SLAC PUB-2768
June 1981
(M)

PROJECTION PURSUIT METHODS FOR DATA ANALYSIS*

Jerome H. Friedman
Stanford Linear Accelerator Center
Stanford, California

Werner Stuetzle
Department of Statistics
Stanford University
and
Stanford Linear Accelerator Center
Stanford, California

I. INTRODUCTION

Multivariate analysis can be thought of as a methodology for detection, description and validation of structure in p-dimensional ($p > 1$) point clouds. Classical multivariate analysis relies on the assumption that the observations forming the point cloud(s) have a Gaussian distribution. All information about structure is then contained in the means and covariance matrices, and the well-known apparatus for estimation and inference in parametric families can be brought to bear. The uncomfortable ingredient in this approach is the Gaussianity assumption. The data may be Gaussian with occasional outliers or even the bulk of the data simply might not conform to a Gaussian distribution. The first case is the subject of robust statistics and is not treated here. We discuss methods that do not involve any distributional assumptions. In this case,

*Work supported by the Department of Energy under contract DE-AC03-76SF00515
(Presented at Army Research Office Conference on Modern Data Analysis, North Carolina, June 1980)

structure cannot be perceived by looking at a set of estimated parameters. An obvious remedy is to look at the data themselves, at the p-dimensional point cloud(s), and to base the description of structure on those views. As perception in more than three dimensions is difficult, the dimensionality of the data first has to be reduced, most simply by projection. Projection of the data generally implies loss of information. As a consequence, multivariate structure does not usually show up in all projections, and no single projection might contain all the information. These points are further illustrated in Chapter 2. It is therefore important to judiciously choose the set of projections on which the model of the structure is to be based. This is the goal of projection pursuit procedures. A paradigm for multivariate analysis based on these ideas is presented in Chapter 3.

By design, projection pursuit methods are ideally suited for implementation or interactive computer graphics systems. The potential of interaction between user and algorithm was convincingly demonstrated in the PRIM-9 system for detection of hypersurfaces and clustering (see Fisherkeller et al [1974]); this system is discussed in Chapter 4. Procedures for multiple regression and multivariate density estimation based on projection pursuit are outlined in Chapters 5 and 6. Common properties of all projection pursuit procedures are discussed in Chapter 7.

2. DETECTION AND DESCRIPTION OF STRUCTURE WITH PROJECTIONS

Our goal is to detect and describe multivariate structure using projections of the data. However, structure, if present,

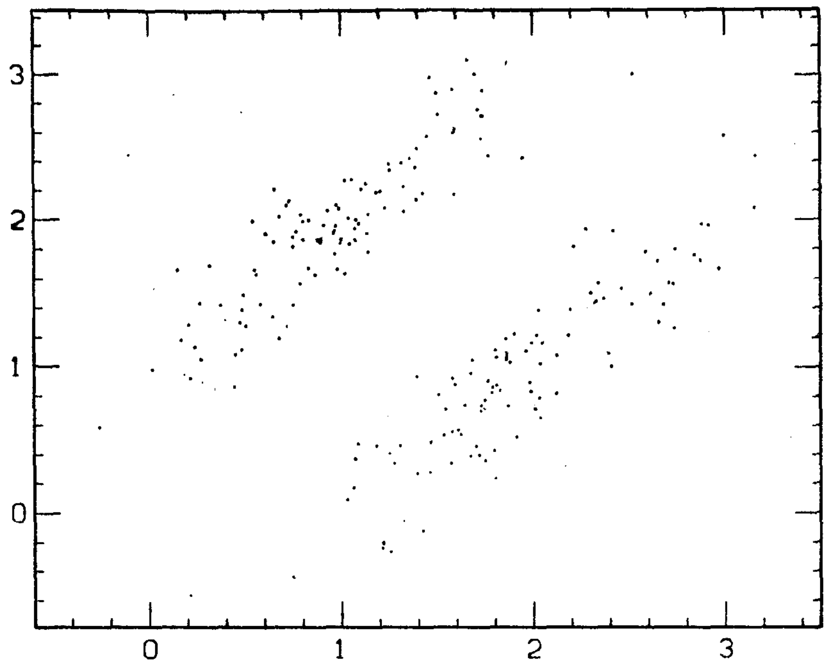


Fig. 1 Structured point cloud in two dimensions

may not be apparent in all projections. This is illustrated by the following examples. Figure 1 shows a point-sample drawn from a bivariate distribution. The apparent structure of the point cloud (separation into two clusters) would be revealed by projection onto the subspace spanned by the vector (1, -1), whereas no structure would be apparent in a projection on the subspace spanned by the vector (1, 1).

The data for Figure 2 are generated from the regression model $Y = X_1 + X_2 + \epsilon$ with (X_1, X_2) uniformly distributed in $[-1,1] \times [-1,1]$ and $\epsilon \sim N(0,0.01)$. Figure 2a shows a projection on the two-dimensional subspace spanned by Y and the linear combination $Z = X_1 + X_2$. This projection clearly shows the association between the predictors X_1 and X_2 and the response Y. A similar plot with $Z = X_1 - X_2$, Figure 2b, is clearly less structured.

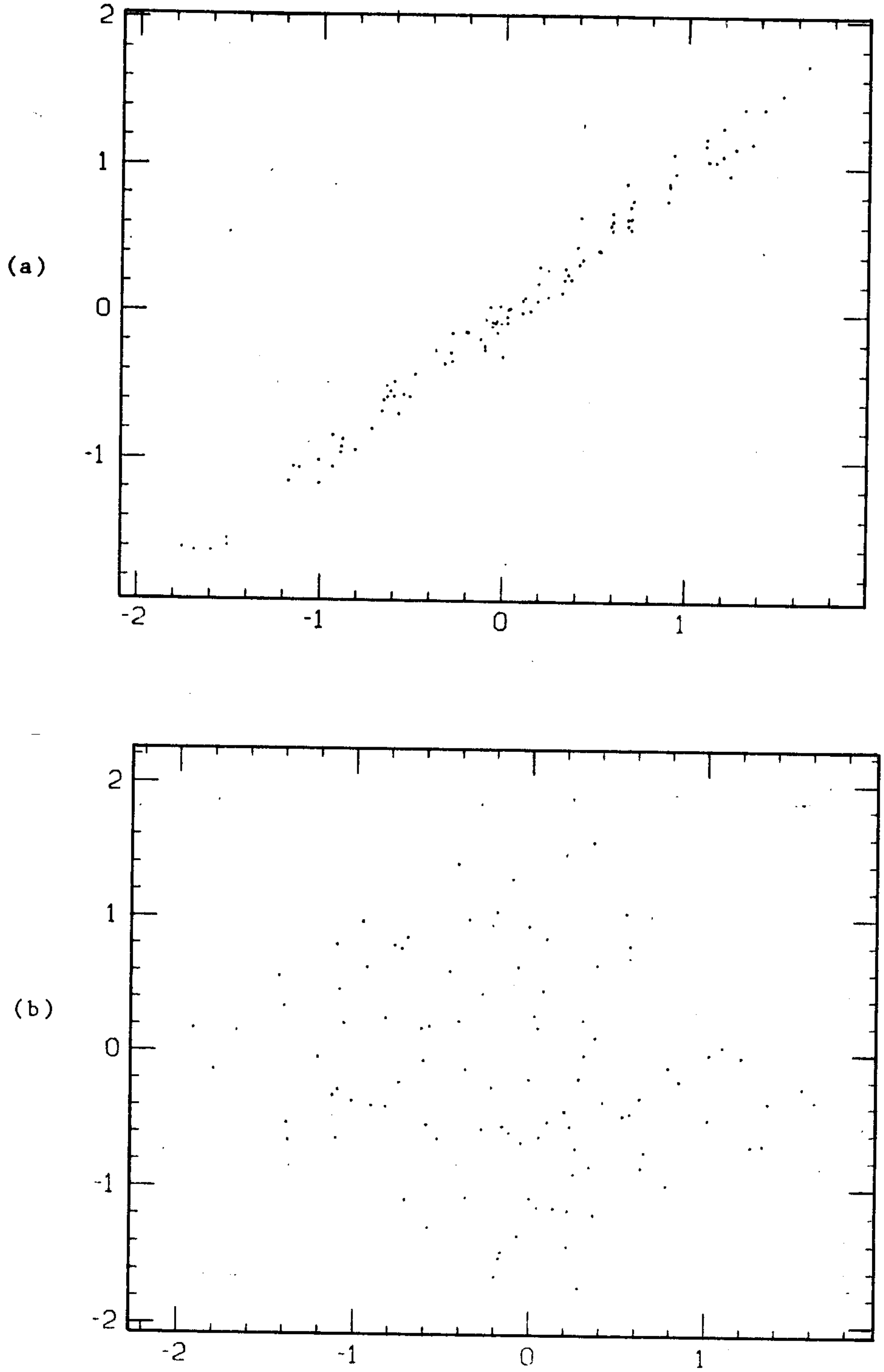


Fig. 2. (a) Projection of data from model $Y=X_1+X_2+\epsilon$ on plane spanned by Y and $Z=X_1+X_2$. (Y is plotted on the vertical axis). (b) Projection of data from model $Y=X_1+X_2+\epsilon$ on plane spanned by Y and $Z=X_1-X_2$.

These examples show that it is important to search for structured projections. This process is called projection pursuit.

It is easy to envision situations where not all the information about the structure is contained in a single projection. Consider the regression example above but with $Y = X_1 . X_2 + \epsilon$. Figures 3a and 3b show two projections with $Z_a = X_1 - X_2$ and $Z_b = X_1 + X_2$. To understand the pictures, note that the simple coordinate transformation $Z_a = X_1 + X_2$, $Z_b = X_1 - X_2$ allows one to express the response as $Y = .25\ (Z_a^2 - Z_b^2)$. It is also interesting to notice that the quadratic dependence on Z_a is washed out due to variability caused by the dependence on Z_b, and vice versa. This suggests that once a structured projection has been found, the structure should be removed so that one obtains a clearer view of what has not yet been uncovered.

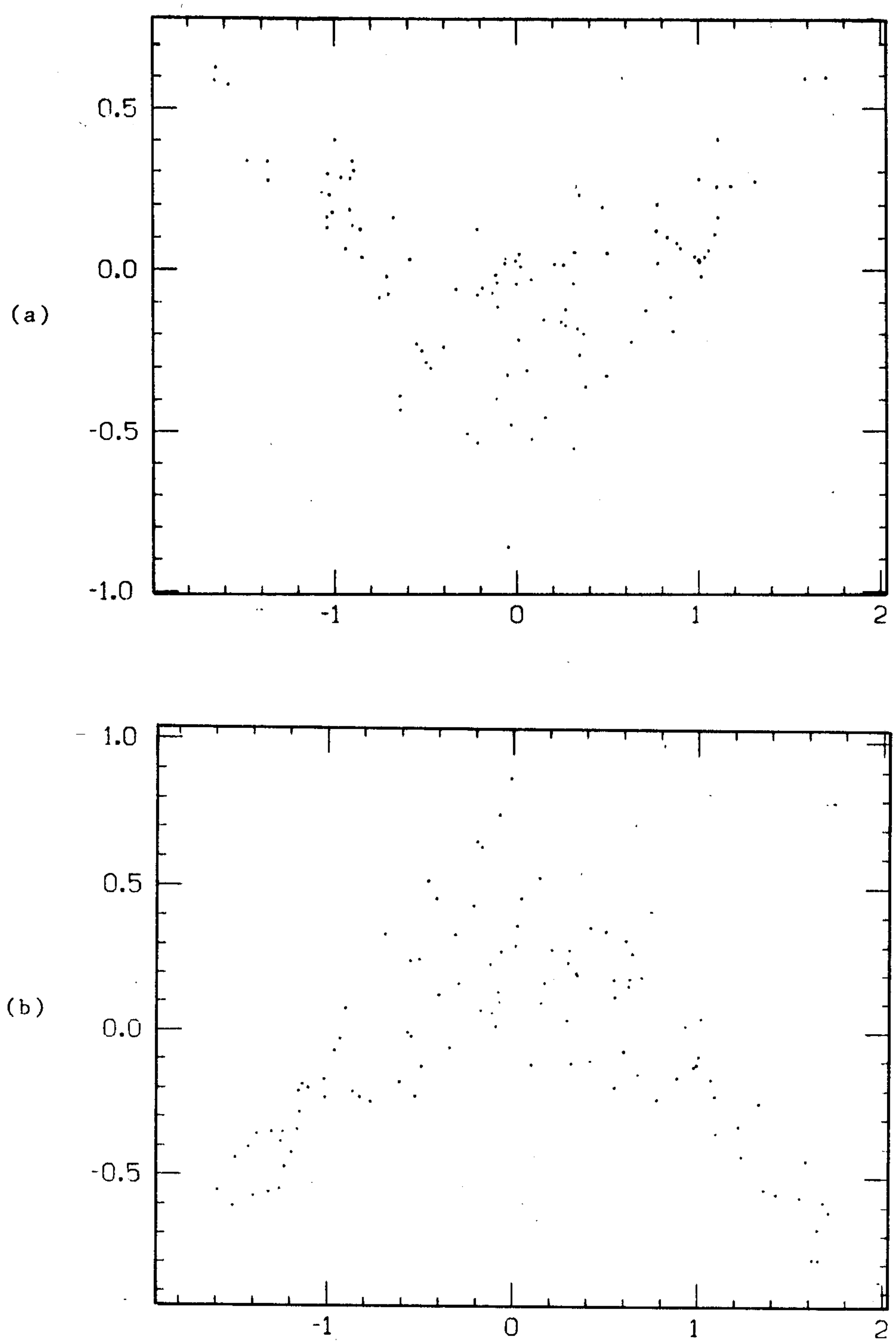


Fig. 3. (a) Projection of data from model $Y=X_1 . X_2+\epsilon$ on plane spanned by Y and $Z_a=X_1+X_2$. (b) Projection of data from model $Y=X_1 . X_2+\epsilon$ on plane spanned by Y and $Z_b=X_1-X_2$.

3. A PROJECTION PURSUIT PARADIGM

The discussion in the previous section motivates the following schema for a class of procedures modeling structure in multivariate data:

```
(i) Choose an initial model.
Repeat
    (ii) Find a projection that shows deviation of the
         data from the current model, indicating pre-
         viously undetected structure (Projection Pursuit).
   (iii) Change the model to incorporate the structure
         found in (ii) (Model Update).
Until the current model agrees with the data in all pro-
jections.
```

Such projection pursuit procedures can be implemented in batch mode. In this case, a figure of merit must be defined, which measures the amount of deviation between model and data revealed in a projection. This figure of merit usually is optimized by numerical search, although in some simple cases optimization can be done analytically. If the optimum figure of merit is less than a threshold, data and model are said to agree. Batch implementations of projection pursuit regression and density estimation are described in Sections 5 and 6.

By construction, projection pursuit procedures are ideally suited for implementation on interactive computer graphics systems. Interaction between program and user can help in

- search for interesting projections
- specification of model update
- termination
- interpretation of structure.

Although projection pursuit procedures are useful in batch mode, their full power comes to bear in an interactive environment.

4. THE PRIM-9 SYSTEM

PRIM-9 (Fisherkeller et al, [1974]) is a system for visual inspection of up to nine-dimensional data, mainly intended for detecting clusters and hypersurfaces. It was implemented on an interactive computer graphics system which allows the modification of pictures in real time and thus makes it possible to generate movie-like effects. Its basic set of operations consists of

Projection: The observations can be projected on a subspace spanned by any pair of the coordinates; the projection is shown on a CRT screen.

Rotation: A subspace spanned by any two of the coordinates can be rotated. If the projection subspace and the rotation subspace share a common coordinate, the rotational motion causes the user to perceive a spatial picture of the data as projected on the three-dimensional subspace defined by the coordinates involved. When the user terminates rotation in a particular plane, the old coordinates in that plane are replaced by the current (rotated) coordinates. This makes it possible to look at completely arbitrary projections of the data, not necessarily tied to the original coordinates.

Masking: Subregions of the p-dimensional observation space can be specified, and only points inside the subregion are displayed. Under rotation, points will enter and leave the masked region.

Isolation: Points that are masked out (i.e., not visible) can be removed, thus splitting the data into two subsets.

The first two operations, projection and rotation, allow the user to perform what one might call "manual projection pursuit". Isolation, the splitting of the data set into subsets, provides a rudimentary form of structure removal. When clustering is detected, the clusters can be separated and each of them examined individually. This process can be iterated.

Although several have been implemented (Stuetzle & Thoma [1978], Donoho et al [1981]), systems like PRIM-9 have not yet found widespread use. The main reason has been the price of the necessary computing equipment. The processing power needed to compute rotations at a reasonable update rate is quite high (on the order of 60000 multiplications per second for 1000 observations and 10 updates of the picture per second). Another major cost has been the graphics device, which must have a sufficiently high bandwidth (typically a megabaud). The situation, however, is rapidly changing. New 16-bit microprocessors provide a speed close to that required for an interactive use of projection pursuit procedures. The price of graphics systems, especially raster scan devices, is falling dramatically. The graphics system at SLAC used for the implementation of PRIM-9 cost $175,000 in 1967. Today the price of a comparable system is $15,000.

5. PROJECTION PURSUIT REGRESSION (PPR)

The goal of regression analysis is to find and describe the association between a response variable Y and predictor variables $X_1 \ldots X_p$, using a sample $\{(y_i, \underline{x}_i)\}_{i=1}^{n}$. PPR attempts to construct a model for this association (or, in more classical terms, to estimate $E(Y|\underline{x})$) from the information contained in projections of the data on two-dimensional subspaces spanned by Y and a linear combination $Z = \underline{\alpha} \cdot \underline{X}$. The algorithm exactly follows our projection pursuit paradigm:

```
(i) Choose an initial model, for example $m_0(\underline{X}) = const.$
Repeat
  (ii) Find a projection that shows deviation of the data
       from the model, i.e., find a direction such that
       the current residuals, $r_i=y_i-m(\underline{x}_i)$, show a dependence on $Z=\underline{\alpha} \cdot \underline{X}$
  (iii) Describe this dependence by a smooth function $s(Z)$.
       Update the model:
         $m(\underline{X}) \leftarrow m(\underline{X}) + s(\underline{\alpha} \cdot \underline{X})$
Until data and model agree in all projections.
```

The model after M iterations has the form

$$m(\underline{X}) = m_0(\underline{X}) + \sum_{m=1}^{M} s_m(\underline{\alpha}_m \cdot \underline{X}). \qquad (1)$$

PPR allows the modeling of smooth but otherwise completely general regression surfaces. So far, a batch version has been implemented. Such an implementation requires the specification of a figure of merit for projections and a method for summarizing a smooth dependence ("smoother"). Smoothing is generally accomplished by local averaging; the value of the smooth s at a particular point z is obtained by averaging the current residuals r_i for those observations with values of z_i close to z. The size of the neighborhoods within which averaging takes place is called the bandwidth of the smoother. A smoother suitable for use with PPR and guidelines for choosing the bandwidth are described and discussed in Friedman and Stuetzle (1981).

A choice for the figure of merit is suggested by figures 2a and 2b. The (inverse) figure of merit is taken to be the residual sum of squares around the smooth of the current residuals versus $\underline{\alpha} \cdot \underline{X}$. It is small in Figure 2a, where the smooth could closely follow the observations, and large in Figure 2b, where the smooth would be roughly constant. This definition of the figure of merit implies that in each iteration the model is updated along the direction for which the update yields the biggest reduction in residual sum of squares.

As with any stepwise procedure, one needs a criterion for stopping the iteration. Stopping too soon can increase the bias of the estimate, while not stopping soon enough can unduly increase its variance. "Optimal" termination of stepwise procedures has been studies (see Stone, [1981]); these methods can be applied here. In practice, the iteration is usually terminated subjectively, based on differences between successive values of the residual sum of squares. In addition, graphical inspection of $s_m(\underline{\alpha}_m \cdot \underline{X})$ can be used to judge whether the corresponding term should be included in the model. If the graph of s_m shows a noisy pattern with no systematic tendency, then its inclusion can only increase the variability of the estimate. On the other hand, a definite dependence indicates that s_m deals with an inadequacy of the present model.

The following example illustrates the operation of PPR. A sample of 200 observations was generated according to the model

$$Y = 10 \sin(\pi X_1 X_2) + 20(X_3 - 0.5)^2 + 10X_4 + 5X_5 + 0X_6 + \epsilon$$

with $(X_1, \ldots, X_6)$ uniformly distributed in $[-1,1]^6$ and $\epsilon \sim N(0,1)$. Figure 4a shows Y plotted against the best single predictor,

X_4, and the corresponding smooth. (The response Y is plotted on the vertical axis, X_4 on the horizontal axis. The "+" symbols represent data points, numbers indicated more than 1 data point. The smooth is represented by the "*" symbols.) Figure 4b shows Y plotted against the linear combination $\underline{\alpha}_1 . \underline{X}$ found in the first iteration with $\underline{\alpha}_1$ = (0.41, 0.51, -0.04, 0.69, 0.31, 0.0). The association is seen to be approximately linear. The model after the first iteration thus is a plane which, in this case, closely coincides with the least squares plane through the data. Figure 4c shows the residuals from this model plotted against the second linear combination $\underline{\alpha}_2 \cdot \underline{X}$ found by the algorithm, with $\underline{\alpha}_2$ = (-0.14, 0.0, 0.99, 0.04, 0.0, -0.03). This iteration is seen to incorporate the quadratic dependence of the response on X_3 into the model. Figure 4d shows the residuals after two iterations plotted against the third linear combination with $\underline{\alpha}_3$ = (.0.70, 0.72, 0.01, 0.03, 0.02, 0.00). Figure 4e shows the residuals after three iterations plotted against the fourth linear combination, with $\underline{\alpha}_4$ = (0.80, -0.59, -0.10, 0.04, 0.01, 0.0). The last two iterations are seen to model the interaction term $\sin(\pi X_1 X_2)$. A further iteration failed to substantially improve the model.

For a more complete discussion of PPR and additional examples, see Friedman and Stuetzle [1981].

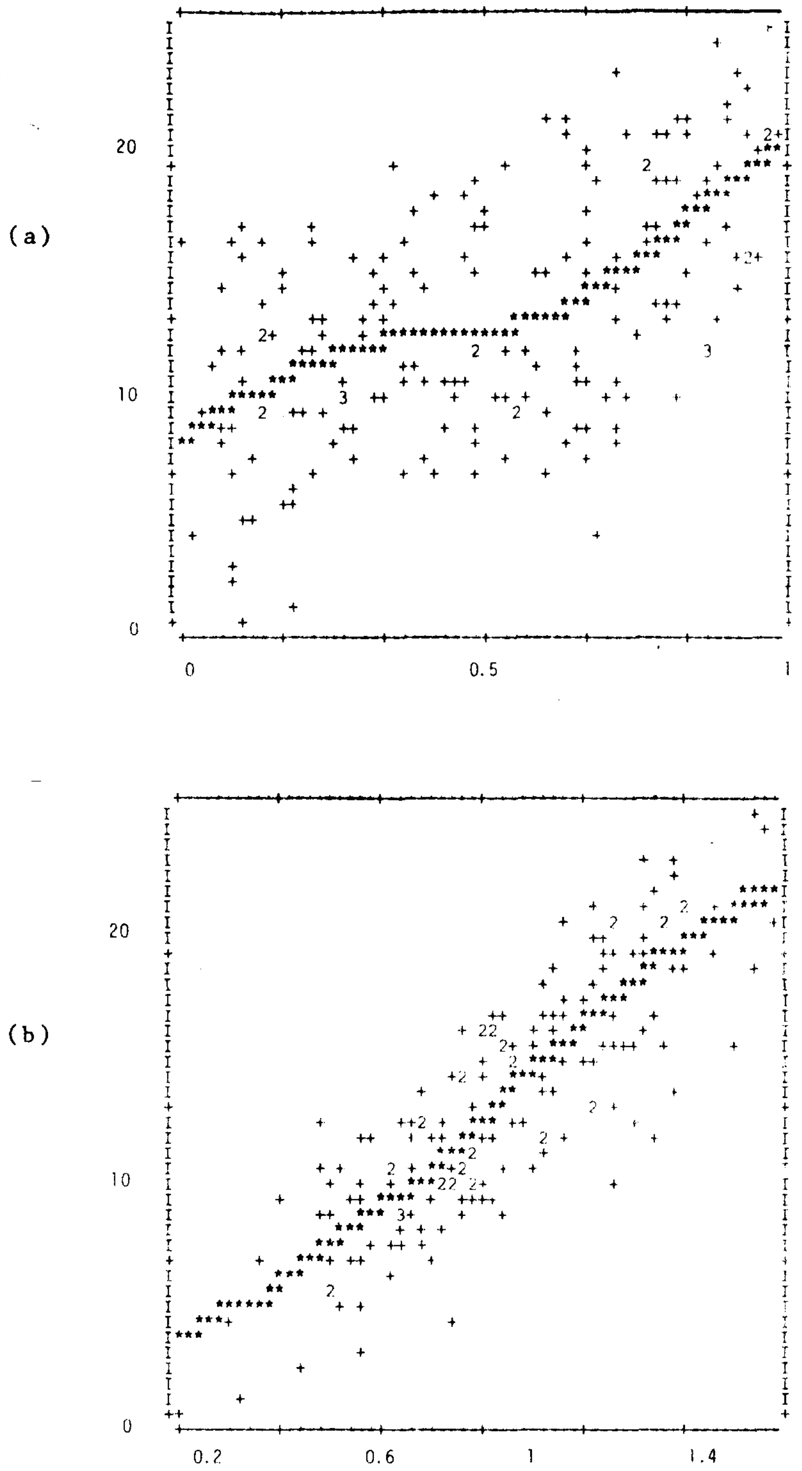
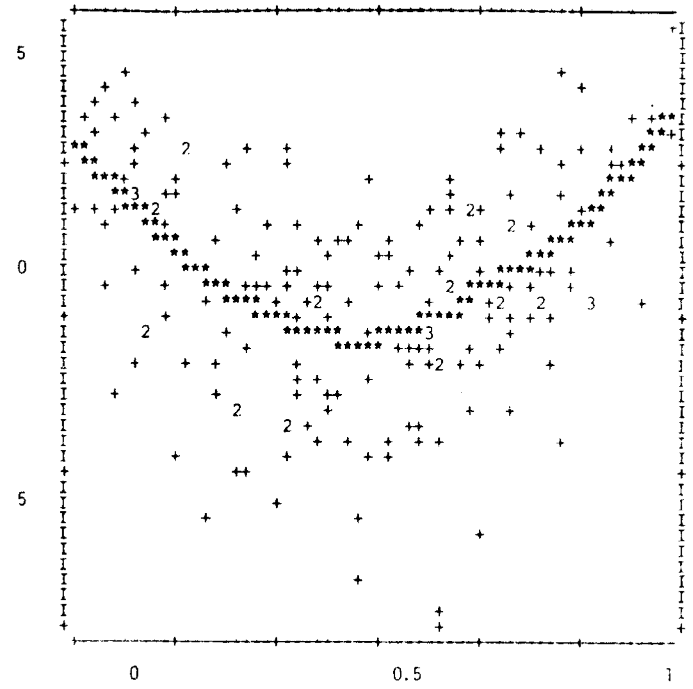


Fig. 4. (a) Scatterplot of Y vs. X_4. (b) Scatterplot of Y vs. $\underline{\alpha}_1\cdot\underline{X}$. (c) Scatterplot of Y - $s_1(\underline{\alpha}_1\cdot\underline{X})$ vs. $\underline{\alpha}_2\cdot\underline{X}$. (d) Scatterplot of Y - $s_1(\underline{\alpha}_1\cdot\underline{X})$ - $s_2(\underline{\alpha}_2\cdot\underline{X})$ vs. $\underline{\alpha}_3\cdot\underline{X}$. (e) Scatterplot of Y - $s_1(\underline{\alpha}_1\cdot\underline{X})$ - $s_2(\underline{\alpha}_2\cdot\underline{X})$ - $s_3(\underline{\alpha}_3\cdot\underline{X})$ vs. $\underline{\alpha}_4\cdot\underline{X}$.

(c)



(d)

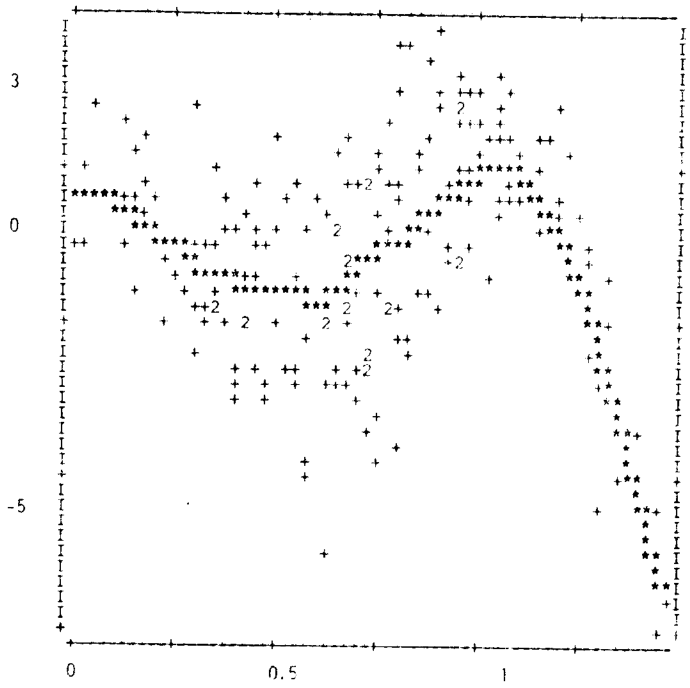


Fig. 4c,d

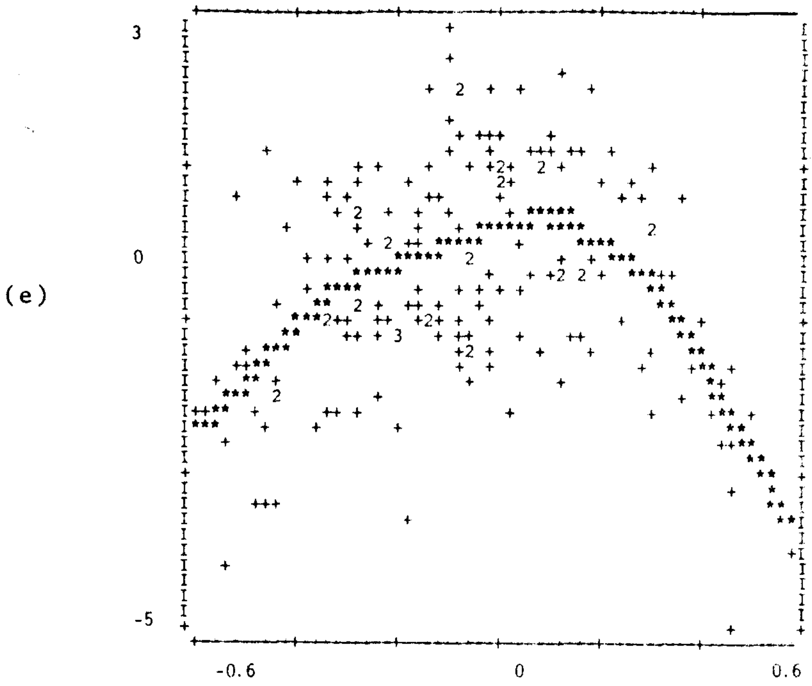


Fig. 4e

6. PROJECTION PURSUIT DENSITY ESTIMATION (PPDE)

The goal of density estimation is to estimate the multivariate distribution of a random vector $\underline{X}$ on the basis of an i.i.d. sample $\underline{x}_1 \ldots \underline{x}_n$. Our procedure again follows the projection pursuit paradigm:

(i) Choose an initial model for the density, for example, m_0 = multivariate normal with sample mean and covariance matrix.

Repeat

(ii) Find a projection that shows deviation of the data from the model; i.e., find a direction such that $m(\underline{\alpha} \cdot \underline{X})$, the model marginal along $\underline{\alpha}$, differs from $p(\underline{\alpha} \cdot \underline{X})$, the (estimated) data marginal along $\underline{\alpha}$.

(iii) Define an "augmenting function" $f(\underline{\alpha} \cdot \underline{X})$ as the quotient of data and model marginals

```
            $f(\alpha \,.\, X) = \frac{p(\underline{\alpha} \,.\, \underline{X})}{m(\underline{\alpha} \,.\, \underline{X})}$
        Update the model so that it and the data agree in
        the marginal along $\underline{\alpha}$:
            $m(\underline{X}) \leftarrow m(\underline{X}) \cdot f(\underline{\alpha} \,.\, \underline{X})$.
Until data and model agree in all projections.
```

The model after M steps of the iteration is of the form

$$m(\underline{X}) = m_0(\underline{X}) \cdot \prod_{m=1}^{M} f_m(\underline{\alpha}_m \,.\, \underline{X}). \tag{2}$$

In step (iii) of the algorithm, the marginal of the data along $\underline{\alpha}$ must be estimated and the marginal of the current model must be computed. The data marginal presents no problem. It can be estimated by projecting the data onto $\underline{\alpha}$ and using a one-dimensional kernel or near neighbor estimate. The analytic computation of the model marginal can be very difficult because it requires a (p-1)-dimensional integration. We perform the integration by Monte Carlo, generating a sample from the model and proceeding as in the estimation of the data marginal.

As in the case of PPR, only a batch version of PPDE has so far been implemented. At each iteration, the direction $\underline{\alpha}$ is chosen such that the update of the current model yields the largest improvement in goodness-of-fit as measured by the likelihood of the sample. Termination rules are analogous to those used in PPR.

The following example illustrates the operation of PPDE. The data for the example are the concentration levels of four hormones in blood measurements of 256 children. The purpose of applying PPDE to these data is to determine if a Gaussian distribution represents a reasonable approximation to the data density. Figures 5a-5d compare the experimental data to a Monte Carlo sample drawn from a Gaussian density with the sample mean and covariance, as projected onto each of the measurement coordinates. The histogram of the experimental data is drawn with solid lines; the histogram of the Monte Carlo data is indicated by "*" symbols. Inspection of these projections indicates that although there are possibly some discrepancies, a Gaussian density might be a reasonable approximation to the data.

Figures 6a-6e show results for three iterations of PPDE. The solution direction $\underline{\alpha}_1$ associated with the first iteration is mainly a combination of the second and third coordinate measurements. The data distribution (Figure 6a) is seen to be somewhat skew and more peaked than the corresponding Gaussian. The discrepancy between the data and the Gaussian model is much more pronounced in this projection than on any of the original coordinate measurements. Figure 6b plots the augmenting function $f_1(\underline{\alpha}_1 \cdot \underline{X})$.

The second linear combination $\underline{\alpha}_2$ mainly involves the third and fourth coordinates. The principal difference between the current model $p_1(\underline{X})$ and the data is seen to be a substantial skewness to the left (Figure 6c). Figure 6d shows the corresponding augmenting function f_2. The linear combination associated with the third projection mostly involves the first and second coordinates. Although this iteration is trying to account for an apparent additional skewness of the data (Figure 6e), the effect is seen to be relatively small and perhaps not significant.

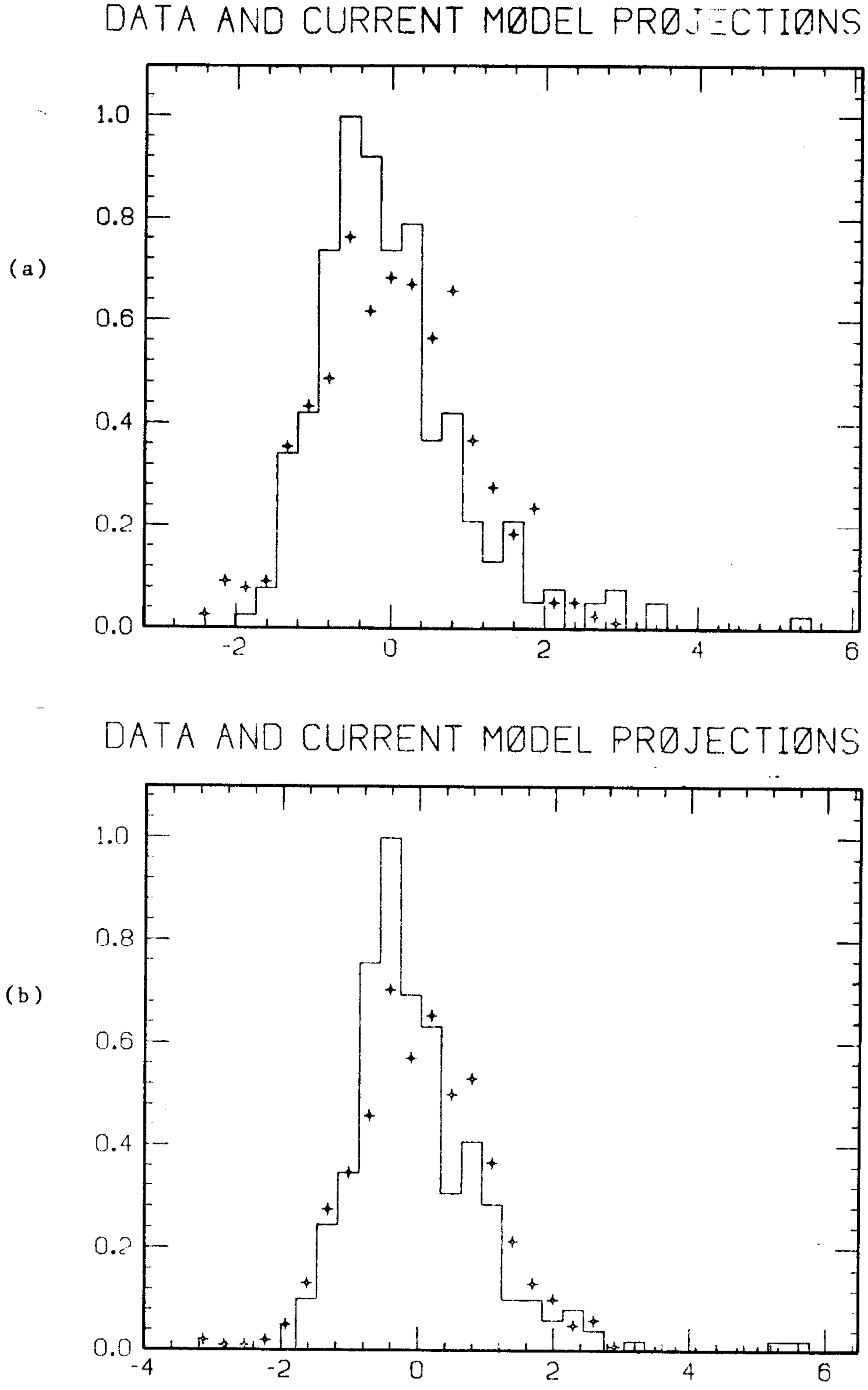


Fig. 5. (a) Hormone data: Histogram of variable 1 with Gaussian Monte Carlo superimposed. (b) Hormone data: Histogram of variable 2 with Gaussian Monte Carlo superimposed. (c) Hormone data: Histogram of variable 3 with Gaussian Monte Carlo superimposed. (d) Hormone data: Histogram of variable 4 with Gaussian Monte Carlo superimposed.

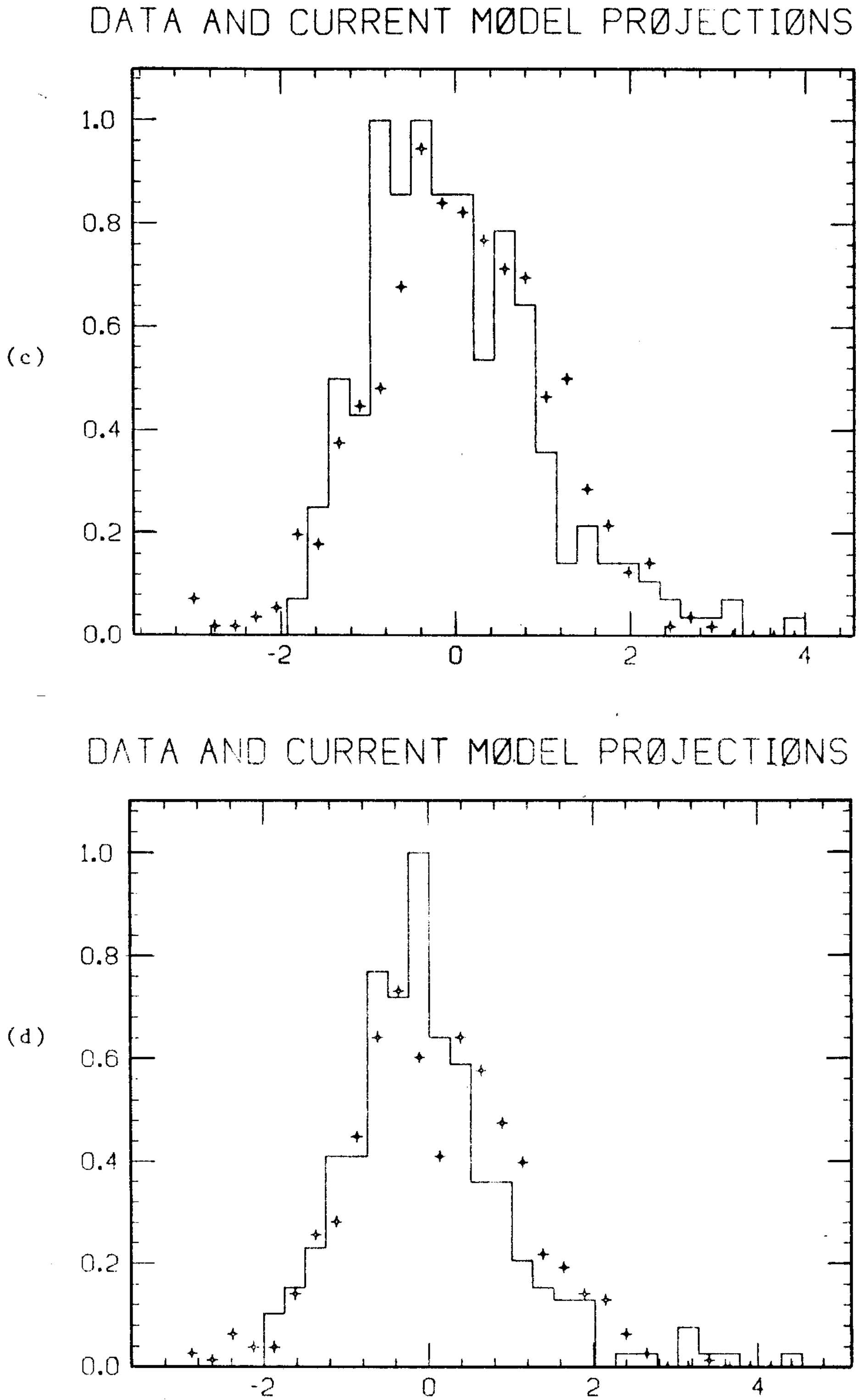

DATA AND CURRENT MØDEL PRØJECTIØNS
(c)
1.0
0.8
0.6
0.4
0.2
0.0
-2
0
2
4
DATA AND CURRENT MØDEL PRØJECTIØNS
(d)
1.0
0.8
0.6
0.4
0.2
0.0
-2
0
2
4

Fig. 5c,d

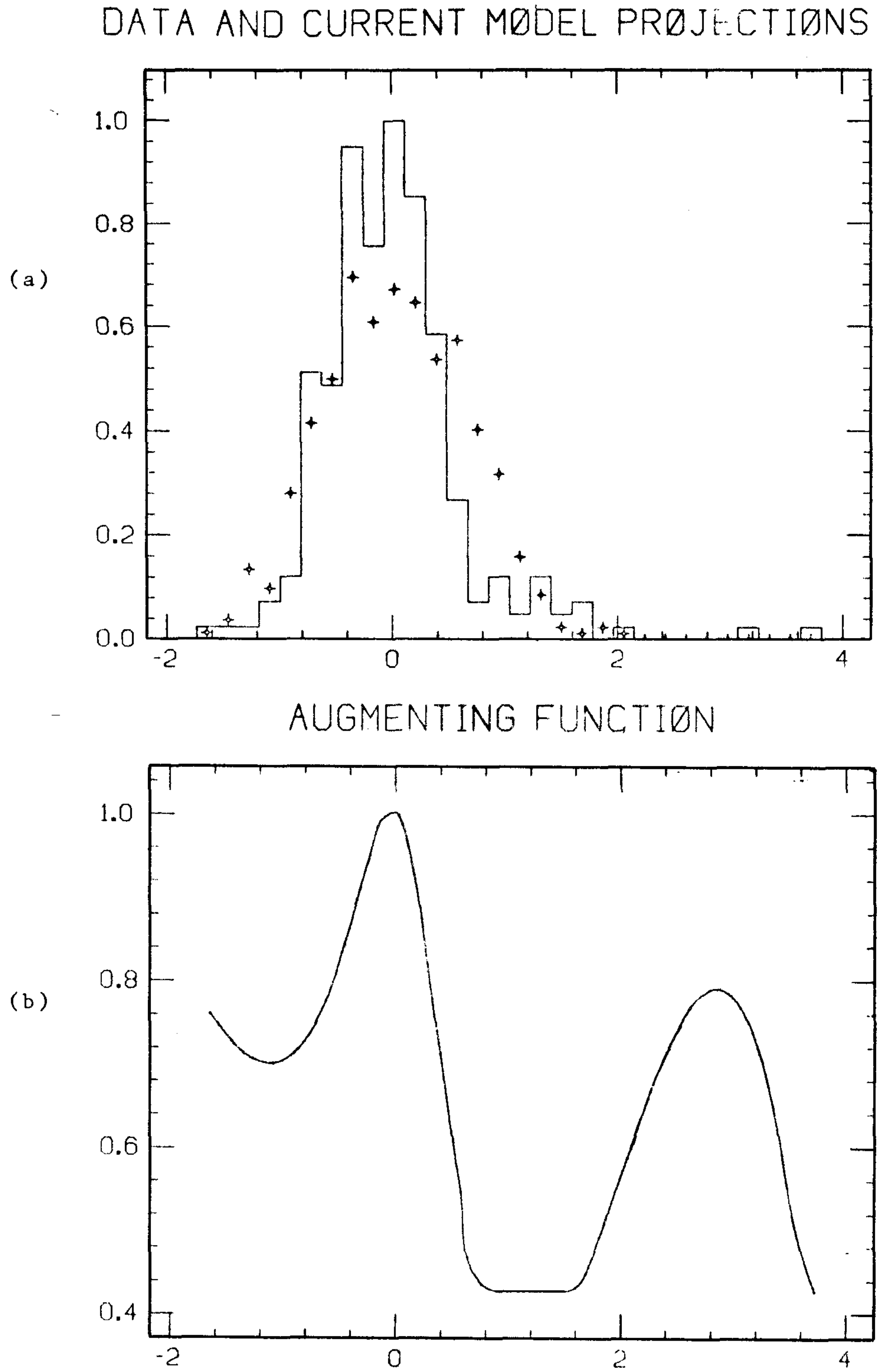


Fig. 6. Hormone Data: (a) Histogram of 1st solution linear combination $\underline{\alpha}_1$=(0.02,0.08,-0.59,0.14) with Gaussian model superimposed. (b) Augmenting function along 1st solution linear combination α_1 (0.02,0.80,-0.59,0.14). (c) Histogram of 2nd sosolution linear combination $\underline{\alpha}_2$=(-0.09,0.18,-0.45,-0.87) with current model Monte Carlo superimposed. (d) Augmenting function along second linear combination α_2=(-0.09,0.18,-0.45,-0.87). (e) Histogram of third solution linear combination α_3=(0.45, 0.88, 0.16,-0.02) with current model Monte Carlo superimposed.

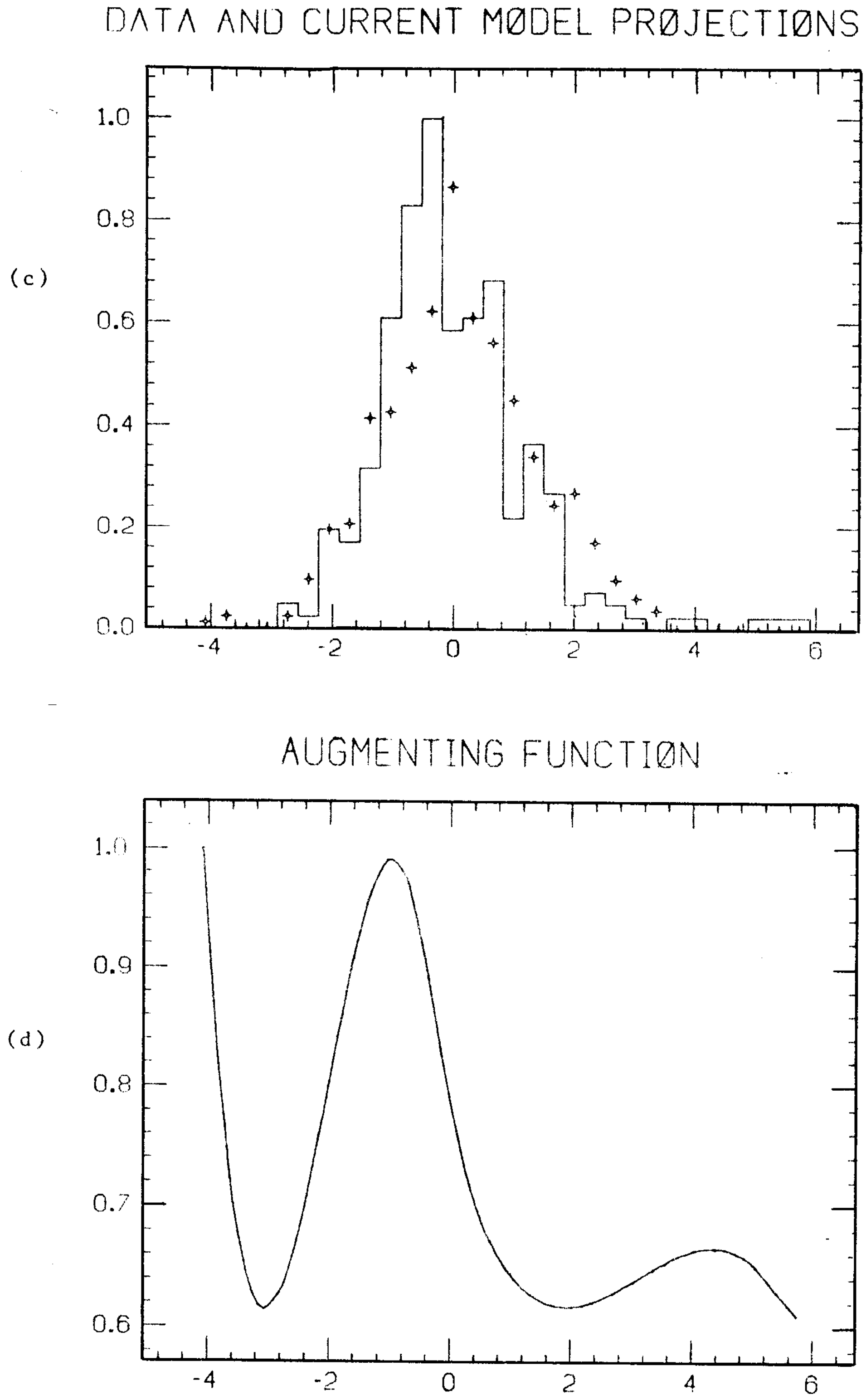

DATA AND CURRENT MØDEL PRØJECTIØNS
(c)
1.0
0.8
0.6
0.4
0.2
0.0
-4
-2
0
2
4
6
AUGMENTING FUNCTIØN
(d)
1.0
0.9
0.8
0.7
0.6
-4
-2
0
2
4
6

Fig. 6c,d.

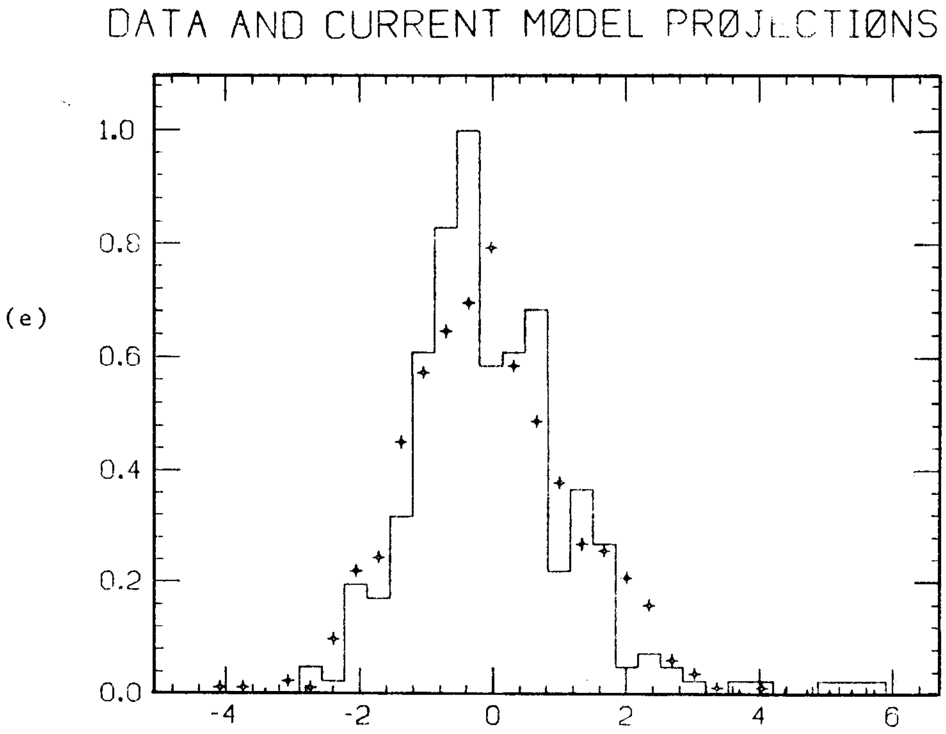


Fig. 6e.

Application of PPDE to these data reveals that a Gaussian model provides a considerably less adequate description than indicated by the coordinate projections alone. The associated graphics gives some insight into the nature of the nonnormality of the data.

7. DISCUSSION

All projection pursuit procedures share some common advantages:

- Since all estimation is carried out in a univariate setting, the large bias of kernel or near neighbor estimates in high dimensions can often be avoided.
- PP procedures do not require specification of a metric in the observation space.
- Bias is encountered with stepwise procedures when many terms are required to provide a good representation of the model

underlying the data, but only a few can be estimated due to insufficient sample size. In these cases, it is important that the first few terms be able to approximate a wide variety of functions so that the most salient features of the data can be modeled. In the limit $M \to \infty$, any regression function can be represented by (1), and any density can be represented by (2) (independent of the initial model), but even for moderate M, functions of those types constitute rich classes. In addition, the choice of the initial model permits the user to introduce any knowledge (s)he may have concerning the data, thereby allowing a further reduction in bias.

- As a data-analytic tool, projection pursuit procedures provide a set of directions $\underline{\alpha}_1 \ldots \underline{\alpha}_M$ for exploring the differences between the initial model and the data. The fact that at each stage the direction is chosen, for which the current model least adequately describes the data, makes them good candidates for that purpose. A graphical comparison of the projections of model and data, along with knowledge of the initial model, can yield considerable insight into the multivariate data distribution. Pictorial representations of each of the augmenting functions s_m, respectively f_m, along with the particular directions over which they vary, can also be quite informative since it is these functions that actually comprise the model.

There are situations in which projection pursuit procedures can be expected not to perform well. Examples of regression functions requiring a large number of terms in equation (1) are those with multiple peaks. Examples of unfavorable density functions are those with highly concave isopleths or with sperically nested isopleths of the same density value.

In addition to regression and density estimation, the projection pursuit paradigm can be applied to the problems of classification and robust estimation of covariance matrices. All projection pursuit procedures use the same set of basic operations, projection pursuit and model update. This should allow the design of an interactive system for analysis of multivariate data that covers a wide range of problems and yet is easy to learn and simple to operate.

1. Fisherkeller, M.A., Friedman, J.H., Tukey, J.W., PRIM-9: An Interactive Multidimensional Data Display and Analysis System, SLAC-PUB-1408, April 1974.
2. Stuetzle, W., Thoma, M., PRIMS-ETH: A Program for Interactive Graphical Data Analysis, Research Report No. 19, Fachgruppe fur Statistik ETH Zurich, October 1978.
3. Donoho, D., Huber, P.J., Thoma, M., (1981). The Use of Kinematic Displays to Represent High Dimensional Data. To appear in Computer Science and Statistics, Proceedings of the 14th Annual Symposium on the Interface.
4. Friedman, J.H., Stuetzle, W. (1981). Projection Pursuit Regression. To appear in JASA.
5. Stone, C.J. (1981). Admissible Selection of an Accurate and Parsimonious Normal Linear Regression Model. Ann. Statist. 9, (to appear).