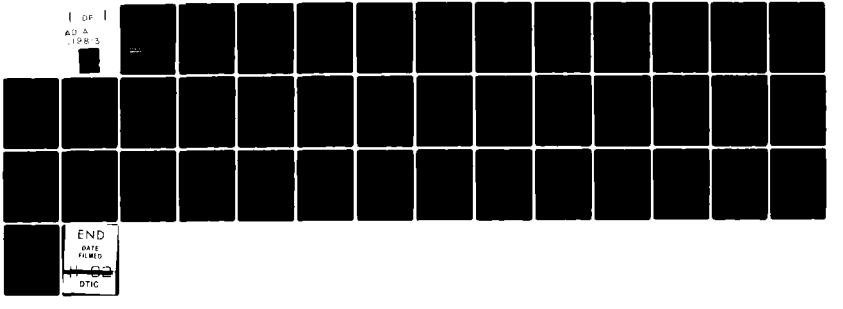
AD-A119 813 STANFORD UNIV CA PROJECT ORION F/G 12/1
PROJECTION PURSUIT DENSITY ESTIMATION.(U)
JUL 81 J H FRIEDMAN, W STUETZLE, A SCHROEDER N00014-81-K-0340
UNCLASSIFIED ORION-002 NL



(2)

PROJECTION PURSUIT DENSITY ESTIMATION

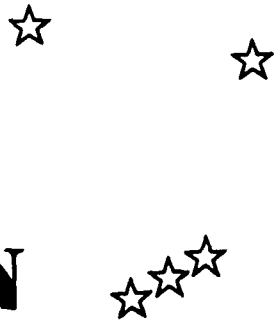
Jerome H. Friedman

Werner Stuetzle

Anne Schroeder

ORION 002

JULY 1981



Project ORION

Department of Statistics
Stanford University
Stanford, California

PROJECTION PURSUIT DENSITY ESTIMATION

by

Jerome H. Friedman

and

Werner Stuetzle

Stanford University
and
Stanford Linear Accelerator Center

and

Anne Schroeder

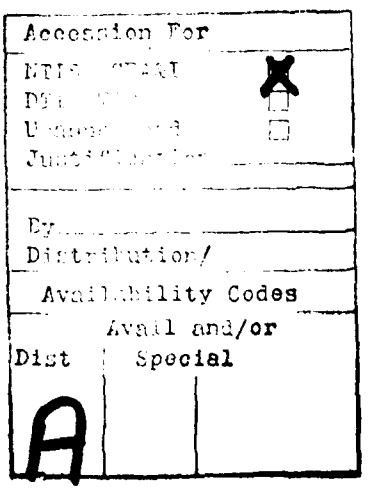
Institut National de Recherche en
Information et Automatique

ORION 002

July 1981

PROJECT ORION

Department of Statistics
Stanford University
Stanford, California



PROJECTION PURSUIT DENSITY ESTIMATION*

Jerome H. Friedman
and
Werner Stuetzle

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

and
Anne Schroeder
INRIA
Institut National de Recherche en
Informatique et Automatique
Le Cheshay, France

ABSTRACT

The projection pursuit methodology is applied to the multivariate density estimation problem. The resulting nonparametric procedure is often less biased than kernel and near neighbor methods and does not require the specification of a metric on the data measurement space. In addition, graphical information is produced that can be used to help gain geometric insight into the multivariate data distribution.

*Work supported by the Department of Energy under contract numbers DE-AC03-76SF00515 and DE-AT03-81-ER108 43, and the Office of Naval Research under contract number ONR N00014-81-K-0340.

1. Introduction

The formal goal of nonparametric density estimation is to estimate the probability density of a random vector $\underline{X}$ on the basis of iid observations $\underline{x}_1 \ldots \underline{x}_N$ without making the assumption that the density belongs to a particular parametric family. Often in practice a more important objective is to gain geometric insight into the data distribution in R^n.

Nonparametric estimation of univariate probability density functions has been extensively studied. Examples of successful methods are the related techniques of kernel estimates (Rosenblatt, 1956; Parzen, 1962), near neighbor estimates (Loftsgaarden and Quesenberry, 1965) and splines (Boneva, Kendall, and Stefanov, 1971). The direct extension of these methods to multivariate settings, however, has not been as successful in practice. This can partly be attributed to their deteriorating statistical performance caused by the so called "curse of dimensionality" (Bellman, 1961) which requires very large bandwidths (radii of neighborhoods) in order to achieve sufficient counts. The resulting estimates are then highly biased. In addition, these methods do not provide any comprehensible information about the structure of the multivariate point cloud.

Our approach to multivariate density estimation is based on the notion of projection pursuit (Friedman and Tukey, 1974, and Friedman and Stuetzle, 1981). It attempts to overcome the problem of large bias by extending the classical univariate density estimation methods to higher dimensions in a manner that involves only univariate estimation. As a by-product graphical information is produced that can be quite helpful in exploring and understanding the multivariate data distribution.

2. Overview

The goal of projection pursuit methods is to estimate multivariate functions by combinations of smooth univariate (ridge) functions of carefully selected linear combinations of the variables.

Our projection pursuit density estimation (PPDE) method constructs estimates of the form

$$p_M(\underline{x}) = p_0(\underline{x}) \prod_{m=1}^{M} f_m(\underline{\theta}_m \cdot \underline{x}) \qquad \underline{x} \in R^n \qquad (1)$$

where:

- p_M is the density estimate (current model) after M iterations of the procedure.
- p_0 is a given multivariate density function to be used as the initial model.
- $\underline{\theta}_m$ is a vector of direction cosines specifying a direction in R^n, thus

$$\underline{\theta}_m \cdot \underline{x} = \sum_{i=1}^{n} \theta_{mi} x_i$$

  is a linear combination of the original coordinate measurements.
- f_m is a univaraite function.

From (1) PPDE is seen to approximate the multivariate density by an initially proposed density p_0, multiplied (augmented) by a product of univariate functions f_m of linear combinations $\underline{\theta}_m \cdot \underline{x}$ of the coordinates. The choice of an initial density is left to the user and should reflect his best initial knowledge of the data. A Gaussian density with the same mean and covariance matrix as the sample is often a natural choice. It is

the purpose of PPDE to choose the directions $\underline{\theta}_m$ and construct the corresponding functions $f_m(\underline{\theta}\cdot\underline{x})$, the product of which estimates the ratio of the data density to the initial model density.

From (1), we obtain the recursion relation

$$p_M(\underline{x}) = p_{M-1}(\underline{x})\ f_M(\underline{\theta}_M\cdot\underline{x}) \quad . \tag{2}$$

Since f_M is used to modify p_{M-1} to obtain p_M, we refer to the f_m as "augmenting functions".

This recursive definition of the model (2) suggests a stepwise approach for its construction. At the Mth iteration there is a current model $p_{M-1}(\underline{x})$ constructed from the previous steps. (For the first step M=1 the current model is the initial model $p_0(\underline{x})$ specified by the user.) Given $p_{M-1}(\underline{x})$ we seek a new model $p_M(\underline{x})$ (2) to serve as a better approximation to the data density $p(\underline{x})$. Thus, a direction $\underline{\theta}_M$ and its corresponding augmenting function $f_{\underline{\theta}_M}(\underline{\theta}_M\cdot\underline{x})$ are chosen to maximize the goodness-of-fit to $p_M(\underline{x})$. In this context we use the log-likelihood

$$W = N\int \log[p_M(\underline{x})]\ p(\underline{x})d\underline{x}$$

as a measure of relative goodness-of-fit. From (2) we see that the likelihood achieves its maximum at the same location as

$$w(\underline{\theta},f_{\underline{\theta}}) = N\int \log[f_{\underline{\theta}}(\underline{\theta}\cdot\underline{x})]\ p(\underline{x})d\underline{x} \quad . \tag{3}$$

This is to be maximized under the constraint that $p_M(\underline{x})$ be properly normalized. That is,

$$\int p_M(\underline{x})d\underline{x} = \int p_{M-1}(\underline{x})\ f_{\underline{\theta}}(\underline{\theta}\cdot\underline{x})d\underline{x} = 1 \quad . \tag{4}$$

For a given direction $\underline{\theta}$, and known $p(\underline{x})$

$$f_{\underline{\theta}}(\underline{\theta}\cdot\underline{x}) = P(\underline{\theta}\cdot\underline{x}) \;/\; P_{M-1}(\underline{\theta}\cdot\underline{x}) \tag{5}$$

is seen to maximize (3) subject to (4). Here $P(\underline{\theta}\cdot\underline{x})$ and $P_{M-1}(\underline{\theta}\cdot\underline{x})$ represent the data and current model marginal densities (respectively) along the (one dimensional) subspace spanned by $\underline{\theta}$. Using this $f_{\underline{\theta}}$ it remains to find a direction $\underline{\theta}$ that causes (3) to achieve a maximum value. This $\underline{\theta}_M$ and its corresponding augmenting function

$$f_M(\underline{\theta}_M\cdot\underline{x}) \equiv f_{\underline{\theta}_M}(\underline{\theta}_M\cdot\underline{x}) \tag{6}$$

define the new model through (2).

In actual applications the data density $p(\underline{x})$ is unknown. We have instead a sample $\underline{x}_1, \underline{x}_2, \ldots \underline{x}_N$ of size N that is assumed to represent N independent samplings from $p(\underline{x})$. This sample is used to estimate the marginal data density $P(\underline{\theta}\cdot\underline{x})$ as well as $w(\underline{\theta}, f_{\underline{\theta}})$. A Monte Carlo technique is used to estimate the current model marginal $P_{M-1}(\underline{\theta}\cdot\underline{x})$. The ratio of the marginal density estimates is used to estimate $f_{\underline{\theta}}(\underline{\theta}\cdot\underline{x})$ (5) and the optimal value $\underline{\theta}_M$ that maximizes the likelihood estimate is determined by a numerical optimization procedure (see Friedman and Stuetzle, 1981).

3. Estimation Procedures.

We now discuss the estimation of the relevant quantities in (3) - (5). First consider the current model marginal $P_{M-1}(\underline{\theta}\cdot\underline{x})$. Without loss of generality we let $\underline{\theta}$ be the first coordinate axis, that is $\underline{\theta}\cdot\underline{x} = x_1$. Then

$$P_{M-1}(x_1) = \int p_{M-1}(\underline{x})dx_2\, dx_3 \ldots dx_n \tag{7}$$

If $P_{M-1}(x_1)$ is continuous then

$$P_{M-1}(x_1) = \lim_{h\to 0} \frac{1}{2h} \int_{x_1-h}^{x_1+h} P_{M-1}(z)dz \tag{8}$$

$$= \lim_{h\to 0} \frac{1}{2h} \int_{-\infty}^{\infty} I(x_1-h \le z \le x_1+h)\, P_{M-1}(z)dz \tag{9}$$

where

$$I(s) = \begin{cases} 1 \text{ if } s \text{ is true} \\ 0 \text{ otherwise.} \end{cases} \tag{10}$$

From (7) one has

$$P_{M-1}(x_1) = \lim_{h\to 0} \frac{1}{2h} \int I(x_1-h \le y_1 \le x_1+h)\, p_{M-1}(\underline{y})d\underline{y}$$

$$= \lim_{h\to 0} \frac{1}{2h} E_{p_{M-1}} [I(x_1-h \le y_1 \le x_1 + h)]. \tag{12}$$

Our estimate of $P_{M-1}(x_1)$ is obtained from (12) by using a small finite value for h (called the span) and employing a Monte Carlo method to estimate the expected value. A Monte Carlo sample $\underline{y}_1, \underline{y}_2, \ldots \underline{y}_{N_s}$ of size N_s is generated with density $p_{M-1}(\underline{x})$ and

$$\hat{P}_{M-1}(x_1) = \frac{1}{2hN_s} \sum_{j=1}^{N_s} I(x_1-h \le y_{1j} \le x_1+h) \tag{13}$$

is taken as our estimate of $P_{M-1}(x_1)$. Since the choice of x_1 as the

direction $\underline{\theta}$ was arbitrary (13) can equally well be written

$$\hat{P}_{M-1}(\underline{\theta}\cdot\underline{x}) = \frac{1}{2hN_s}\sum_{j=1}^{N_s} I(\underline{\theta}\cdot\underline{x} - h \le \underline{\theta}\cdot\underline{y}_j \le \underline{\theta}\cdot\underline{x}+h) \tag{14}$$

for any $\underline{\theta}$. Note that the same Monte Carlo sample can be used for all $\underline{\theta}$ and $\underline{x}$. In Appendix 2 we discuss in detail procedures for generating a Monte Carlo sample from the density $p_{M-1}(\underline{x})$.

By assumption the data represent a sample from $p(\underline{x})$ that can be used, in analogy with (14), to estimate the data marginal $P(\underline{\theta}\cdot\underline{x})$ as

$$\hat{P}(\underline{\theta}\cdot\underline{x}) = \frac{1}{2hN}\sum_{i=1}^{N} I(\underline{\theta}\cdot\underline{x} - h \le \underline{\theta}\cdot\underline{x}_i \le \underline{\theta}\cdot x+h). \tag{15}$$

From (5) our estimate of $f_{\underline{\theta}}(\underline{\theta}\cdot\underline{x})$ becomes

$$\hat{f}_{\underline{\theta}}(\underline{\theta}\cdot\underline{x}) = \frac{N_s \sum_{i=1}^{N} I(\underline{\theta}\cdot\underline{x} - h \le \theta\cdot\underline{x}_i \le \underline{\theta}\cdot\underline{x}+h)}{N \sum_{j=1}^{N_s} I(\underline{\theta}\cdot\underline{x} - h \le \theta\cdot\underline{y}_j \le \underline{\theta}\cdot\underline{x}+h)}. \tag{16}$$

This is just the ratio of the fraction of real counts to the fraction of Monte Carlo counts in an interval of width 2h centered at $\underline{\theta}\cdot\underline{x}$. In order to help stabalize the denominator we choose h to always include exactly kN_s Monte Carlo observations. In this case (16) becomes

$$\hat{f}_{\underline{\theta}}(\underline{\theta}\cdot\underline{x}) = \frac{1}{kN}\sum_{i=1}^{N} I(\underline{\theta}\cdot\underline{x} - h \le \theta\cdot\underline{x}_i \le \underline{\theta}\cdot\underline{x}+h). \tag{17}$$

Using (17) the log-likelihood (3) can be estimated as

$$\hat{w}(\underline{\theta}, f_{\underline{\theta}}) = \sum_{i=1}^{N} \log \hat{f}_{\underline{\theta}}(\underline{\theta} \cdot \underline{x}_i). \tag{18}$$

The direction $\underline{\theta}_M$ that maximizes (18) along with its corresponding augmenting function $f_M(\underline{\theta}_M \cdot \underline{x})$ (6) define the Mth term of the PPDE model (1).

4. Redundant Variable Elimination

For purposes of interpretation, it is desirable that models be parsimonious. That is, models should involve only as many variables as are required for an adequate description of the data. For models constructed by projection pursuit this means that each solution linear combination $\underline{\theta}_M$ should involve only those predictor variables that are necessary. Due to sampling fluctuations, it often happens that several variables enter into a solution linear combination (usually with small coefficients), but their removal will not substantially effect the quality of the solution. This is especially true if some of the variables are highly correlated.

Redundant variables entering into a solution linear combination $\underline{\theta}_M$ can be eliminated by the following (reverse) stepwise procedure. Each non-zero coefficient is in turn set to zero, keeping all other coefficients at their solution values; the corresponding augmenting function is computed and the log-likelihood is obtained. That variable x_U for which this (deleted) log-likelihood w_U is largest becomes a candidate for elimination. Let x_L be the variable for which the deleted log-likelihood w_L is smallest, and w_C be the log-likelihood for the complete solution (no variables deleted).

If

$$w_C - w_U > \beta(w_C - w_L) \tag{19}$$

then the elimination procedure stops and the complete solution is accepted. Otherwise x_U is deleted (coefficient set to zero) and the above procedure is repeated (next iterative pass) for all variables with non zero coefficients. This iterative procedure terminates when the candidate variable for an iteration x_U cannot be deleted [(19) true]. The quantity β is a user specified parameter with a value between 0.1 and 0.2.

5. Termination Criteria

As with any stepwise procedure, one needs a criterion for stopping the iteration at some (Mth) step. Stopping too soon can increase the bias of the estimator, while not stopping soon enough can unduly increase its variance. "Optimal" termination of stepwise procedures has been studied (see Stone, 1974, and references therein); these methods can be applied here. In practice, stepwise procedures are usually terminated subjectively, based on inspection of the successive values of goodness-of-fit criterion.

PPDE can provide several additional aids in judging whether a new step enhances the model enough to be included. One can compare $P_{M-1}(\underline{\theta}_M \cdot \underline{x})$ (the current model marginal projected onto $\underline{\theta}_M \cdot \underline{x}$) and $P(\underline{\theta}_M \cdot \underline{x})$ (the actual data marginal projected on $\underline{\theta}_M \cdot \underline{x}$). The ratio of these two densities would be the Mth augmenting function. Since $\underline{\theta}_M$ is chosen to maximize (in the likelihood sense) the difference between the projected model and data densities, their comparison in this projection represents a genuine comparison of the full multivariate densities for equality. Our experience indicates that graphical comparisons are most effective.

Graphical inspection of $f_M(\underline{\theta}\cdot\underline{x})$ can be additionally used to judge whether it should be included in the model. If the graph of $f_M(\underline{\theta}\cdot\underline{x})$ versus $\underline{\theta}_M\cdot\underline{x}$ displays a noisy pattern with no systematic tendency, then its inclusion will likely increase the variance of the density estimate. On the other hand, a definite tendency indicates that $f_M(\underline{\theta}_M\cdot\underline{x})$ is dealing with a genuine inadequacy of the present model.

6. Expression of the Results

From the formal point of view, the result of applying PPDE is an estimate of the data density specified by the initial model, a series of directions (unit vectors) $\underline{\theta}_m \in R^n$, and a corresponding set of augmenting functions $f_m(\underline{\theta}_m\cdot\underline{x})$. The augmenting functions can be stored as specific values associated with each observation. Due to their inherent smoothness however, this representation is highly redundant and a bit cumbersome. For this reason, we approximate each augmenting function by a cubic spline function

$$s_m(z) = \sum_{\ell=1}^{L} \beta_{\ell m} B_\ell(z). \tag{20}$$

The $B_\ell(z)$ are basic cubic B-splines (deBoor, 1979), the $\beta_{\ell m}$ are determined by a least squares fit of s_m to the "observations" $[\underline{\theta}\cdot\underline{x}_i, f_m(\underline{\theta}\cdot\underline{x}_i)]$ $(1 \le i \le N)$. The number of knots L is chosen to be inversely proportional to the span k (17). The internal knots are placed such that equal numbers of Monte Carlo observations fall between each pair.

7. Examples

We illustrate the use of PPDE by applying it to two examples. (A FORTRAN program implementing the PPDE procedure is available from the authors.) In all examples, the initial model $p_0(\underline{x})$ was taken to be Gaussian

with the sample mean and covariance matrix; the logarithm of the likelihood of the data sample under the initial model was arbitrarily set to zero; the size of the Monte Carlo sample was taken to be twice the data sample size.

The first example is especially simple and was chosen primarily to facilitate the exposition of the functioning of the algorithm. Here, 75 observations were generated in two dimensions from each of three Gaussian distributions, with unit covariance matrix and centers at the vertices of an equal lateral triangle of length six on each side. Figure 1a depicts the true density function by showing a few of its isopleths in the plane. In all, there are 225 data points in two dimensions. Figure 1b shows a scatterplot of the actual data. Since the (simulated) data for this example is only two-dimensional, it is possible to monitor the progress of the PPDE procedure as it attempts to iteratively construct the two-dimensional density from one-dimensional projections.

Figure 2a shows some isopleths of the initial model approximation $p_0(\underline{x})$ - Gaussian with the sample mean and covariance matrix. Figure 2b plots the data (solid histogram) and a Monte Carlo sample drawn from $p_0(\underline{x})$ (stars) as projected on the first solution linear combination $\underline{\theta}_1 = (0,1)$. Figure 2c shows a plot of the first augmenting function. Figures 2b and 2c, as well as the increase of the log-likelihood $\Delta W_1 = 79.2$, indicate that the model after the first iteration, $p_1(\underline{x}) = p_0(\underline{x})\ f_1(\underline{\theta}_1 \cdot \underline{x})$, is a substantial improvement over the initial model. Figure 2d shows some isopleths of $p_1(\underline{x})$ in the plane. Comparison of Figures 2a and 2d verifies this accessment.

Figure 3a plots the data and a Monte Carlo sample drawn from $p_1(\underline{x})$ as projected on the second solution linear combination $\underline{\theta}_2 = (0.85, 0.53)$. Figure 3b shows the second augmenting function $f_2(\underline{\theta}_2 \cdot \underline{x})$ versus $\underline{\theta}_2 \cdot \underline{x}$. The increase

in log-likelihood $\Delta W_2 = 108.6$ and Figures 3a and 3b indicate that $p_2(\underline{x}) = p_0(\underline{x})\ f_1(\underline{\theta}_1 \cdot \underline{x})\ f_2(\underline{\theta}_2 \cdot \underline{x})$ is a substantial improvement over $p_1(\underline{x})$. This verified by Figure 3c which shows some isopleths of $p_2(\underline{x})$. The three peak structure of $p(\underline{x})$ -the true data density- is now reproduced in the estimate $p_2(\underline{x})$.

Figure 4a and 4b show the data and Monte Carlo -drawn from $p_2(\underline{x})$- projected on the third solution linear combination $\underline{\theta}_3 = (1,0)$, and the corresponding augmenting function $f_3(\underline{\theta}_3 \cdot \underline{x})$. The log-likelihood improvement $\Delta W_3 = 16.2$, as well as Figures 4a and 4b, indicate that $p_3(\underline{x}) = p_2(\underline{x})\ f_3(\underline{\theta}_3 \cdot \underline{x})$ provides at best a little improvement over $p_2(\underline{x})$. Figure 4c shows isopleths of $p_3(\underline{x})$. From Figures 4b and 4c, we see that $p_3(\underline{x})$ has slightly increased the size of the peak centered at (3.0, 2.5) which was a little underestimated by $p_2(\underline{x})$.

The small increase in log-likelihood for the third iteration as compared to the increases corresponding to the previous iterations, as well as inspection of Figures 4a and 4b, indicate that $p_3(\underline{x})$ does not provide sufficient enhancement of the density estimate to warrant its acceptance. We would thus terminate the iterative procedure and take as our final density estimate, $p_2(\underline{x})$. More iterations would be able to provide further refinement of the density estimate if the sample size were large enough to control the additional variance that would be introduced. In order to exactly reproduce the data density of this example, an infinite number of iterations would be needed, requiring infinite sample size. As seen from Figure 3c, however, only two augmenting functions of these particular linear combinations $(\underline{\theta}_1, \underline{\theta}_2)$ provide a reasonably good approximation.

Table 1 compares PPDE to ten nearest neighbor density estimation for this same problem of three unit covariance Gaussian clusters

centered at the vertices of an equal lateral triangle of length six on a side. In addition to the two-dimensional example, we also consider this problem in five and ten dimensions. The quantity shown in Table 1, for both estimates, is the expected squared difference between the true density and the estimate, divided by the variance of the true density, and then subtracted from one. A value of one for this quantity means that the estimate is perfect, while a value near zero indicates that the variation of the estimate from the true density is nearly as large as the variation of the true density itself. The values shown in Table 1 are the averages of the results obtained in ten Monte Carlo replications of this example.

Table 1 indicates that in two dimensions their performance is comparable. However, as the dimension of the measurement space increases, the performance of the ten nearest neighbor estimate degrades much more rapidly than PPDE, "explaining" only about 20% of the density variation in ten dimensions.

The next example illustrates the use of PPDE in a purely data analytic setting. For this example, we use data from the diabetes study of Reaven and Miller (1979). For each of 145 subjects in the study, five variables were measured: (1) relative weight, (2) a measure of glucose tolerance, (3) a second measure of glucose tolerance - glucose area, (4) a measure of insulin secretion - insulin area, and (5) a measure of how glucose and insulin interact - SSPG. Variables (2) and (3), the two measures of glucose tolerance, exhibit a high degree of linear association ($r = 0.96$) so that only variables (1), (3), (4), and (5) are considered.

Our purpose with this example is to see how well the four-dimensional data density $p(\underline{x})$ can be represented as a product of two two-dimensional marginal densities $p_{ab}(x_a,x_b)$ $p_{cd}(x_c,x_d)$. If the data density could be factored into such a product for a specific pairing of variables, (ab) (cd),

then all of the data sturcturing in the full four-dimensional space would be apparent by viewing the two scatterplots - variable a versus variable b, and variable c versus variable d.

Unlike the previous example, the initial model is not explicitly defined. Here the initial model is the factored approximation

$$p_o(\underline{x}) \equiv p_{ab}(x_a, x_b)\, p_{cd}(x_c, x_d) \tag{21}$$

with a specific choice for the variables a, b, c and d. The two-dimensional marginal densities in (21) are taken to be the actual data as projected onto the subspaces spanned by (x_a, x_b) and (x_c, x_d), respectively. Since it is not our purpose to provide explicit density estimates, it is not necessary to have an explicit (computable) representation for $p_o(\underline{x})$. All that is necessary is that we be able to draw a sample from it. Such a sample of size N (here N = 145) is generated by randomly permuting the observation labels of the $(x_a x_b)$ pairs with respect to the $(x_c x_d)$ pairs. Let $(r_1, r_2, \ldots, r_N)$ be a random permutation of the integers $(1,2,\ldots,N)$. The Monte Carlo sample from the initial model is taken to be the four tuples:

$$x_{1a}\ x_{1b}\ x_{r_1 c}\ x_{r_1 d}$$

$$x_{2a}\ x_{2b}\ x_{r_2 c}\ x_{r_2 d}$$

$$\vdots$$

$$x_{Na}\ x_{Nb}\ x_{r_N c}\ x_{r_N d}$$

As many Monte Carlo observations as needed can be obtained by repeating this procedure with different random permutations $(r_1, r_2, \ldots, r_N)$.

Table 2 shows the increase in log-likelihood achieved by PPDE, after two and four iterations, starting with the four different initials models (21) specified by the four distinct groupings (a,b), (c,d). It is clear that the least improvement was associated with

$$p_0(\underline{x}) = p_{13}(x_1,x_3)\ p_{24}(x_2,x_4), \tag{22}$$

indicating that this factorization gives the best representation of the actual data density. Figure 5a shows a scatterplot of x_1 versus x_3 and Figure 5b shows x_2 versus x_4 for the data sample. The results in Table 2 indicate that this is the best pair of plots to view the four-dimensional data structuring. However, starting with an initial density equal to the product of the two (two-dimensional) densities shown in Figures 5a and 5b, PPDE was able to construct a model with substantially greater likelihood (Table 2) indicating that Figures 5a and 5b do not reveal all of the data structuring in the full four-dimensional space.

This is verified in Figure 5c where the 145 data points (solid) and 145 Monte Carlo points drawn from $p_0(\underline{x})$ (22) ("+" signs) are shown projected on the plane spanned by the first two solution directions $\underline{\theta}_1 = (-0.29, -0.37, 0.38, 0.80)$ and $\underline{\theta}_2 = (-0.08, 0.92, -0.37, -0.11)$. The horizontal axis is $\underline{\theta}_1 \cdot \underline{x}$ and the vertical axis is $\left(\underline{\theta}_2 - (\underline{\theta}_2 \cdot \underline{\theta}_1)\ \underline{\theta}_1\right) \cdot \underline{x} / |\underline{\theta}_2 - (\underline{\theta}_2 \cdot \underline{\theta}_1)\ \underline{\theta}_1|$. The data are seen to have somewhat the same shape as the factored approximation (22) but to be more tightly concentrated, especially in the circular "ball" centered at (0,0).

8. Discussion

As a formal estimator of a multivariate density function PPDE shares advantages common to projection pursuit procedures. Since all estimation is carried out in a univariate setting, the high bias inherent in other multivariate nonparametric density estimators can often be avoided. PPDE does not require the specification of a metric in the n-dimensional data space. Also, the PPDE estimate can be represented in a concise functional form [(1)] and [(20)].

Bias is encountered with stepwise procedures when many terms are required to provide a good representation of the true data density, 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 density can be modeled. In the limit $M \to \infty$, any density function can be represented by (1) (for any p_0), but even for moderate M, functions of this form constitute a rich class. In addition, the choice of initial model p_0 permits the user to introduce any knowledge he may have concerning the density function, thereby allowing a further reduction in bias.

The success of PPDE will, of course, depend on the particular nature of the actual data density. Examples of density functions requiring large M in (1) are those with highly concave isopleths or spherically nested isopleths of the same density value (unless, of course, this structure is anticipated and incorporated in the initial model p_0).

APPENDIX 1: Back Adjustment of the Augmenting Functions.

The basic iterative procedure described in Section 2 is (using the language of linear regression) "stagewise" in that each $\underline{\theta}_M$ and its corresponding f_M are chosen as the solution to an optimization problem holding all previous $\underline{\theta}_m$ and f_m ($m<M$) fixed. It is sometimes possible to improve the goodness-of-fit of the model (without decreasing the number of degrees of freedom) by refitting all f_m ($m \leq M$) after each f_M is included in the model. This is done in an iterative manner similar to the basic (outer) iteration procedure except that the directions $\underline{\theta}_m$ ($1 \leq m \leq M$) are held fixed to avoid the (costly) numerical optimization. At each stage m of this inner iterative procedure $f_m(\underline{\theta}\cdot\underline{x})$ is readjusted to maximize the log-likelihood (3) given all $f_j(\underline{\theta}_j\cdot\underline{x}), j\neq m$. One complete pass through this inner iteration produces a new set of augmenting functions comprising a model with (possibly) higher likelihood. Since this pass has changed each f_m it is possible that yet another pass can increase the likelihood still further. Thus, the passes themselves are iterated until no increase in likelihood is observed.

We now discuss the calculation of a new $f_m(\underline{\theta}_m\cdot\underline{x})$ given $f_j(\underline{\theta}_j\cdot\underline{x})$ $j\neq m$. Let

$$p_{(m)}(\underline{x}) = p_o(\underline{x}) \prod_{j\neq m} f_j(\underline{\theta}_j\cdot\underline{x}) = p_M(\underline{x})/f_m(\underline{\theta}_m\cdot\underline{x}). \tag{A1}$$

We seek a new function $f'_m(\underline{\theta}_m\cdot\underline{x})$ that maximizes

$$w'(f'_m) = N\int \log[f'(\underline{\theta}_m\cdot\underline{x})]\, p(\underline{x})d\underline{x} \tag{A2}$$

subject to the constraint

$$\int p_{(m)}(\underline{x}) f'_m(\underline{\theta}_m\cdot\underline{x})d\underline{x} = 1. \tag{A3}$$

the solution is

$$f'_m(\underline{\theta}_m \cdot \underline{x}) = \frac{P(\underline{\theta}_m \cdot \underline{x})}{P_{(m)}(\underline{\theta}_m \cdot \underline{x})} \tag{A4}$$

where the numerator and denominator represent the corresponding projected marginal densities. These marginal densities are estimated in the manner described in Section 3. The resulting estimate for f'_m then replaces f_m in the new model $p_M(\underline{x})$.

APPENDIX 2: Monte Carlo Sampling

To apply PPDE, it is necessary to draw a Monte Carlo sample from the initial model $p_0(\underline{x})$, respectively from the current model $p_{M-1}(\underline{x})$. For many choices of $p_0(\underline{x})$, there exist special algorithms that allow efficient direct sampling (see Everett and Cashwell, 1973). Densities for which this is not the case can be sampled using the accept/reject method.

Suppose a Monte Carlo sample drawn from density $q(\underline{x})$ is available and one wishes a sample drawn from $p(\underline{x})$. Let

$$\gamma = \max_{\underline{x}} \frac{p(\underline{x})}{q(\underline{x})} . \tag{A5}$$

Consider each Monte Carlo observation $\underline{x}_i$ in turn. For each, draw a random number r_i in the interval [0,1]. If $r_i\gamma \leq p(\underline{x}_i)/q(\underline{x}_i)$, then the ith observation is accepted; otherwise it is rejected. The accepted Monte Carlo observations will be a random sample drawn from $p(\underline{x})$.

This accept/reject method can be used to draw a random sample from any density $p(\underline{x})$. The efficiency of the procedure (number accepted, divided by number accepted plus number rejected) will be greater the closer $q(\underline{x})$ resembles $p(\underline{x})$ in the sense of low variability of $p(\underline{x})/q(\underline{x})$.

The form of $p_{M-1}(\underline{x})$ [(1) and (2)] permits reasonably efficient sampling with the accept/reject method. First the random sample drawn from $p_{M-2}(\underline{x})$, —available from the previous iteration— is considered. From (1) and (A4), we have

$$\frac{p_{M-1}(\underline{x})}{p_{M-2}(\underline{x})} = \prod_{m=1}^{m-2} \frac{f'_m(\underline{\theta}_m \cdot \underline{x})}{f_m(\underline{\theta}_m \cdot \underline{x})} \quad f'_{M-1}(\underline{\theta}_{M-1} \cdot \underline{x}) \tag{A6}$$

We estimate the maximum value of (A6) by its largest value over the data sample. Applying the accept/reject procedure to the p_{M-2} Monte Carlo sample using (A6), yields a (smaller) sample drawn from $p_{M-1}(\underline{x})$. The remaining Monte Carlo observations are drawn from $p_{M-1}(\underline{x})$ by applying the accept/reject procedure to a sample from $p_0(\underline{x})$ using

$$\frac{p_{M-1}(\underline{x})}{p_0(\underline{x})} = \prod_{m=1}^{M-1} f_m'(\underline{\theta}_m \cdot \underline{x}) \quad . \tag{A7}$$

Again, the maximum value of (A7) is estimated by its largest value over the data sample.

These maximum value estimates are clearly biased. The result of this bias is to introduce a small (usually negligible) additional bias to the resulting density estimates. This bias contribution can be eliminated with the following procedure. Let $\hat{\gamma}$ be the estimated maximum value for $r(\underline{x})$. Monte Carlo observations $\underline{y}$ for which $r(\underline{y}) \leq \hat{\gamma}$ are accepted or rejected as described above. If $r(\underline{y}) > \hat{\gamma}$ then the observation is included in the sample L times where L is the integer part of $r(\underline{y})/\hat{\gamma}$. The quantity $r(\underline{y}) - L$ is then used with the standard accept/reject procedure to determine if $\underline{y}$ is accepted yet another time.

References

Bellman, R.E. (1961) "Adaptive Control Processes," Princeton Univ. Press Princeton, New Jersey.

Boneva, L.I., Kendall, D.G., and Stefanov, I. (1971) Spline Transformations. J. Roy. Statist. Soc. B. 33, 1-70.

de Boor, C. (1978) "A Practical Guide to Splines" Springer - Verlag.

Everett, C.J. and Cashwell, E.D. (1972). A Monte Carlo Sampler. Report LA-5061-MS, Los Alamos Scientific Laboratory, New Mexico.

Friedman, J.H. and Tukey, J.W. (1974) A Projection Pursuit Algorithm for Exploratory Data Analysis. IEEE Trans. Computers C-23, 881-890.

Friedman, J.H. and Stuetzle, W. (1981) Projection Pursuit Regression. J. American Statist. Assoc. December.

Loftsgaarden, D.O. and Quesenberry, C.P. (1965) A Nonparametric Density Function Ann. Math. Statist., 36 1049-1051.

Parzen, E. (1962). On the Estimation of a Probability Density Function and the Mode. Ann. Math. Statist. 33, 832-837.

Reaven, G.M. and Miller, R.G. (1979) An Attempt to Define the Nature of Chemical Diabetes Using Multidimensional Analyses. Diabetologia 16, 17-24.

Rosenblatt, M. (1965) Remarks on Some Nonparametric Estimates of a Density Function. Ann. Math. Statist. 27, 832-837.

Stone, H.M. (1974) Cross-Validatory Choice and Assessment of Statistical Predictions. J. Roy. Statist. Soc. B-36, 111-147.

TABLE 1

Comparison of PPDE and 10 nearest neighbor density estimation (first example).

$$1.0 - \left\{ E\left[(\hat{p}-p)^2 \right] \; / \; \mathrm{Var}(p) \right\}$$

Dimension	PPDE	10 nearest neighbors
2	76	77
5	75	57
10	64	22

TABLE 2

Increase in (log) likelihood of PPDE solutions from factored initial model $p_o(\underline{x}) = p_{ab}(x_a, x_b)\, p_{cd}(x_c, x_d)$

(third example)

Combination				Number of Iterations	
a	b	c	d	2	4
1	2	3	4	85.7	124.1
1	3	2	4	47.1	76.3
1	4	2	3	85.8	122.1

FIGURE CAPTIONS

Figure 1:

(a) Isopleths of a Two Dimensional Density

(b) A Sample of Size 225 Drawn From the Density

Figure 2:

(a) Isopleths of Initial Model - Gaussian With Sample Mean and Covariance

(b) Data (Histogram) and Monte Carlo From Initial Model $p_0(\underline{x})$ Along First Solution Linear Combination $\underline{\theta}_1 = (0,1)$.

(c) First Augmenting Function f_1.

(d) Isopleths of $p_1(\underline{x})$.

Figure 3:

(a) Data (Histogram) and Monte Carlo from $p_1(\underline{x})$ Along Second Solution Solution Linear Combination $\underline{\theta}_2 = (0.85, 0.53)$.

(b) Second Augmenting Function f_2.

(c) Isopleths of $p_2(\underline{x})$.

Figure 4

(a) Data (Histogram) and Monte Carlo from $p_2(\underline{x})$ Along Third Solution Linear Combination $\underline{\theta}_3 = (1,0)$.

(b) Third Augmenting Function f_3.

(c) Isopleths of $p_3(\underline{x})$.

Figure 5:

(a) Diabetes data: x_1 versus x_3 .

(b) Diabetes data: x_2 versus x_4 .

(c) Diabetes Data (Solid) and Monte Carlo From Factored Model $p_0(\underline{x})$ ("+") Projection Onto Plane Spanned by First Two Solution Linear Combinations.

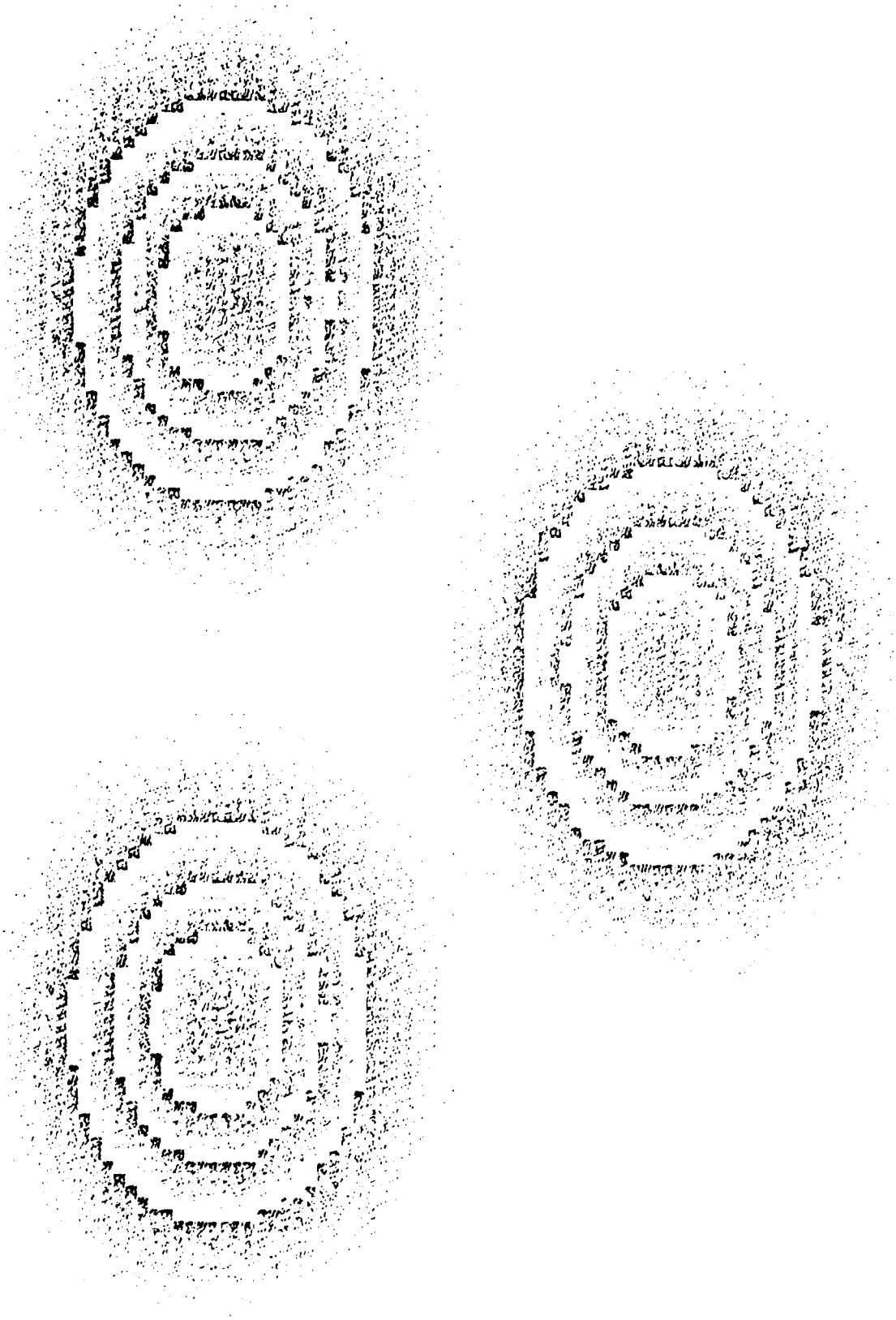


Figure 1a

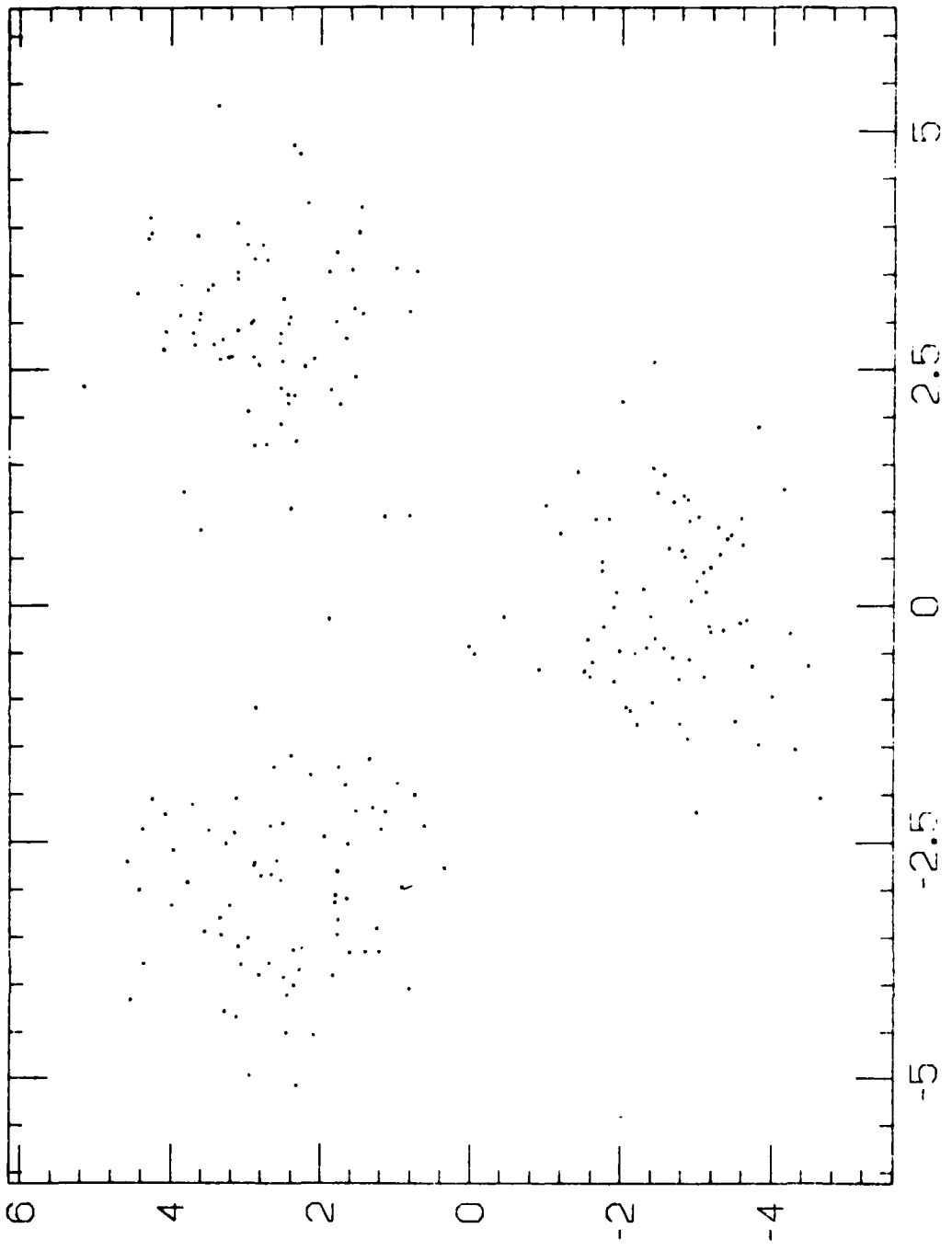


Figure 1b

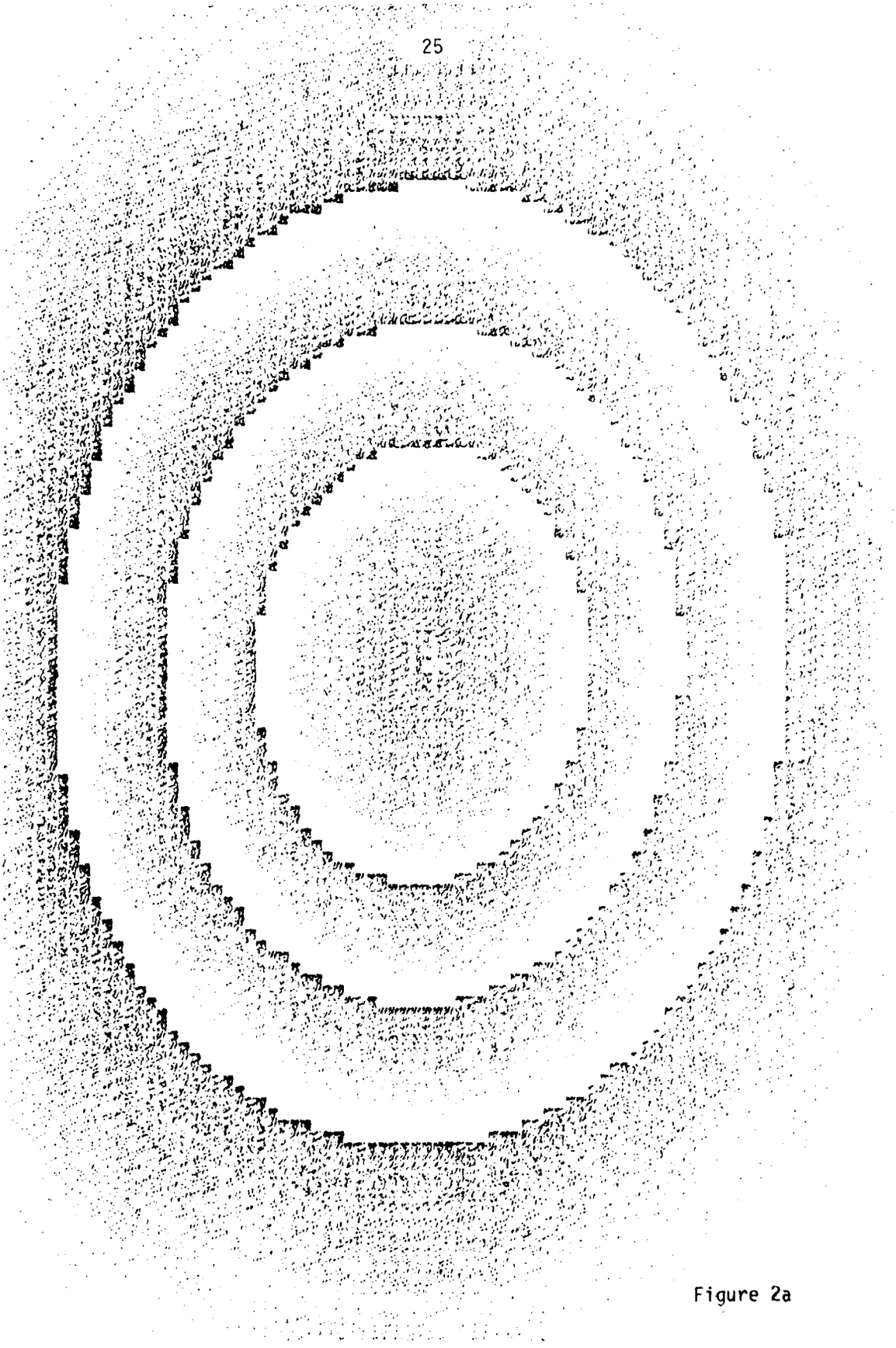


Figure 2a

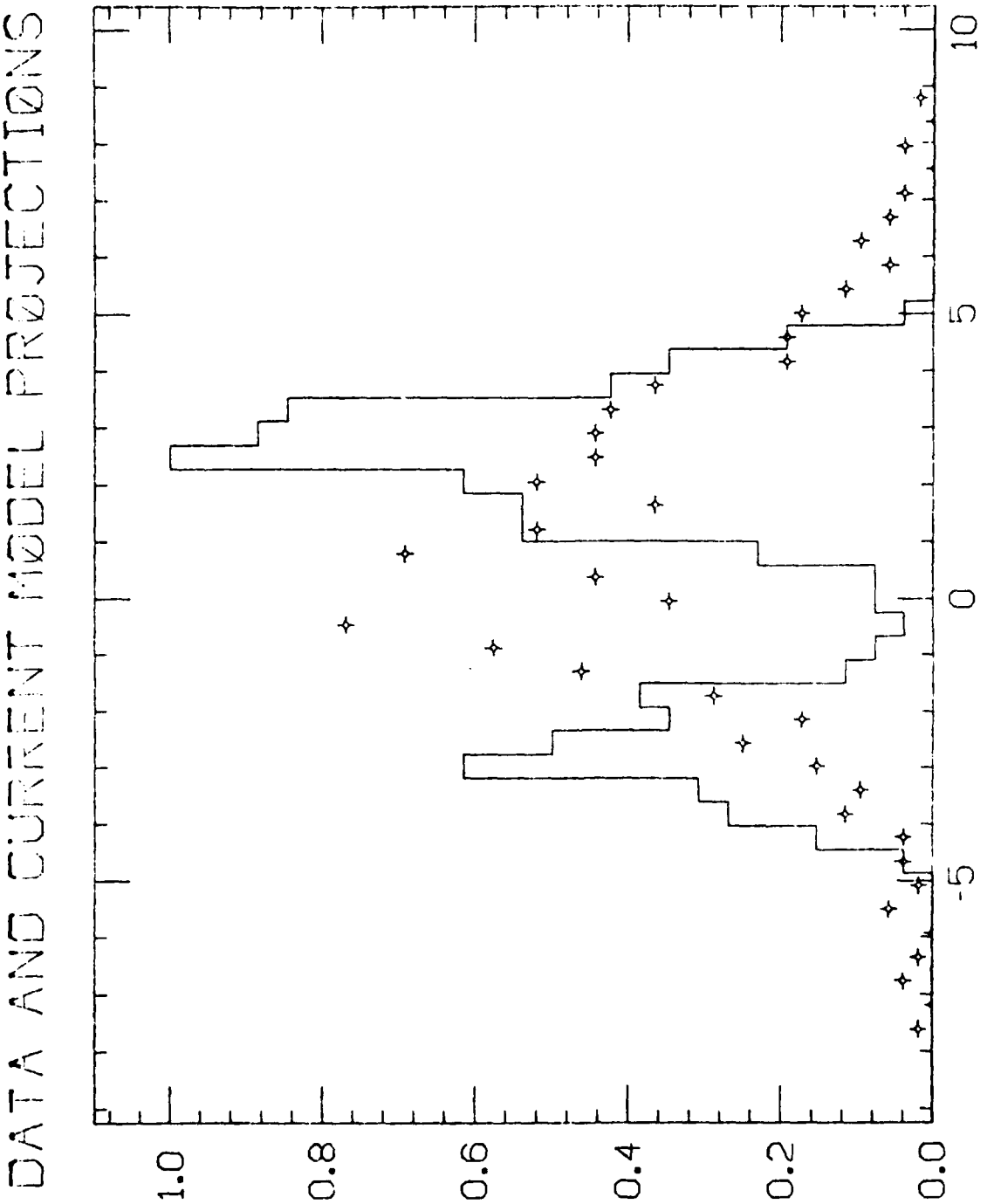


Figure 2b

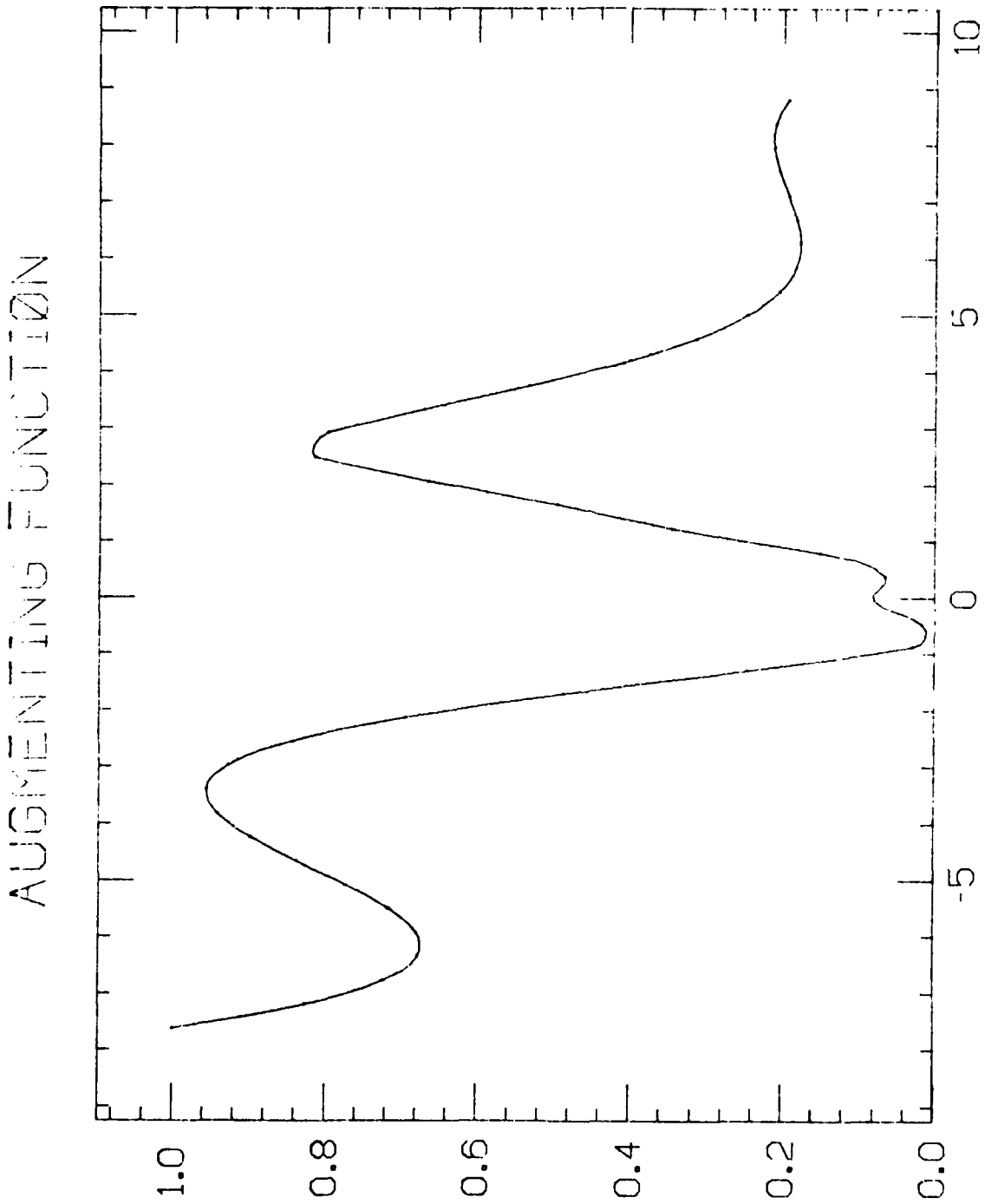


Figure 2c

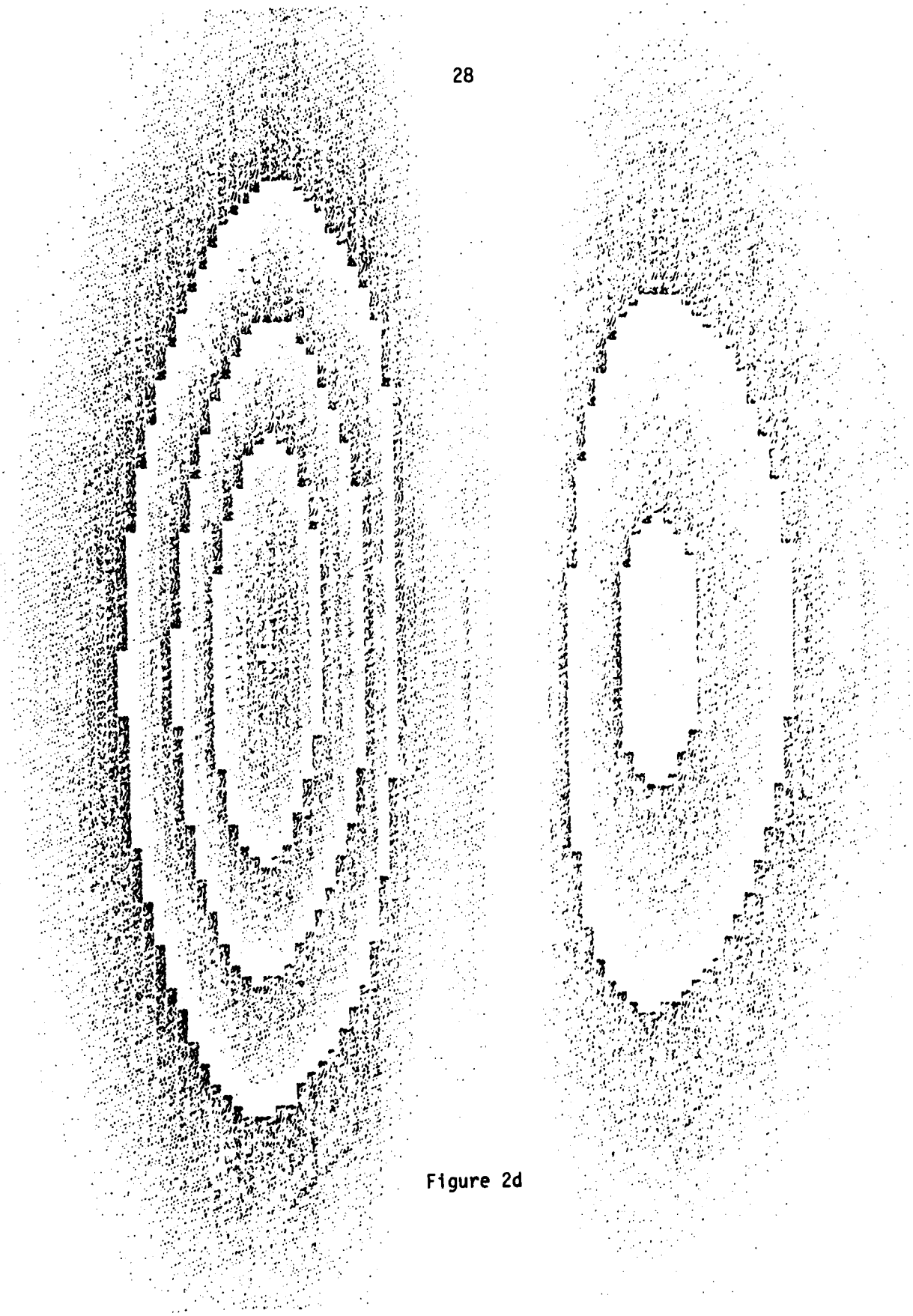


Figure 2d

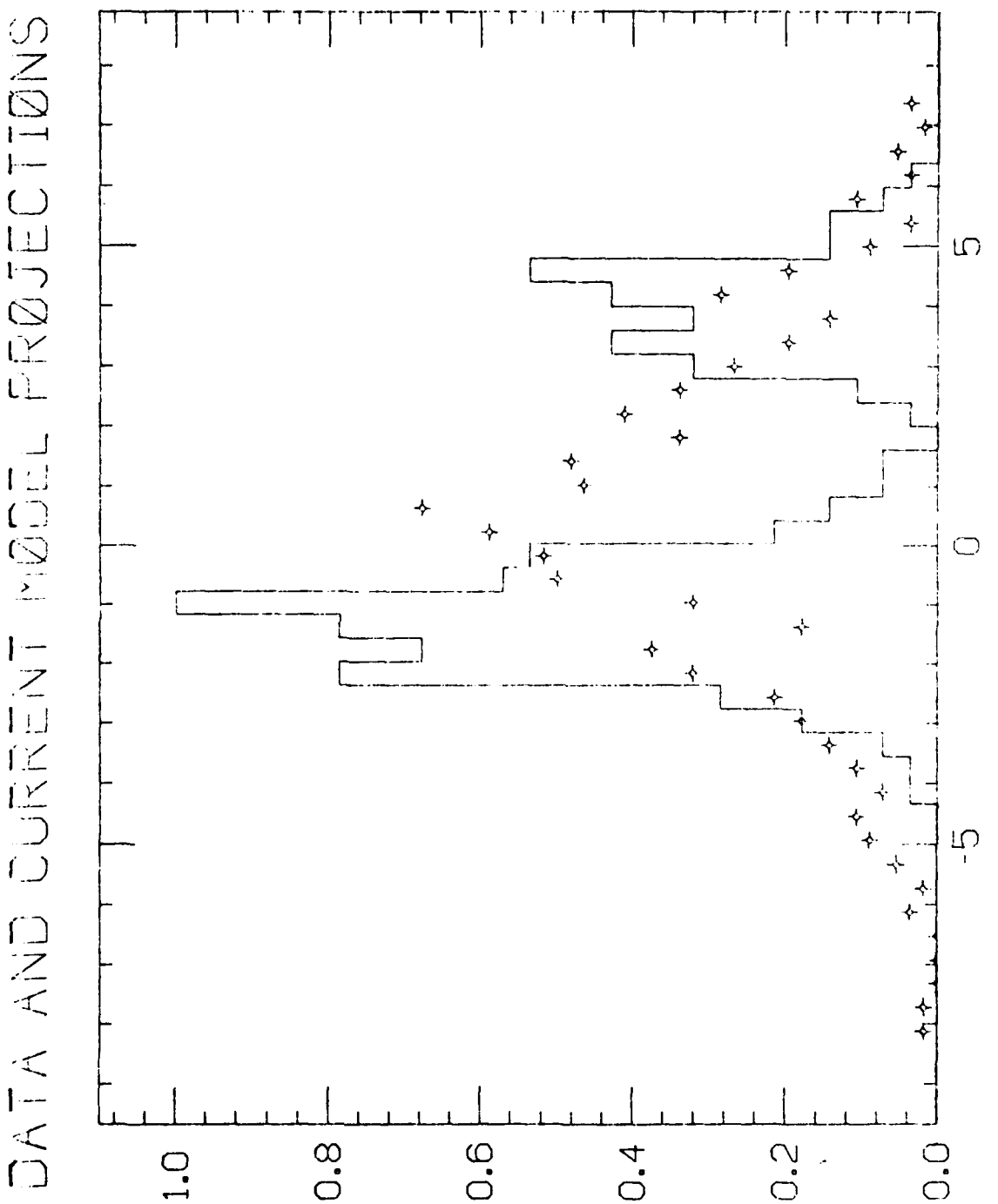


Figure 3a

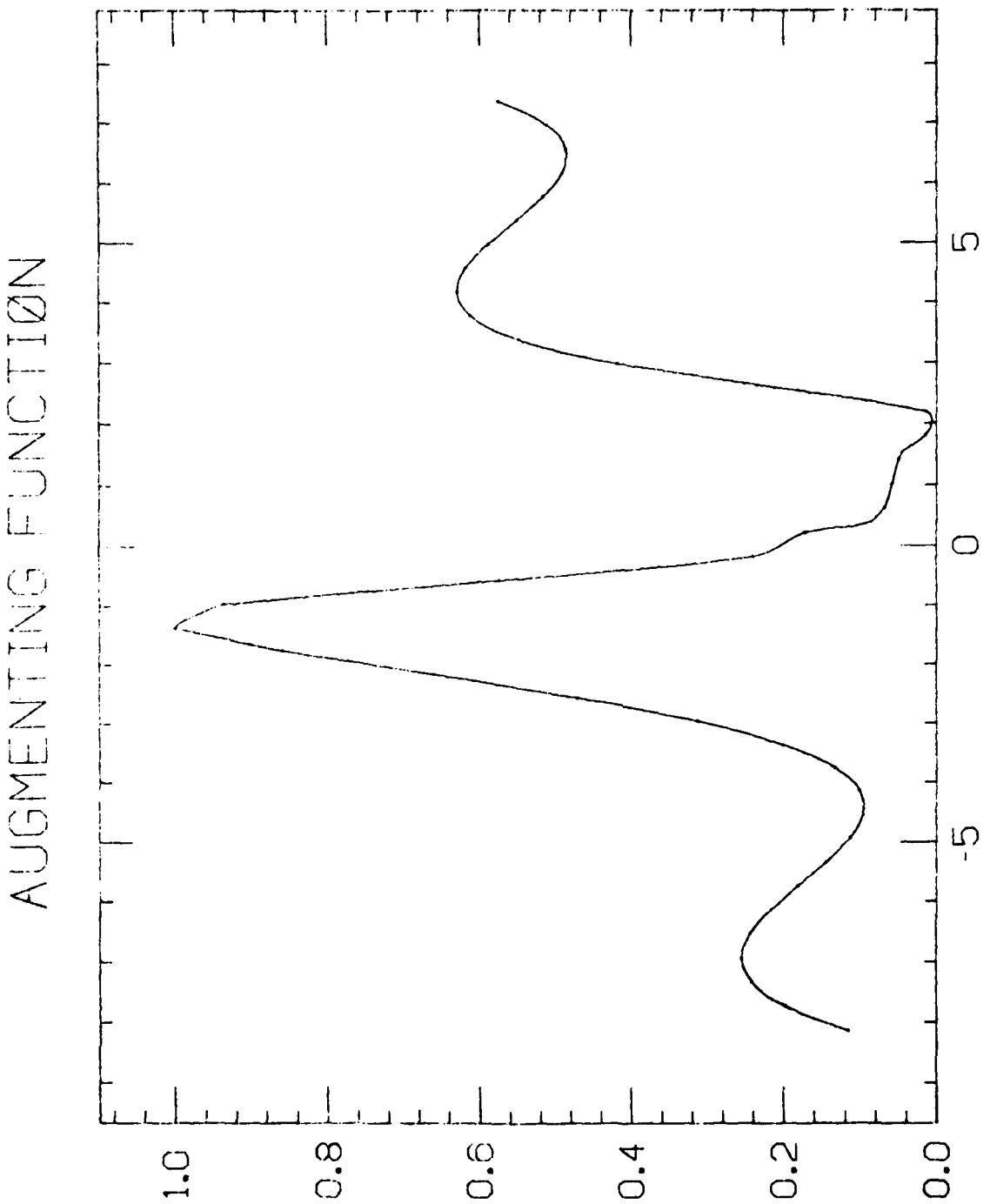


Figure 3b

Figure 3c

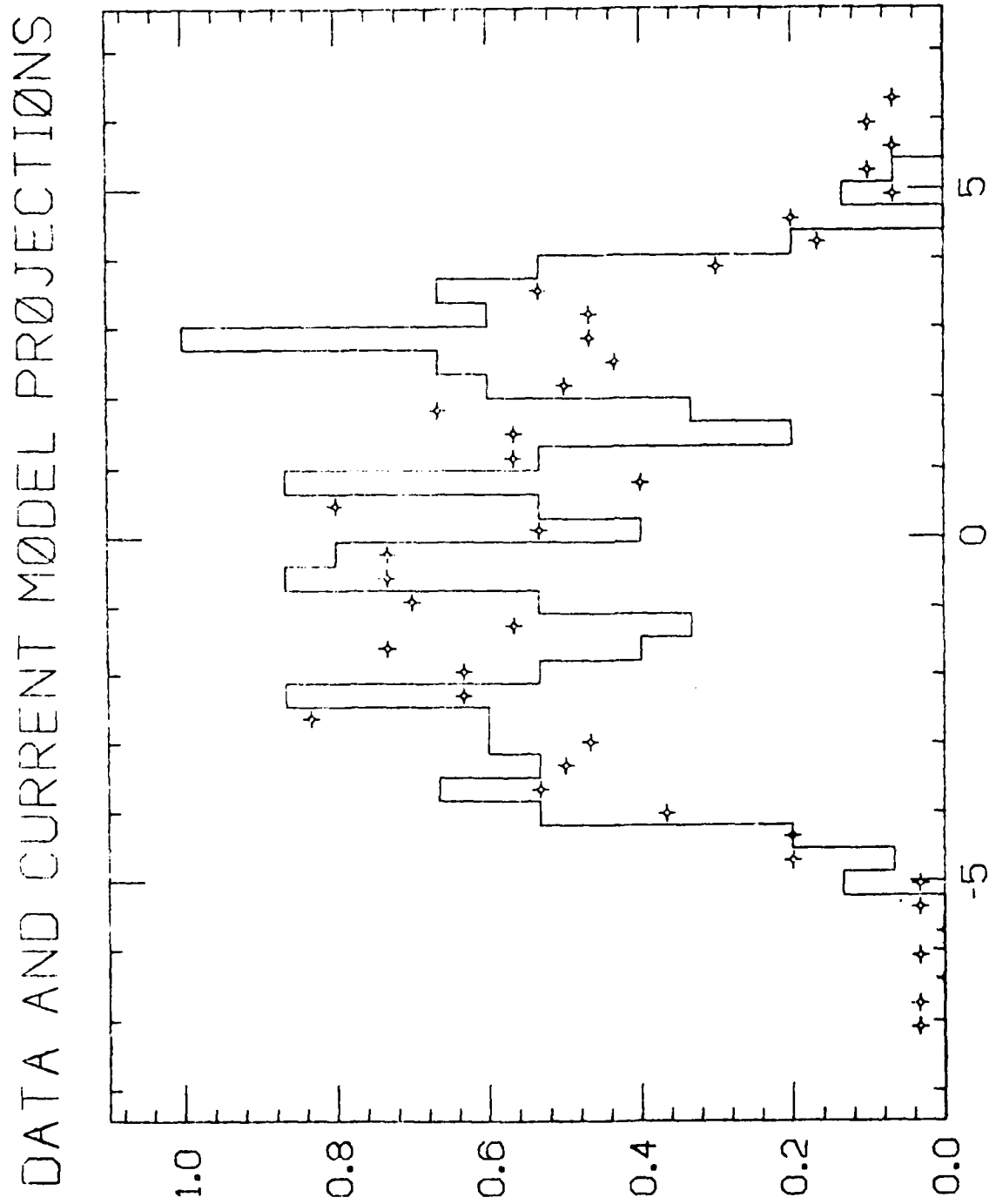


Figure 4a

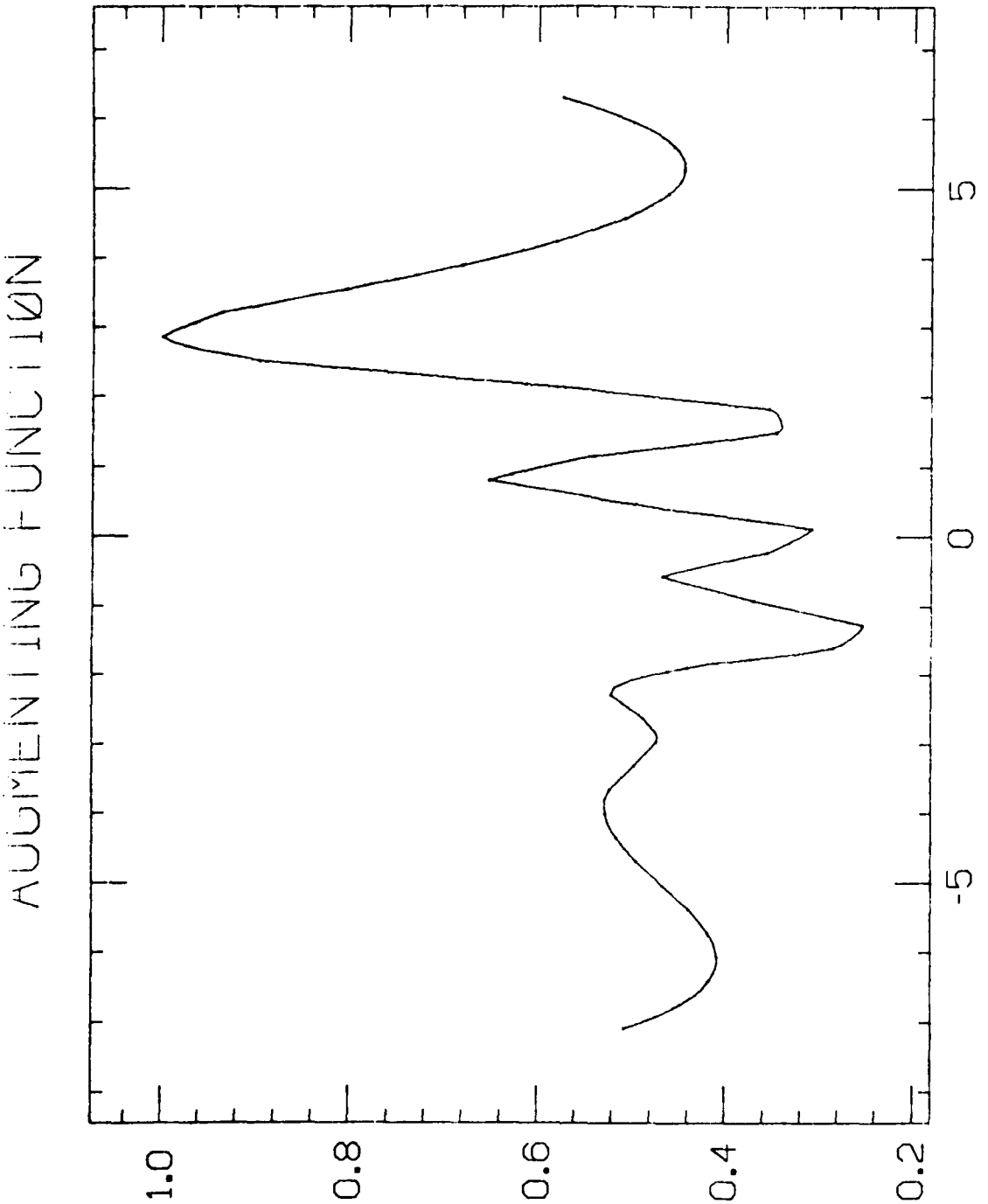


Figure 4b

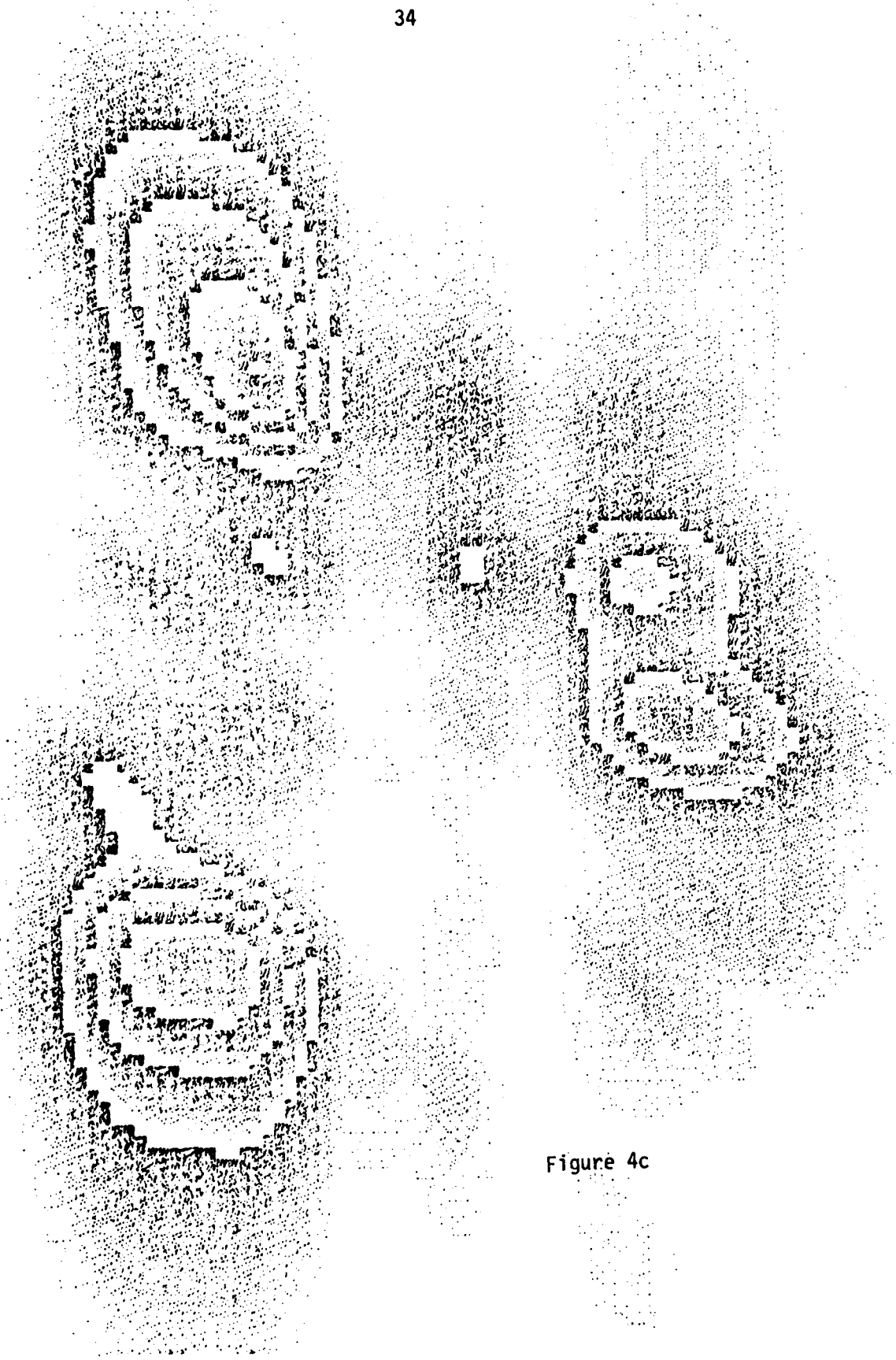


Figure 4c

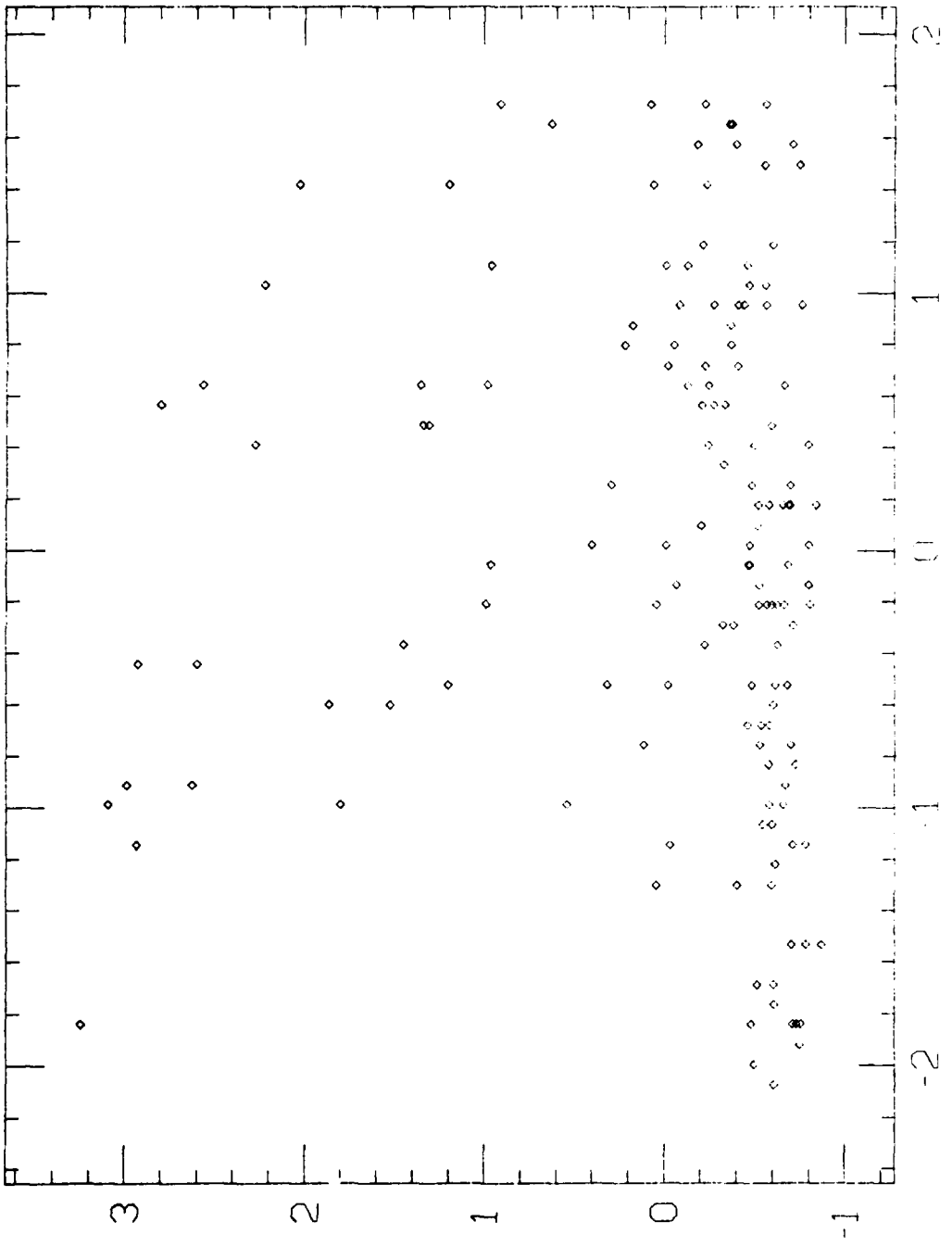


Figure 5a

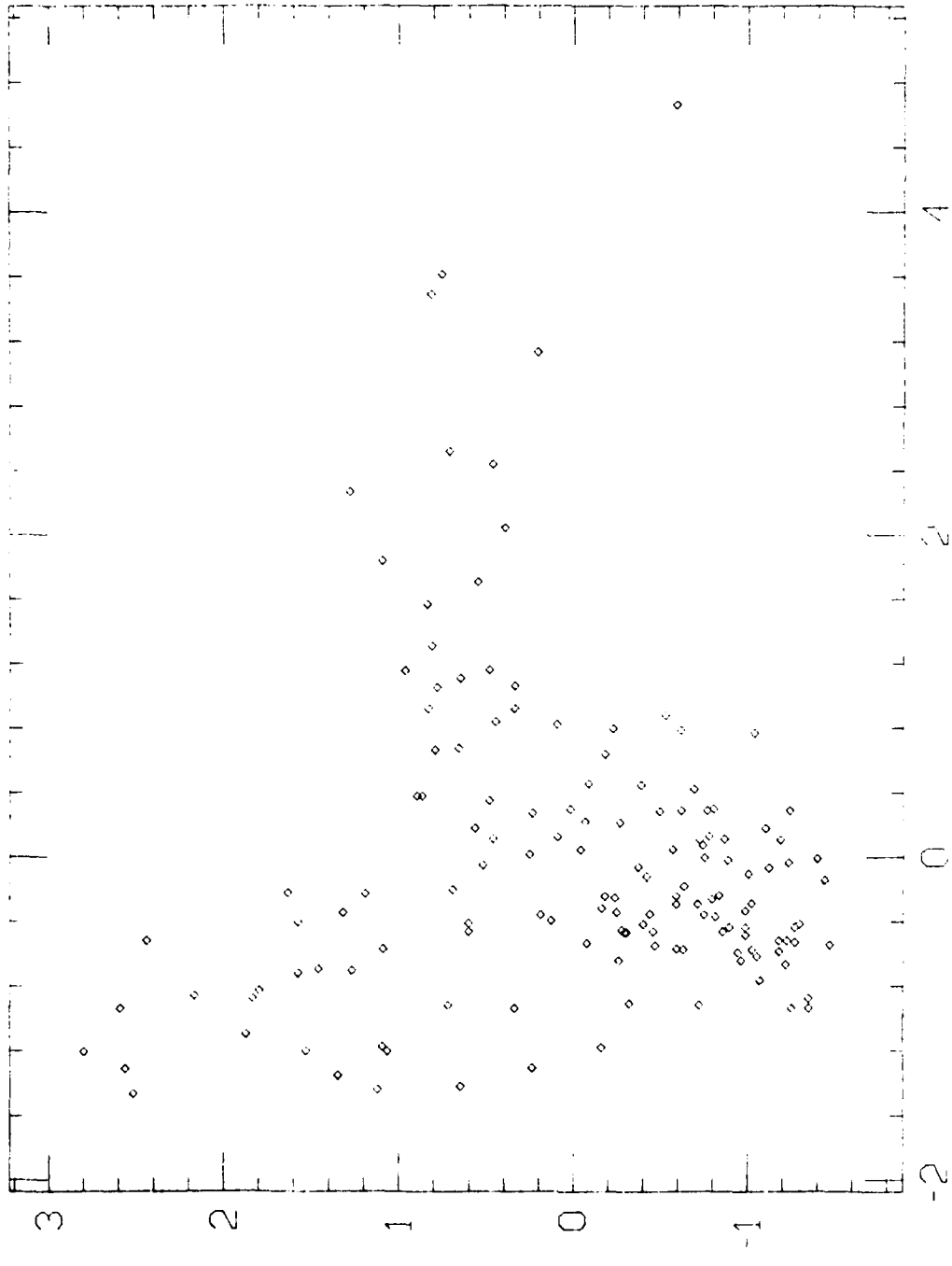


Figure 5b

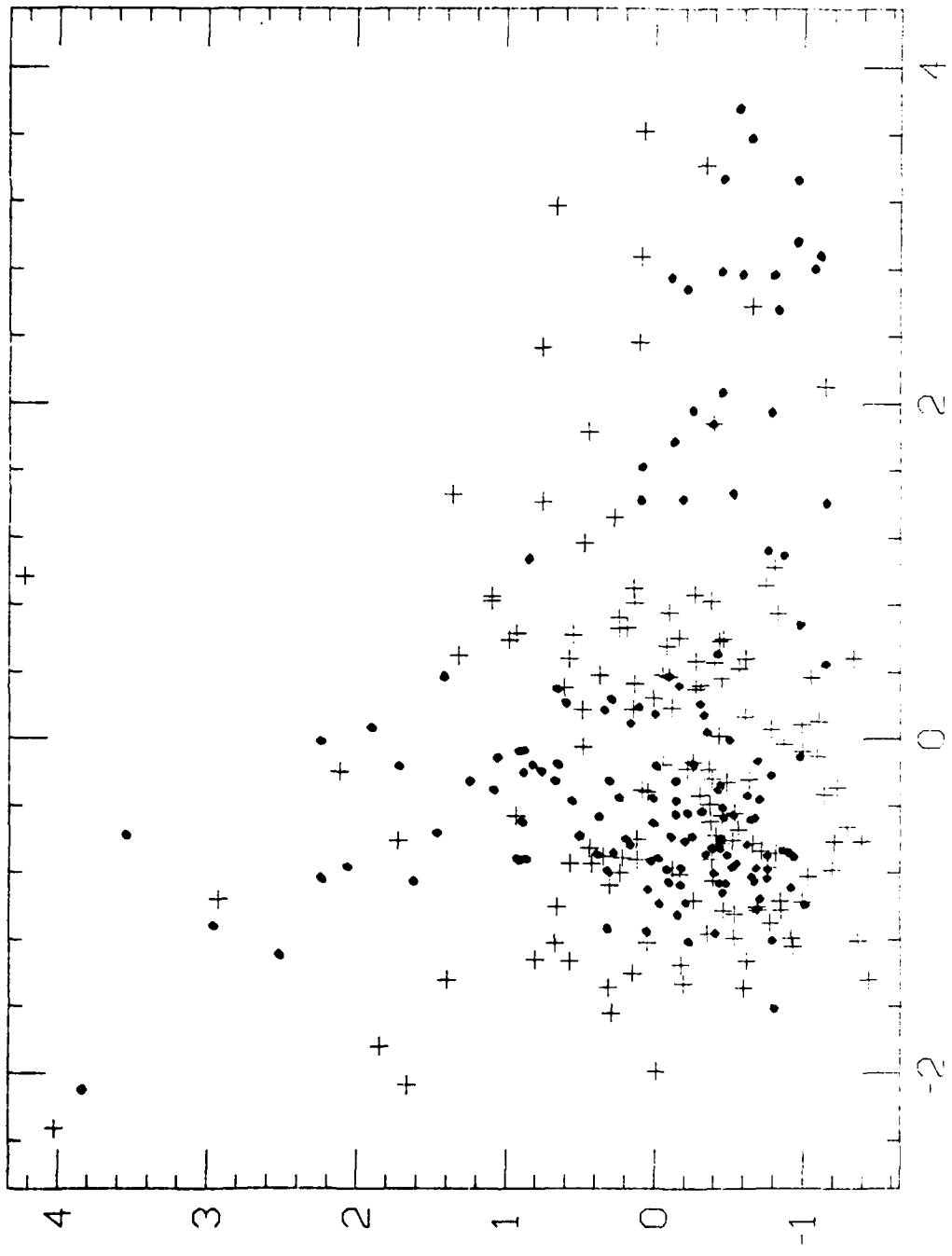


Figure 5c

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE		READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER ORION 002	2. GOVT ACCESSION NO. AD-A119 813	3. RECIPIENT'S CATALOG NUMBER
4. TITLE (and Subtitle) PROJECTION PURSUIT DENSITY ESTIMATION		5. TYPE OF REPORT & PERIOD COVERED Technical 6. PERFORMING ORG. REPORT NUMBER
7. AUTHOR(s) Jerome H. Friedman Werner Stuetzle Anne Schroeder		8. CONTRACT OR GRANT NUMBER(s) N00014-81-K-0340
9. PERFORMING ORGANIZATION NAME AND ADDRESS Department of Computer Science Stanford University Stanford, California 94305		10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS
11. CONTROLLING OFFICE NAME AND ADDRESS Department of the Navy Office of Naval Research Arlington, Virginia 22217		12. REPORT DATE July 1981 13. NUMBER OF PAGES 37
14. MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office)		15. SECURITY CLASS. (of this report) 15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release; distribution unlimited.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

The view, opinions, and/or findings contained in this report are those of the author(s) and should not be construed as an official Department of the Navy position, policy, or decision, unless so designated by other documentation.

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

The projection pursuit methodology is applied to the multivariate density estimation problem. The resulting nonparametric procedure is often less biased than kernel and near neighbor methods and does not require the specification of a metric on the data measurement space. In addition, graphical information is produced that can be used to help gain geometric insight into the multivariate data distribution.

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE

S/N 0102-LF-014-6601

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)


END
DATE
FILMED
DTIC