using the (Legendre polynomial) moment expansion. This should help make the method more practical to those with fairly modest computing resources.

A powerful aid in interpreting the output of this projection pursuit procedure would be a means for connecting the various solution plots (views of the data) so that particular observations or groups of observations in one plot could be identified in the other views. One could then easily

identify hierarchical clustering as well as many more types of complex structure from the several views provided by the different projection prusuit solutions. With a color terminal supporting dynamic graphics one could use the color m and n plotting technique (McDonald, 1984) to great advantage. With a black and white terminal (again supporting dynamic graphics) the scatterplot brushing techniques (Becker and Cleveland, 1985) would be very useful. In the absence of either of these alternatives the static m and n plotting technique (Diaconis and Friedman, 1983) might be of some use. In the absence of these powerful garphical techniques the varying views can be connected by more laborious methods involving isolating points in one view and plotting their positions in the other views.

For the solution projections presented in the previous section the structure (nonnormality) was fairly striking and easily recognized from simple point (scatter) plot representations of the projected densitites. This is not always the case. Human visual perception is not very good at distinguishing varying densities of points. Only the (local) presence or absence of points is easily recognized. Fairly striking density variation is often difficult to see in a scatterplot. For this reason it is often helpful to view a graphical representaton of a smoothed density estimate of the projected solution distributions. Structure easily missed in a scatterplot can be quite evident in such displays. There are several good methods for (smooth) density estimation in two dimensions (Scott, 1985). These density estimates can be represented graphically as contour plots, color relief maps, or isometric projections of the three dimensional surface of the (estimated) density versus the projection coordinates.

As seen in the examples, exploratory projection pursuit solutions can sometimes both discover interesting nonlinear effects and suggest straightforward interpretations for them. More often the interpretation of the discovered structure is elusive and requires a great deal of study and further investigation. In this sense applying projection pursuit to a data set can often raise more questions that it (immediately) answers. This is the primary purpose of an exploratory technique. The discovery of strong (nonlinear) effects will usually cause the analyst to look harder at his data with hopefully a corresponding gain in insight and in understanding.

FORTRAN programs implementing the exploratory projection pursuit procedures described above are available from the author.

Acknowledgments

Helpful discussions with Persi Diaconis, Brad Efron, Iain Johnstone and Art Owen are gratefully acknowledged. I thank John Tukey for reigniting my interest in this subject.

References

Anderson, T.W. (1958). Introduction to Multivariate Analysis. Wiley, New York.

Becker, R.A. and Chambers, J.M. (1984). *S*: An Interactive environment for Data Analysis and Graphics. Wadsworth, Inc., Belmont, CA.

Becker, R.A. and Cleveland, W.S. (1985). Brushing a scatter plot matrix: high-interaction graphical methods for analyzing multidimensional data. AT&T Bell Laboratories Statistical Research Reports No. 7.

Bellman, R.E. (1961). Adaptive control processes, Princeton University Press.

Belsey, D.A., Kuh, E. and Welsh, R.E., (1980). Regression Diagnostics. John Wiley and Sons.

Buja, A. and Asimov, D. (1985). Grand tour methods: an outline. Proceedings: 17th Symposium on the Interface of Computer Science and Statistics.

Chen, Z. and Li, G. (1981). Robust principal components and dispersion matrices via projection pursuit. Dept. of Statistics, Harvard University, report PJH-8.

Devlin, S.J., Gnanadesikan, R., and Kettenring, J.R. (1981). Robust estimation of dispersion matrices and principal components. *J. Amer. Statis. Assoc.* **76**.

Diaconis, P. and Friedman, J.H. (1983). M and N plots. In: Recent Advances in Statistics (Eds: M.H. Rizvi, J.S. Rustagi, and D. Siegmund), Academic Press.

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

Donoho, D.L. and Ramos, E. (1982). PRIMDATA: data sets for use with PRIMH. Dept. of Statistics, Harvard University, Technical Report.

Donoho, A.W., Donoho, D.L. and Gasko, M. (1986). MACSPIN^{TM}: A tool for dynamic display of multivariate data. D^2 Software, Inc., Austin, TX, Report.

Fisherkeller, M.A., Friedman, J.H., and Tukey, J.W. (1975). PRIM-9: An interactive multidimensional data display and analysis system. Stanford Linear Accelerator Center, Report SLAC-PUB-1408.

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

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

Gill, P.E., Murray, W. and Wright, M.H. (1981). Practical Optimization. Academic Press, London.

Harrison, D. and Rubenfeld, D.L. (1978). Hedonic prices and the demand for clean air. *J. of Environ. Econ. and Mngt.* **5**.

Huber, P.J. (1981). Projection pursuit. Dept. of Statistics, Harvard University, Research report number PJH-6.

Huber, P.J. (1985). Projection Pursuit. *Ann. Statis.* **13** 435-475.

Jones, M.C. (1983). The projection pursuit algorithm for exploratory data analysis. University of Bath Ph.D. Dissertation.

Kennedy, W.J. and Gentle, J.E. (1980). Statistical Computing. Marcel Dekker, Inc.

McDonald, J.A. (1984). Interactive graphics for data analysis. Dept. of Statistics, Stanford University, Report number ORION 11.

Mosteller, F. and Tukey, J.W. (1977). Data Analysis and Regression. Addison-Wesley.

Scott, D.W. (1985). Average shifted histograms: effective nonparametric density estimators in several dimensions. *Ann. Statis.* **13** 1024-1040.

Switzer, P. (1970). Numerical classification. *Geostatistics* (ed: D.F. Merrian) Plenum Press, New York 31-43.

Tukey, J.W. and Tukey, P.A. (1985). Computer graphics and exploratory data analysis: an introduction. Proceedings: Annual meeting of National Computer Graphics Assoc., Dallas, TX April, 1985.

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

Table 1

Measurement variables for the states data.

Y_1 : population estimate as of July 1, 1975

Y_2 : average income (1974)

Y_3 : illiteracy rate (1970)

Y_4 : life expectancy (1969-1971)

Y_5 : homicide rate (1976)

Y_6 : high school graduation rate (1970)

Y_7 : average number of days below freezing temperature (1931-1960) in capital or large city

Table 2

Null projection index distributions for the data examples. Projection index (of order J) values obtained by running two–dimensional projection pursuit in the subspace spanned by the largest q principal components, on 20 random samples of size N, drawn from a p–dimensional standard normal.

Table 2a: p=7, q = 4, N=50, J=2

0.022, 0.025, 0.028, 0.028, 0.030, 0.030, 0.030, 0.030, 0.030, 0.033
0.033, 0.038, 0.038, 0.045, 0.058, 0.060, 0.060, 0.065, 0.065, 0.098

Table 2b: p=10, q =10, N=392, J=6

0.043, 0.045, 0.048, 0.050, 0.050, 0.053, 0.053, 0.053, 0.053, 0.055
0.055, 0.055, 0.058, 0.058, 0.060, 0.060, 0.060, 0.063, 0.070, 0.070

Table 2c: p=13, q =13, N=506, J=6

0.043, 0.043, 0.043, 0.045, 0.045, 0.045, 0.045, 0.045, 0.045, 0.048
0.048, 0.048, 0.048, 0.050, 0.053, 0.053, 0.053, 0.053, 0.055, 0.063

Table 3

First projection pursuit solution for the states data.

Solution 1: Projection index = 0.19

α = -0.52, 0.26, 0.08, -0.60, 0.41, 0.16, 0.31

β = -0.05, 0.73, -0.30, 0.10, 0.0, 0.58, 0.16

Table 4

States comprising the smaller cluster associated with lower values of the vertical axis, listed in descending values of the vertical coordinate.

New Mexico	:	-1.32
Texas	:	-1.52
Tennessee	:	-2.01
West Virginia	:	-2.03
Georgia	:	-2.08
Kentucky	:	-2.24
North Carolina	:	-2.29
Alabama	:	-2.72
Arkansas	:	-2.72
South Carolina	:	-2.98
Louisiana	:	-3.15
Mississippi	:	-3.48

Table 5

Measurement variables for automobile data

Y_1	:	gallons per mile (fuel inefficiency)
Y_2	:	number of cyliners in engine
Y_3	:	size of engine (cubic inches)
Y_4	:	engine power (horse power)
Y_5	:	automobile weight
Y_6	:	acceleration (time from 0 to 60 mph)
Y_7	:	model year
Y_8	:	American (0/1)
Y_9	:	European (0/1)
Y_{10}	:	Japanese (0/1)

Table 6

First two projection pursuit solutions for the automobile data.

Solution 1: Projection index = 0.31

α = -0.72, -0.08, 0.56, 0.30, -0.20, 0.10, -0.13, 0.00, 0.00, 0.02

β = 0.00, -0.01, 0.91, 0.14, -0.39, 0.08, -0.01, -0.03, -0.02, - 0.01

Solution 2: Projection index = 0.15

α = -0.10, 0.18, -0.70, -0.15, 0.22, -0.06, -0.17, -0.23, 0.45, -0.32

β = 0.03, 0.09, 0.00, -0.30, 0.53, -0.01, -0.01, 0.34, 0.19, -0.69

Table 7

Neighborhood variables comprising the Boston housing data.

Y_1	:	log (per capita crime rate)
Y_2	:	fraction of land zoned for big lots
Y_3	:	fraction of nonretail business land
Y_4	:	(nitrogen oxide concentration)^2 (pphm)^2
Y_5	:	(average number of rooms)^2
Y_6	:	fraction of owner-occupied units built before 1940
Y_7	:	log (weighted distances to five employment centers)
Y_8	:	log (index of access to radial highways)
Y_9	:	full-value property-tax rate
Y_{10}	:	pupil - teacher ratio
Y_{11}	:	log (0.4 - (fraction black population - 0.63)^2)
Y_{12}	:	log (fraction of lower status population)
Y_{13}	:	log (median value of owner-occupied homes)

Table 8

First five projection purusit solution planes for Boston neighborhood data.

Solution 1: Projection index = 0.69

α = 0.13, -0.41, -0.50, -0.24, -0.04, -0.02, 0.20, 0.26, 0.24, -0.04, -0.50, 0.14, 0.19

β = 0.0, 0.996, 0.05, -0.02, 0.0, 0.02, 0.02, 0.04, -0.04, -0.02, 0.0, 0.01, 0.01

Solution 2: Projection index = 0.51

α = 0.12, 0.23, -0.76, -0.06, 0.05, 0.01, 0.04, 0.09, 0.37, 0.41, -0.09, 0.10, 0.12

β = 0.17, 0.08, 0.0, 0.16, 0.03, -0.13, -0.08, 0.40, 0.83, 0.24, 0.06, -0.01, -0.11

Solution 3: Projection index = 0.40

α = -0.21, -0.23, -0.25, 0.16, -0.19, -0.09, 0.38, -0.31, 0.70, 0.0, 0.04, 0.08, -0.15

β = 0.17, -0.44, -0.79, 0.07, -0.03, -0.02, 0.0, 0.10, 0.29, -0.23, 0.03, -0.02, 0.03

Solution 4: Projection index = 0.25

α = -0.41, 0.18, -0.23, 0.22, -0.15, 0.66, 0.18, -0.01, 0.30, 0.10, 0.08, 0.10, 0.30

β = 0.02, -0.09, 0.13, 0.07, -0.45, 0.0, 0.22, -0.07, 0.05, 0.07, 0.09, 0.02, 0.84

Solution 5: Projection index = 0.34

α = -0.03, -0.12, -0.60, -0.01, -0.08, 0.10, -0.42, 0.05, 0.06, 0.57, -0.05, -0.29, -0.12

β = 0.25, 0.06, -0.55, 0.71, 0.0, -0.02, 0.17, -0.07, -0.29, 0.0, -0.01, 0.0, 0.02

FIGURE 1

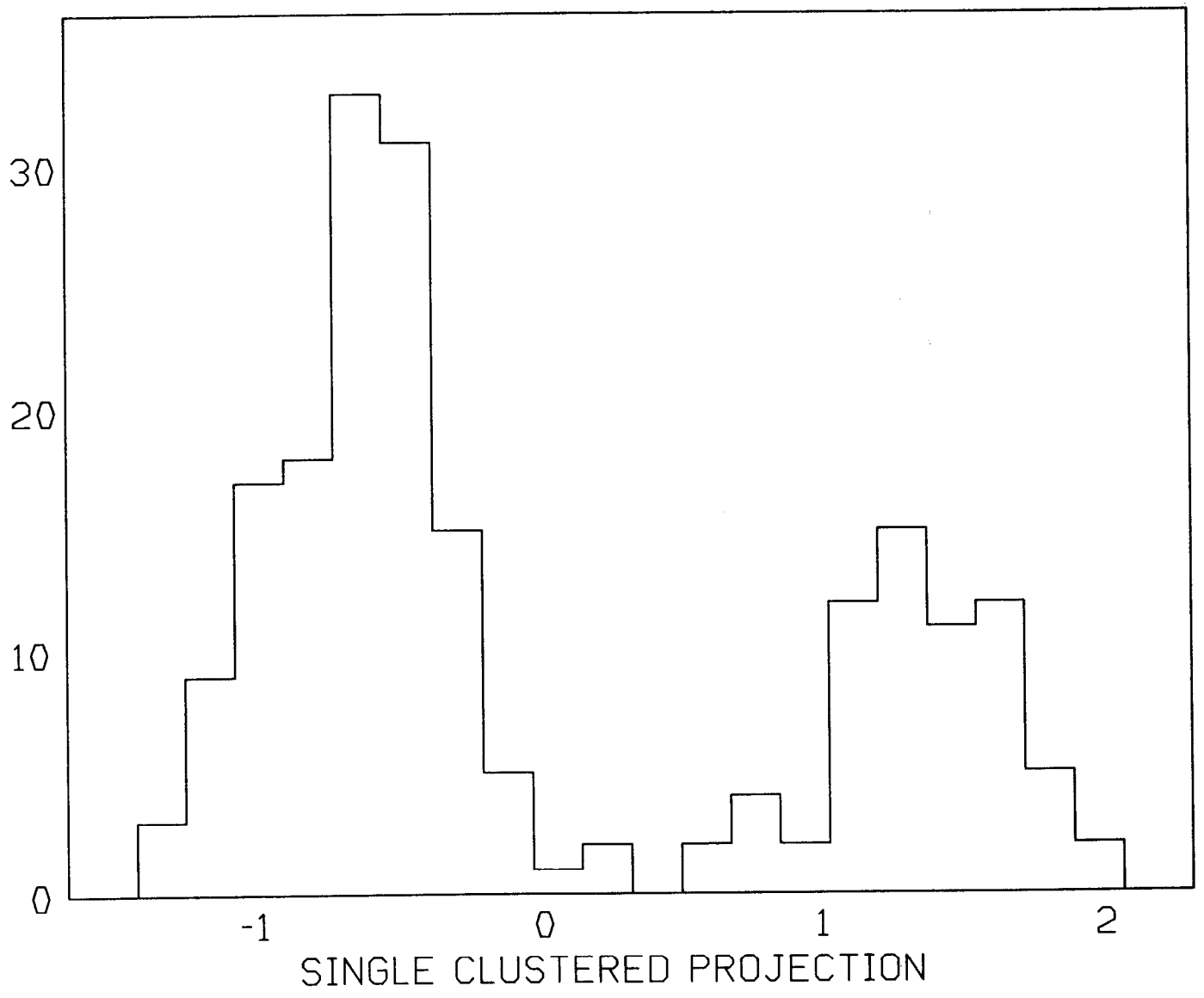


FIGURE 2A

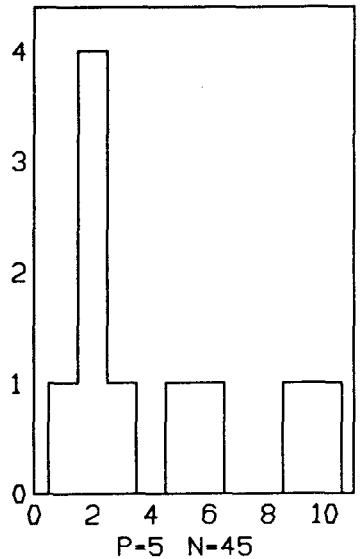


FIGURE 3A

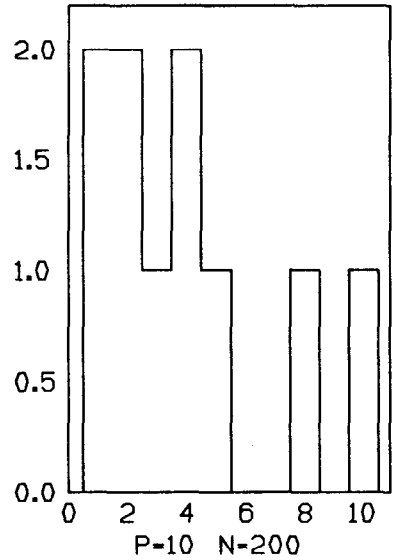


FIGURE 4A

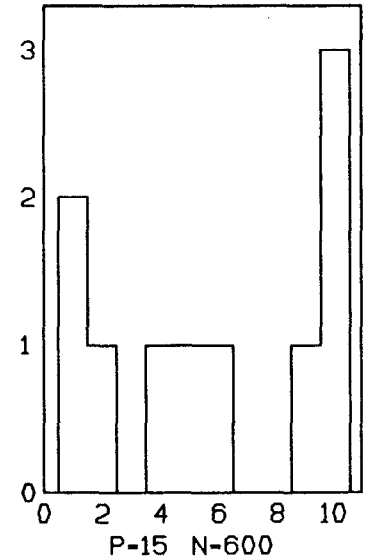


FIGURE 2B

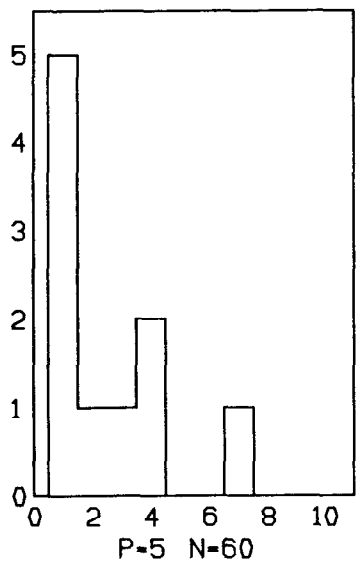


FIGURE 3B

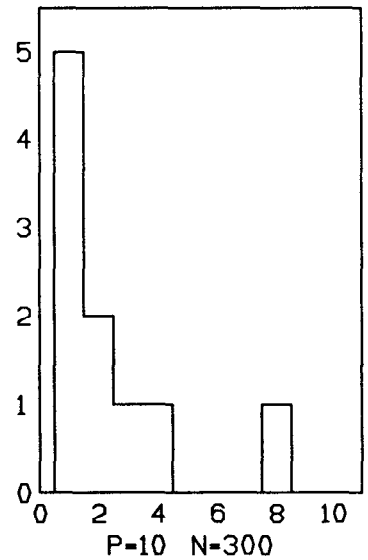


FIGURE 4B

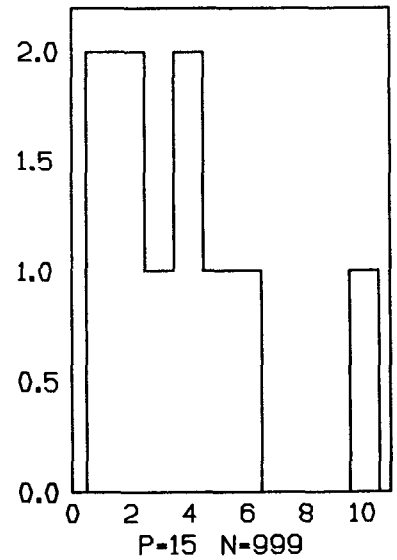


FIGURE 5A

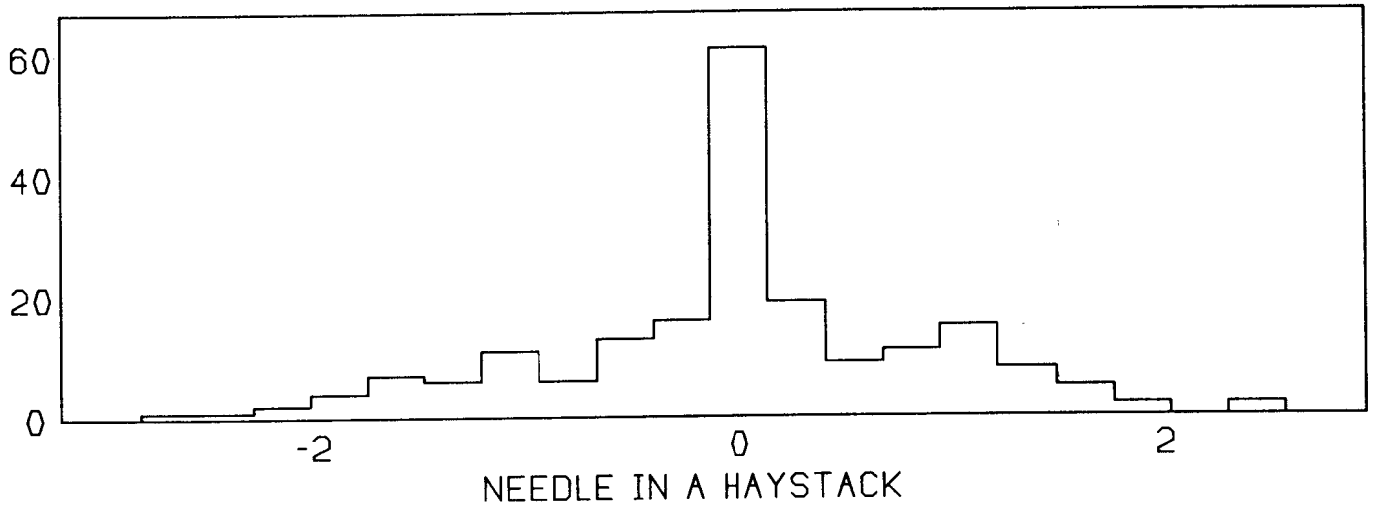


FIGURE 5B

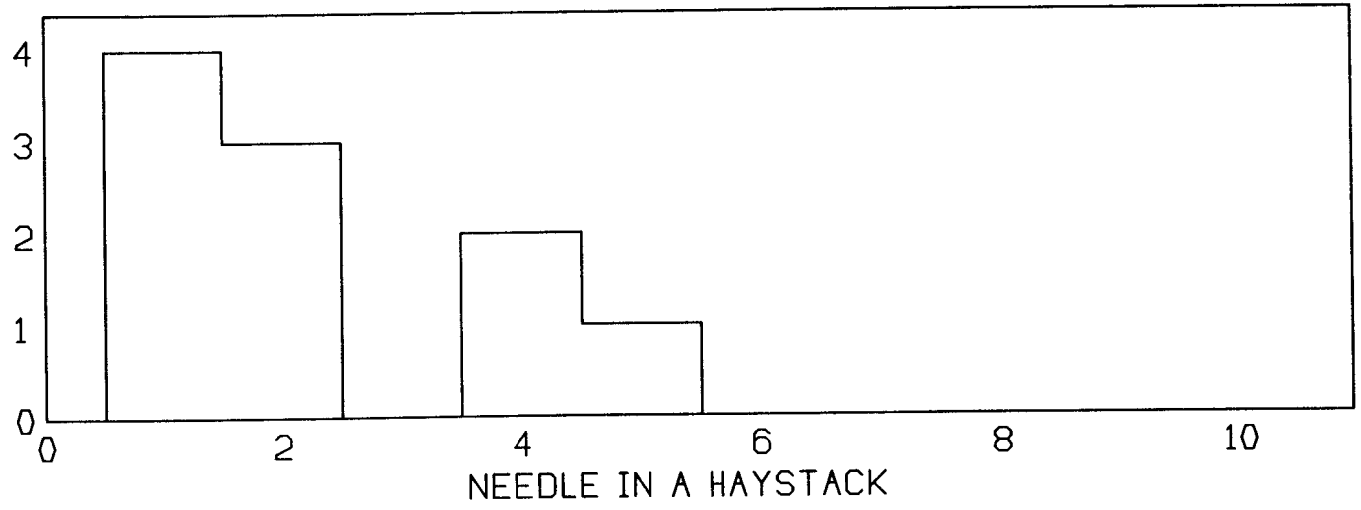


FIGURE 6

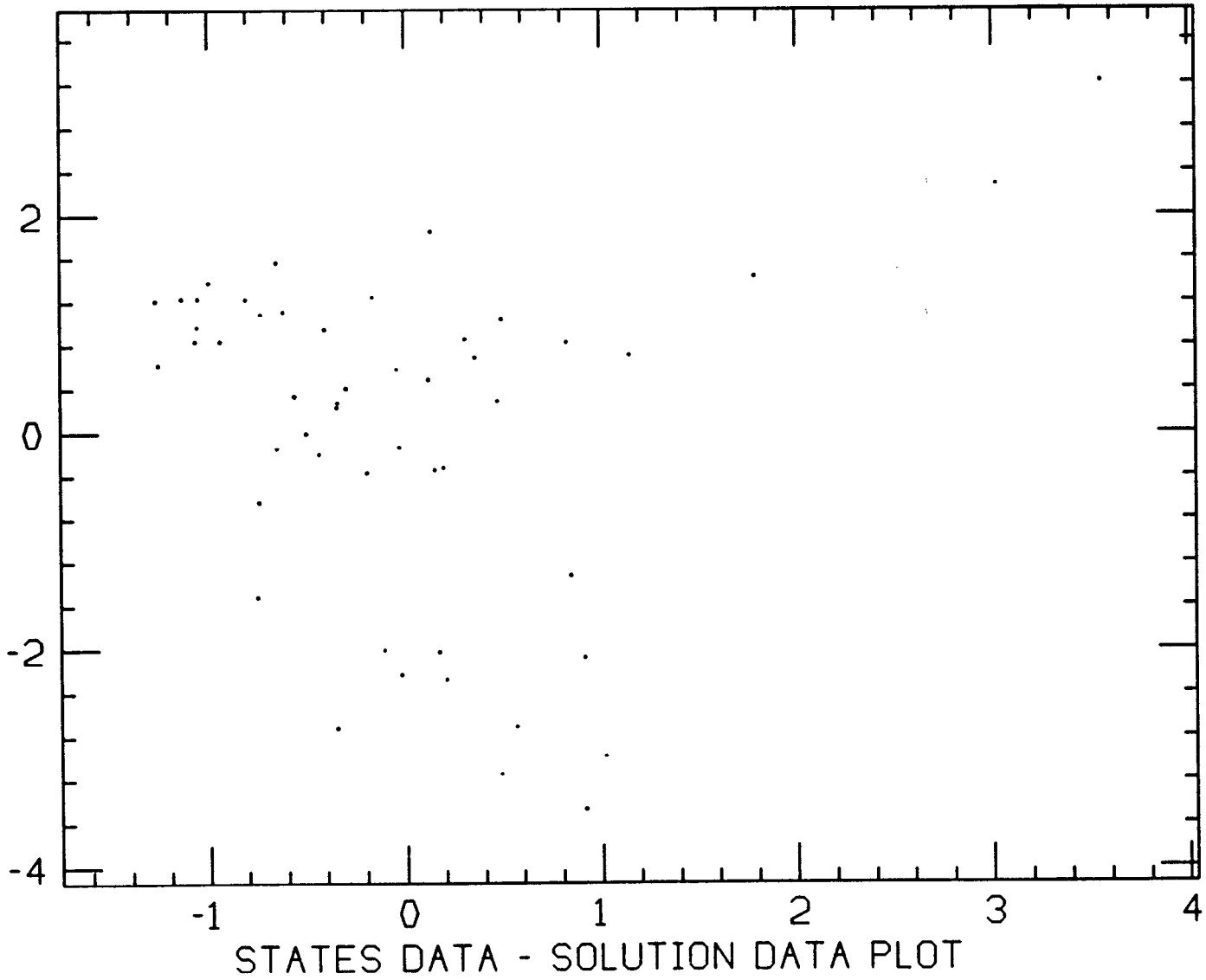
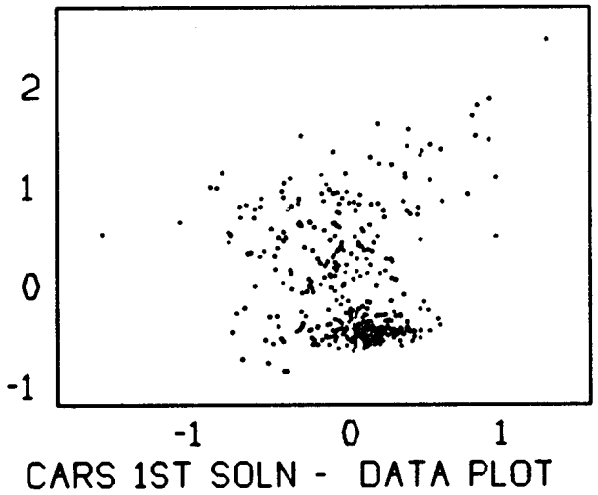
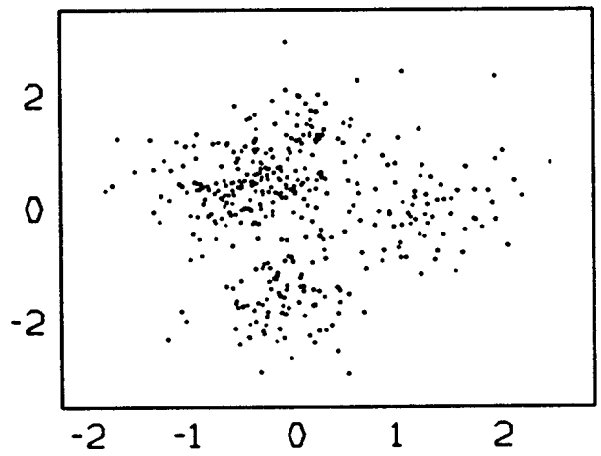


FIGURE 7A



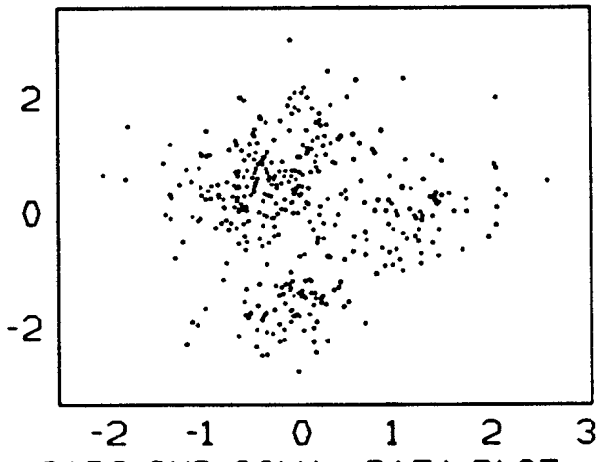
CARS 1ST SOLN - DATA PLOT

FIGURE 7B



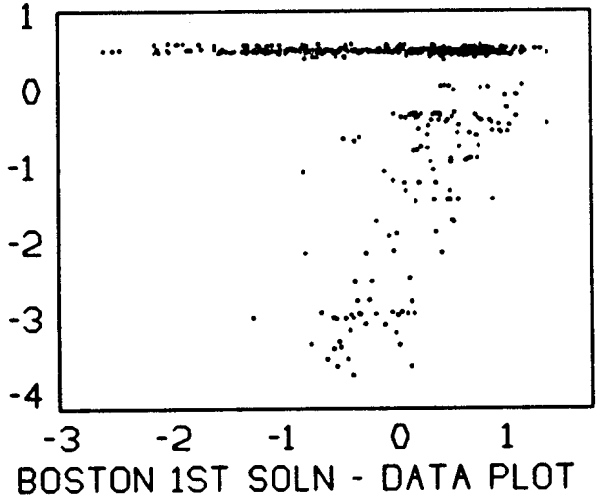
CARS 2ND SOLN - ADJUSTED PLOT

FIGURE 7C



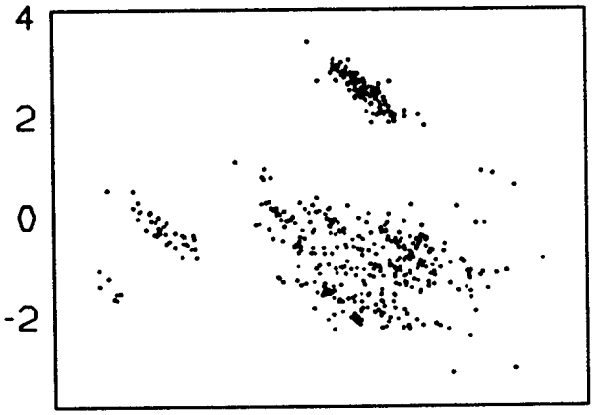
CARS 2ND SOLN - DATA PLOT

FIGURE 8



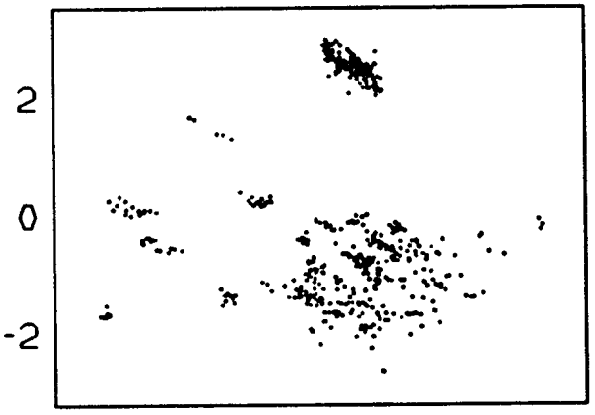
BOSTON 1ST SOLN - DATA PLOT

FIGURE 9A



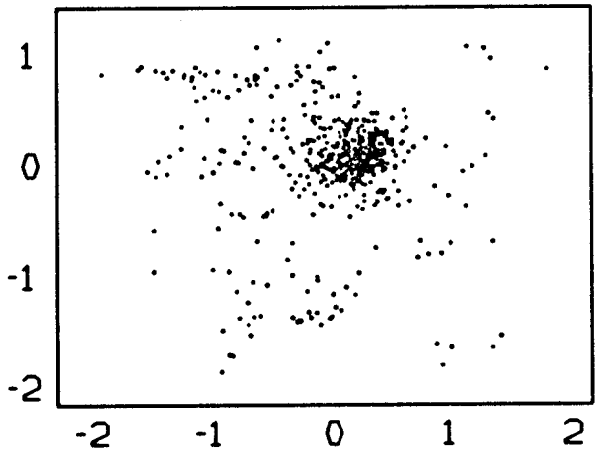
BOSTON 2ND SOLN - ADJUSTED PLOT

FIGURE 9B



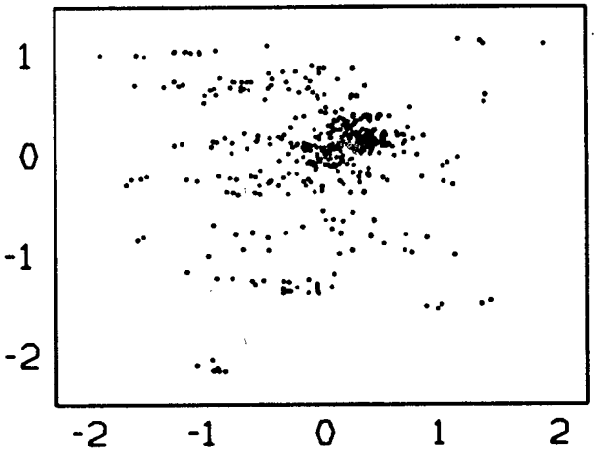
BOSTON 2ND SOLN - DATA PLOT

FIGURE 10A



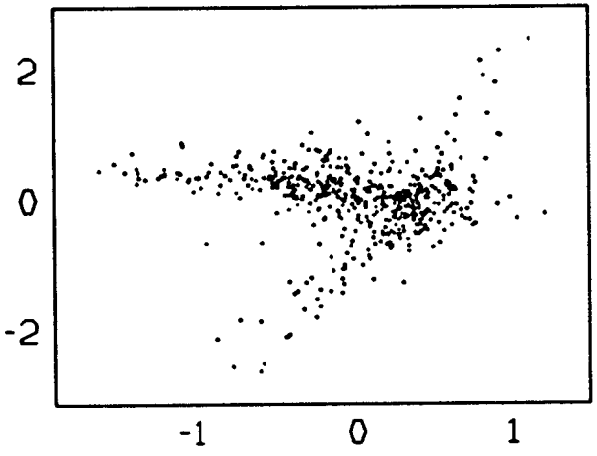
BOSTON 3RD SOLN - ADJUSTED PLOT

FIGURE 10B



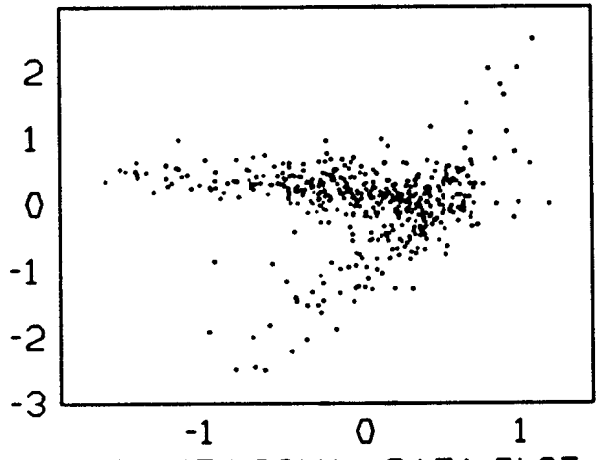
BOSTON 3RD SOLN - DATA PLOT

FIGURE 11A



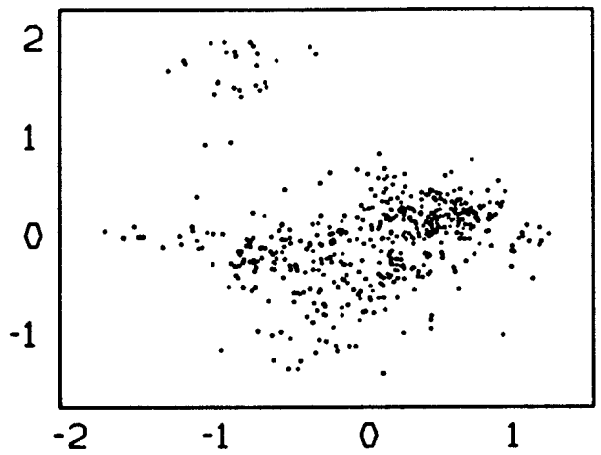
BOSTON 4TH SOLN - ADJUSTED PLOT

FIGURE 11B



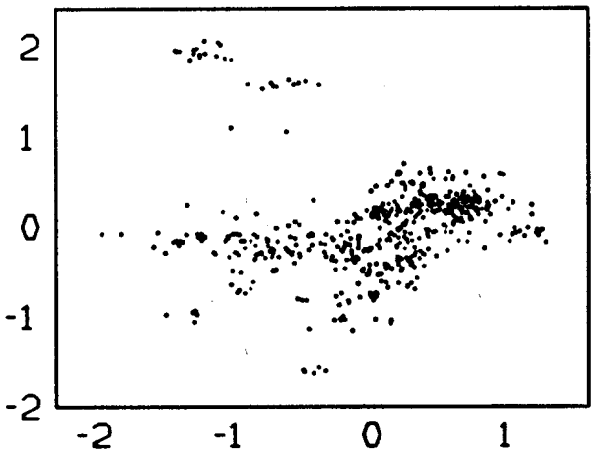
BOSTON 4TH SOLN - DATA PLOT

FIGURE 12A



BOSTON 5TH SOLN - ADJUSTED PLOT

FIGURE 12B



BOSTON 5TH SOLN - DATA PLOT