SLAC PUB-2193
August 1978

GRAPHICS FOR THE MULTIVARIATE TWO-SAMPLE PROBLEM*

Jerome H. Friedman
Stanford Linear Accelerator Center
Stanford, California

and

Lawrence C. Rafsky
ADP Network Services
Ann Arbor, Michigan

ABSTRACT

Some graphical methods for comparing multivariate samples are presented. These methods are based on minimal spanning tree techniques developed for multivariate two-sample tests. The utility of these methods is illustrated through examples using both real and artificial data.

(Submitted to The Journal of the American Statistical Association)

*Work partially supported by the Department of Energy

I. INTRODUCTION

Comparing two multivariate samples is a basic statistical procedure. Friedman and Rafsky (1978) (hereafter referred to as FR78) have presented several nonparametric multivariate two-sample tests sensitive to general alternatives. Such tests yield a significance level for the (null) hypothesis that the two samples are identically distributed. Whether or not the null hypothesis can be rejected, an understanding of the difference between the two samples can be helpful. Although the techniques of FR78 were discussed strictly within the framework of hypothesis testing, we show that they lend themselves to graphical representations that often provide such insight.

Graphical methods for the multivariate two-sample problem are of two types:

(1) Methods that produce planar point plots from multivariate data. These are applied to the pooled sample, and points in the resulting plot are then labeled with their original sample identity.

(2) Methods that intrinsically use sample identities in constructing the two-dimensional display.

Included in (1) are projections on pairs of coordinates as well as the generally more revealing "multidimensional scaling" techniques (see Kruskal - Wish, 1978, and references contained therein). These latter techniques -- also known as "nonlinear mapping" and "smallest space analysis" -- attempt to position the planar points so that interpoint distances are closely related (in some sense) to corresponding distances in the original multivariate space. In the present context, it is the ("zero-intercept linear") metric versions -- which try to make corresponding distances actually equal -- that would be most appropriate (see also Kruskal, 1977, p. 304-311).

Other type (1) techniques involve projection on new coordinates judiciously chosen after examining the full multivariate data set (Projection Pursuit, Friedman-Tukey, 1974), as well as familiar and generalized eigenanalysis (see e.g., Gnanadesikan, 1977, p. 5-26, 48-62). It is clear that type (1) methods, without point labeling, are also useful for examining the structure of a single multivariate sample. There are a variety of other graphic displays -- not point plots -- designed for multivariate data that, although originally intended for other purposes, could be applied to the pooled or separate samples (glyphs and variations, faces, function plots, etc.: see Gnanedesikan, 1977, especially p. 63-39, 207-255). We find them less helpful for the two-sample problem, especially with many observations, since identity labeling is less convenient and differences within-sample tend to obscure those between-sample. And even if sample differences are properly revealed, the information provided by such displays often lacks direct geometric interpretability. (See Note 2 of Section 6).

Type (2) methods include univariate two-sample graphics (e.g., P-P and Q-Q plots, Wilk and Gnanadesikan, 1968) applied to individual marginals, as well as projections onto planes defined by the Fisher linear discriminant axis separating the two samples and a second linear discriminant axis determined as Fisher's but constrained to be orthogonal to it (see Friedman-Tukey, 1974, and Sammon, 1970).

We first present a type (2) graphic technique that is a multivariate generalization of P-P plots. We also give a modification designed particularly for scale differences. We then describe a type (1) procedure with aims similar to metric multidimensional scaling (the term "planing" is convenient for any such approach). Our methods are illustrated on five examples.

II. MINIMAL SPANNING TREES

The multivariate two-sample problem begins with a sample $X_1, \ldots X_m$ from F_X and a sample $Y_1, \ldots Y_n$ from F_Y, both defined on R^p. The procedures in FR78 for testing $F_X=F_Y$ start by constructing the minimal spanning tree (MST) of the N=m+n pooled sample points. MSTs are natural objects: they are a collection of straight line segments ("edges") with sample points as endpoints, having minimum total (R^p Euclidean) distance while providing a path between every sample point pair. It is immediate that MSTs have precisely N-1 edges and do indeed form a tree, i.e., have no cycles. MSTs can be determined quite rapidly (the full interpoint distance matrix need not be computed: see Appendix, FR78) and provide compact descriptions of point sets. Formal graph theory definitions, both for MSTs and for terms used below, may be found in FR78. But graph theory is blessed with terminology so vivid that few readers will require any assistance.

III. MULTIVARIATE P-P PLOTS

The multivariate Smirnov generalization presented in FR78 involves sequencing the pooled data points in R^p (after building an MST) and applying the standard univariate Smirnov test to the resulting sequence. Two sequencing procedures were presented, one for general use, and the other for use when pure scale differences are expected. The first sequencing which generalizes the standard Smirnov test, begins by rooting the MST at an end point of a tree diameter (longest path in the tree) and then traverses the tree (roughly by following that diameter) in a special preorder. The preorder visits shallow subtrees of a given point before deeper ones. The order in which points are visited in this traversal defines the sequence. The traversal is designed so that points generally close in R^p are also close in the sequence. This is illustrated in Figure 0. Shown are 50 points in the plane, their MST, and the derived sequence. Note that the tree diameter, unique in this case, is the 19 edge path from point 1 to point 50.

Concentrations of one sample or the other in portions of such a sequence tend to represent local concentrations of corresponding points in R^p. Location shifts in R^p will thus be reflected by location shifts in the sequence, and likewise for scale. Inferences concerning the nature of the difference between the two samples can, therefore, be based on a P-P plot of the points taken in sequence order. (Observe that a P-P plot of two (univariate) samples depends only on the ranks of the m sample X and n sample Y points.)

Our multivariate P-P plot will be effective to the extent that our sequencing accurately reflects the multivariate interpoint distances. Evidence presented in FR78, and the examples below, indicate that the sequencing defined achieves its goal reasonably well. Interpretations of our P-P plot are as in the univariate case: nearly identical samples will have plots that are close to the graph of y=x, location differences will lead to plots that lie predominately above or below this line, while scale differences will be revealed by plots that tend to lie first on one side of y=x, intersect the line, and then lie on the other side for the remainder of the plot. There are, of course, a wide variety of distributional differences that lend themselves to analogous interpretations.

Our second sequencing, which generalizes a "radial" Smirnov test, ranks points on their depth in the MST when the tree is rooted at a center. (In Figure 0, points 20 and 21 are centers: points "halfway" along a diameter.) Depth (of a point) is the number of points encountered on the MST path between the point and the center root, excluding the point itself. Note that a center of an MST tends to lie near the geometric center of the multivariate point sample and, in many ways, is analogous to the median of a univariate sample. Since the edges of an MST tend to be directed along density gradients, the path between a point and an MST center will tend to follow density gradients of the pooled distribution. For spherically symmetric distributions (and their affine transformations) depth can thus be interpreted as the number of points encountered on a "steepest density descent" path from the center and is, therefore, estimating a quantity similar to (weighted) distance from the mode.

Concentrations of points from one sample in portions of this second type of sequence tend to represent concentrations at a given (weighted) distance from the center. Scale differences in R^p will appear as location differences in this sequence, and we can construct from the sequence a useful "radial" P-P plot.

Multivariate rank P-P plots based on these sequencings are obviously good adjuncts to the corresponding Smirnov tests: the tests are based on the statistic max $|F_X - F_Y|$, the maximum taken over points in sequence order. Their utility in general exploratory analysis can, of course, be established only with use. Examples using artificial and real data are presented and discussed in Section V.

IV. MST PLANING

The multivariate "runs" test presented in FR78 is based on a direct relationship between sample identities and the MST, without the intermediate stage of establishing a sequence. In order to graphically present the results of this procedure, one needs a graphical representation of the MST. It is always possible to plot the multivariate data points in two dimensions so that the N-1 MST edge lengths are exactly preserved. One can view this plot (with point labels identified) for possible interpretation.

Preserving the N-1 MST edge lengths does not completely specify the planar locations of the N observations. Each point is free to rotate in a circle about its parent at a radius equal to the (multivariate) distance from its parent. To constrain the locations of all the points, it is necessary to expand the graph so that each point is part of a cycle.

Lee, Slagel and Blum (1977) present an ingenious technique for expanding the graph and constraining the planar locations of the data points. Their procedure begins by rooting the MST arbitrarily, and sequentially "mapping" (placing in the plane) the points in the order in which they are visited in a breadth first search of the tree, starting at the root (i.e., one depth at a time; see Aho, Hopcroft, Ullman, 1974, Ch. 5). The position of each point in the plane is fixed by the requirement that the (multivariate) Euclidean distances between the point and both its parent and the root be exactly preserved in the plane. Since these three points form a triangle in p-space, their three interpoint distances specify the location of a point being mapped up to a sign ambiguity: one can place a point either to the right or left of the line joining the root and its parent, and still preserve the sides of the triangle. This ambiguity can be resolved: choose the location that minimizes the discrepancy (absolute difference between multivariate and planar distance) between the point and a third point already mapped.

Our multivariate planing procedure is a modification of this technique:

1. Root the MST at a tree center
2. Map this center, its farthest Euclidean distance daughter, and another (arbitrary) daughter into the plane preserving their three interpoint distances.
3. For each remaining daughter of the center
   Map it to preserve its distance to the center and to the farthest daughter from the center (resolve sign ambiguity with another daughter already mapped)
   end For

   Comment: This completes the mapping of the depth zero (center) and depth one (daughter) points
4. For each depth ≥ 2
   While points at this depth remain unmapped
   4.1 Find point at this depth not already mapped that is farthest from its parent

4.2 Map this point: Attempt to preserve its distance to its parent and to the farthest point already mapped (resolve sign ambiguity by minimizing discrepancy with grandparent).
<u>If</u> attempt fails
Map the point to preserve distance from parent and minimize discrepancy with farthest point.

4.3 <u>For</u> each sister of point mapped in Step 4.2
Map to preserve distance from parent and from point mapped in Step 4.2 (resolve sign ambiguity with grandparent).
<u>Comment</u>: Triangle inequality insures success

<u>end</u> <u>while</u>

<u>end</u> <u>For</u>

This mapping procedure yields a two-dimensional representation of the multivariate observations in which all (multivariate) MST edge distances are preserved, as well as distances from each point to its sister farthest from their common parent. This causes the interpoint distances between MST neighbors to be reproduced fairly well. In addition, global information is introduced by constraining one of the daughters to reproduce as accurately as possible a relatively large interpoint distance (that to the farthest point so far mapped). Starting with the tree center and mapping radially outward with increasing depth insures that many different points serve as this global (farthest so far mapped) reference point. The resulting configuration of points in the plane is intended to reflect the interpoint distance relationships in R^p. In order to gauge the degree to which this goal has been achieved, we compute and append to the plot a planing error ("absolute value stress") that is the normalized total discrepancy:

$$E = \frac{\sum_{i<j} | D_{ij} - d_{ij} |}{\sum_{i<j} D_{ij}} \tag{1}$$

where D_{ij} is the p-dimensional Euclidean distance between the ith and jth points and d_{ij} is their distance in the plane. If the multivariate observations all lie in a two-dimensional linear subspace of R^p, E has the value zero and our planing procedure gives a perfect representation.

While in no way attempting to systematically minimize (1), our planing technique nonetheless competes, in some sense, with procedures that do minimize similar stress quantities (usually a function of squared discrepancies to insure differentiability). These procedures, employing gradient techniques, or in the case of "zero-intercept linear" metric scaling, a standard eigenanalysis, are remarkably effective but somewhat time-consuming. MST-based planing, on the other hand, is exceptionally fast: planing 10,000 points is a perfectly reasonable task.

It is apparent, however, from examining the recommended procedure that MST-planing does a better job of matching small interpoint distances than larger ones. We may conclude, therefore, that MST-planing is even more competitive with those multidimensional scaling approaches that minimize "weighted stress" -- larger distances getting smaller weight than small ones. This behavior might well be fortuitous: giving less weight to large distances allows the planar placement to concentrate on local information with the well known result that the final display is more responsive to general nonlinear singularities. But it must be admitted that more powerful (although costly) "parametric mapping" techniques exist for uncovering such structure (see Shepard-Carrol, 1966, and Gnanadesikan, 1977, p. 35-48).

Multidimensional scaling procedures can produce, moreover, minimum-stress point placements in higher dimensions and are, therefore, used to uncover well-fitting lower-than-p-dimensional configurations ($p>2$). (They can also be used to find any dimensional metric representations when the interpoint distances are given directly and not formed originally from N points in p-space - for instance, subjectively determined dissimilarities. But in this case, non-metric techniques, requiring only that fitted and original distances be monotonically related, might be more appropriate). There is no bar to extending the sequential mapping technique of planing to three or more dimensions: full dimensional simplexes would play the role of triangles. In R^3, for instance, a point could be mapped preserving an MST edge length and the distances to two points previously mapped. We have not yet implemented this idea, however, and detailed experimentation remains to be done.

Nonetheless, our MST-planing serves well in its intended role: a graphical adjunct to the multivariate Wald-Wolfowitz test of FR78. The test statistic used is the number of MST edges linking points from the X sample to points from the Y sample, and these edges can be directly shown on the planing. For nearly identical samples, these edges will obscure the plot, but for differing samples there should be few enough edges to be manageable. (The edges could themselves be sampled and fewer shown). They will assist the eye in picking out regions of overlap, and will also highlight stray X's in a dense Y region (and vice-versa). This highlighting is particularly important when color is not available for distinguishing the sample identities of the plotted points. When these edges, moreover, link points that do not appear close in the plane, they reveal distortion and local contributions to overall planing error.

V. EXAMPLES

To evaluate the utility of these procedures and gain insight into their properties, we apply them to four data sets.

The first two are artificially generated so that the graphics can be judged in light of known multivariate structure. The third data set is the well known Iris data of E. Anderson (Fisher, 1936). The fourth was collected in a particle physics experiment.

The data of Figure 1 are two samples of 100 observations from ten-dimensional product double exponental distributions with unit scale parameter, separated in location by a distance of three units. Figure 1a displays the "standard" multivariate P-P plot, Figure 1b the "radial" P-P plot. These plots clearly suggest that the two samples differ primarily in location.

Figure 1c shows our planing of these data with sample identities indicated. The value E = 0.53 indicates that there is considerable distortion of the p-dimensional interpoint distances. This is not surprising since the data are fully ten-dimensional by construction. In spite of this, one can still observe that the first sample (indicated by small squares) predominately populates the lower right of the plot, while the second sample (indicated by small crosses) dominates the upper left, leading to the inference that the two samples differ in location.

Figure 1d shows the same plot along with (multivariate) MST edges that connect points from different samples. This gives an indication of those observations from one sample that lie near to or in regions dominated by the other sample.

The data of Figure 2 are similarly ten-dimensional product double exponential samples; here the scale of one is three times that of the other. The multivariate P-P plots and the planing reflect this quite well. Comparing Figures 2c and 2d, we note that adding the X-Y edges helps the eye see clearly the outer ring of X's.

We now consider the Iris data, 50 observations in four dimensions on each of three species. Figure 3 is a planing of the full data set with the three species identified after plotting. It is instructive to compare this plot to the CRIMCOORDS projection which uses a three-group discrimination to define the display coordinates (Gnanadesikan, 1977, p. 221). One should also compare the Projection Pursuit display (Friedman-Tukey, 1974, p. 886) and Sammon's "nonlinear map" (a type of weighted-stress minimization, Sammon, 1969, p. 403). Lee, Slagle, and Blum (1977, p. 292) reproduce their MST-based planing of the Iris data, but they have apparently not used the entire data set.

Figure 4 focuses on the two similar species (Versicolor and Virginica). Here one sees that the two samples approach each other quite closely but overlap very little. Moreover, the Iris Virginica measurements appear to be more spread out in 4-space and have a somewhat different shape than Iris Versicolor.

The data in our final example are measurements of the energy deposited by particles as they pass through a particle detector comprised of crystals of sodium iodide (Richardson, 1978). The amount of energy absorbed by the 13 crystals closest to where a particle passed is recorded. The object is to try to distinguish between various types of particles that pass through the apparatus, based on the relative values of these 13 energies. In particular, one wishes to know whether the apparatus can be used to distinguish between electrons and negative pi mesons.

In order to test this, data for 200 electrons (first sample, represented by squares) and 200 pi mesons (second sample, represented by crosses) are examined. Figure 5a indicates a strong location difference between these two samples in the 13-dimensional space. Figure 5b indicates that there may also be a substantial scale difference. Figures 5c and 5d show a somewhat complicated configuration. The electrons are all concentrated near the origin while the pi mesons tend to have two components, one also concentrated near the origin and the other small component with larger scale distributed over the extent of the plot. Figures 5e and 5f magnify the central region near the origin. Here one sees a rather complex structure which is not easily categorized as either a location or scale difference but is probably closer to a location difference.

The displays of Figure 5 indicate a considerable difference between the distributions of the electrons and negative pi mesons in the 13-dimensional observation space. The rather low planing error (E = 0.22) indicates that the two-dimensional plot does not grossly misrepresent the relationships between the points in R^{13}. One can conclude that the apparatus could successfully differentiate electrons from pi mesons but that a simple decision rule, such as a linear discriminant, would be inadequate. Nearest neighbor or piecewise linear discrimination might be more effective. (The recursive partitioning decision rule of Friedman (1977) was able to achieve an error rate of 2%).

The examples of this section indicate that our graphical methods can be useful. They present to the data analyst more information than the value of a test statistic and can serve as a useful adjunct to two-sample testing. FORTRAN programs implementing these ideas are available from either author.

VI. NOTES

1. The methods suffer from some limitations that should be kept in mind when interpreting results. They are based on interpoint distance relationships in the pooled sample and, in particular, depend upon the order of the sorted N(N-1)/2 interpoint distances (or general dissimilarities). As discussed in FR78, our methods are resistant to moderate changes in this order but are not fully robust. Moreover, equality of two distributions in a p-dimensional coordinate space is uneffected by arbitrary strictly monotone transformations on any coordinate. But procedures based on interpoint distances are invariant only to global

shifts (translating each point by the same amount) and global expansion or contraction (scaling each coordinate by the same amount). Best choice of relative coordinate scaling, or perhaps a more complicated transformation, depends on the particular situation.

2. It is clear that any type (1) graphic technique can be applied as well to the k-sample problem: the planar display is constructed from the pooled sample and then labeled with sample identities. We did just that, in fact, when we planed the Iris data. We observed that it was helpful to proceed stagewise: use the planing to identify a well separated sample, remove it, plane again, etc. Undoubtedly this is a good procedure generally; the speed of planing allows for many repetitions (and, therefore, also facilitates error analysis through cross-validation or subsampling). The stagewise approach would be particularly useful when clustering, i.e., using planing on a single sample to uncover clusters of observations. If we had no knowledge of sample identities in the Iris data, stagewise planing and visual examination would have recovered the Setosa group perfectly -- as any reasonable procedure must -- and would (depending on individual judgment) most likely identify two more clusters: one almost purely of Virginica and the other about a 30% - 70% mix of Virginica and Versicolor. One could, moreover, proceed by separately planing the two groups resulting from the first stage, then plane the resulting four, etc., similar to the recursive partitioning approach of Friedman (1977).

Our earlier comments on the lack of geometric interpretability for the displays that are not point plots apply equally well to the k-sample or the clustering situations. Mezzich and Worthington, 1978, ran an experiment asking thirteen subjects to view, individually, the same data set displayed using linear and circular profiles, Andrews curves, principal component projection, Chernoff faces, and nonmetric multidimensional scaling. The data set consisted of four groups, each containing eleven observations in 17-dimensional space. The group identities of the observations were not shown; subjects were asked to uncover the clustering structure. An analysis of their responses demonstrates emphatically that only the multidimensional scaling display allowed an accurate reconstruction and --probably because of its familiar geometric interpretability-- was rated by the subjects as easy to use. (Principal component projection --also a point plot-- was ranked second in ease-of-use score; the others were far behind.) It is dangerous, however, to draw general conclusions from the experiment. Several design defects weaken its verisimilitude: (1) the four groups were very well separated, (2) the variables were all identically scaled ordered categories, (3) the subjects were told to look precisely for four clusters consisting of eleven observations, and (4) the subjects were untrained in statistics data analysis.

3. The sequencings used to construct the P-P plots (and the Kolmogorov-Smirnov statistics of FR78) depend on a choice of center and diameter. (Similarly, planing depends on a choice of center.) Since we choose arbitrarily, the sensitivity of our approach to these choices might well be questioned. While altering, say the choice of diameter changes the sequencing and the resulting plot, a variety of simulations have convinced us that our techniques possess "impression robustness" --i.e., the data analytic conclusions we draw from these displays are almost always unaffected.

4. It is often held that Q-Q plots, which reflect (univariate) location differences by non-zero intercepts, scale differences by non-unit slopes, and more general differences by deviations from straightness, are data analytic tools superior to P-P plots. There is some interest, therefore, in multivariate generalizations of the Q-Q plot. Our sequencings could yield such a generalization directly but, since Q-Q plots depend on the actual values of the order statistics in each sample, we would have to depend on actual distances along the sequence instead of merely the ordering it induces. These distances are unreliable since most multivariate data produce an enormous stress when forced to lie in R^1 (stress defined by analogy with (1)).

It is possible to define and generalize a "rank" Q-Q plot: replace in the univariate case every data value by its rank in the pooled sample, and then construct an ordinary Q-Q. It is clear that this is equivalent to applying a monotone transformation ($w \to F_{xy}(w)$, where F_{xy} is the empirical cdf of the pooled sample) to each axis. Examples have convinced us, however, that the "stretching" involved destroys all interpretability.

5. It is important to realize that our planing procedure is the result of heuristic arguments, Gedanken experiments, and extensive simulations. We have attempted to strengthen the Lee-Slagel-Blum procedure in a number of ways, trying to arrive at the most reasonable tradeoff of planer placement optimality and computing effort, but the end product is only one of many plausible fast approximations to (zero-intercept linear) metric multidimensional scaling. An alternative that we abandoned introduces more global information by preserving an MST edge distance and the distance to the farthest point already mapped for <u>every</u> point, not just for

one among every group of sisters. We found the decrease in planing error achieved, did not warrant the additional computing required.

ACKNOWLEDGMENTS

We thank William S. Cleveland, Joseph Kruskal, John Tukey, Paul Tukey, and a referee for helpful comments.

REFERENCES

Aho, A., Hopcroft, J. and Ullman, J. (1974), The design and analysis of computer algorithms, Addison-Wesley.

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

Friedman, J.H. (1977), A recursive partitioning decision rule for nonparametric classification. IEEE Trans. Comp., C-26, 404-408.

Friedman, J.H. and Rafsky, L.C. (1978), Multivariate generalizations of the Wald-Wolfowitz and Smirnov two-sample tests. Stanford Linear Accelerator Center, Stanford University, Report SLAC-PUB-2116, (to appear in Annals of Statistics, 1979).

Friedman, J.H. and Tukey, J.W. (1974), A projection pursuit algorithm for exploratory data analysis. IEEE Trans. Comp., C-23, 881-890.

Gnanadesikan, R. (1977), Methods for statistical data analysis of multivariate observations, John Wiley & Sons.

Kruskal, J.B. and Carroll, J.D. (1969), Geometric models and badness-of-fit functions. In Multivariate Analysis II (P.R. Krishnaiah, ed.), 639-670, Academic Press.

Kruskal, J.B. (1976), "Multidimensional scaling and other methods for discovering structure." In Statistical Methods for Digital Computers (K. Enstein, A. Ralston, H. Wilf, eds.), John Wiley & Sons.

Kruskal, J.B. and Wish, M. (1978), Multidimensonal Scaling, Sage Publications.

Lee, R.C.T., Slagle, J.R., and Blum, H. (1977), A triangulation method for the sequential mapping of points from N-space to two-space. IEEE Trans. Comp., C-26, 228-292.

Mezzich, J.E. and Worthington, D.R.L. (1978), A comparison of graphical methods of multidimensional psychiatric diagnostic data. In Graphical Representation of Multivariate Data (P.C.C. Wang, ed.), 123-141, Academic Press.

Richardson, M. (1978), Pion-electron separation with the crystal ball. Stanford Linear Accelerator Center, Stanford University, Crystal Ball Note CB308 (unpublished).

Sammon, J.W. (1969), A nonlinear mapping for data structure analysis. IEEE Trans. Comp., C-18, 401-409.

Sammon, J.W. (1970), An optimal discriminant plane. IEEE Trans. Comp., C-19, 826-829.

Shepard, R.N. and Carroll, J.D. (1966), Parametric representation of nonlinear data structures. In Multivariate Analysis (P.R. Krishnaiah, ed.), 561-592, Academic Press.

Wilk, M.D. and Gnanadesikan, R. (1968), Probability plotting methods for the analysis of data. Biometrika, 55, 1-17.

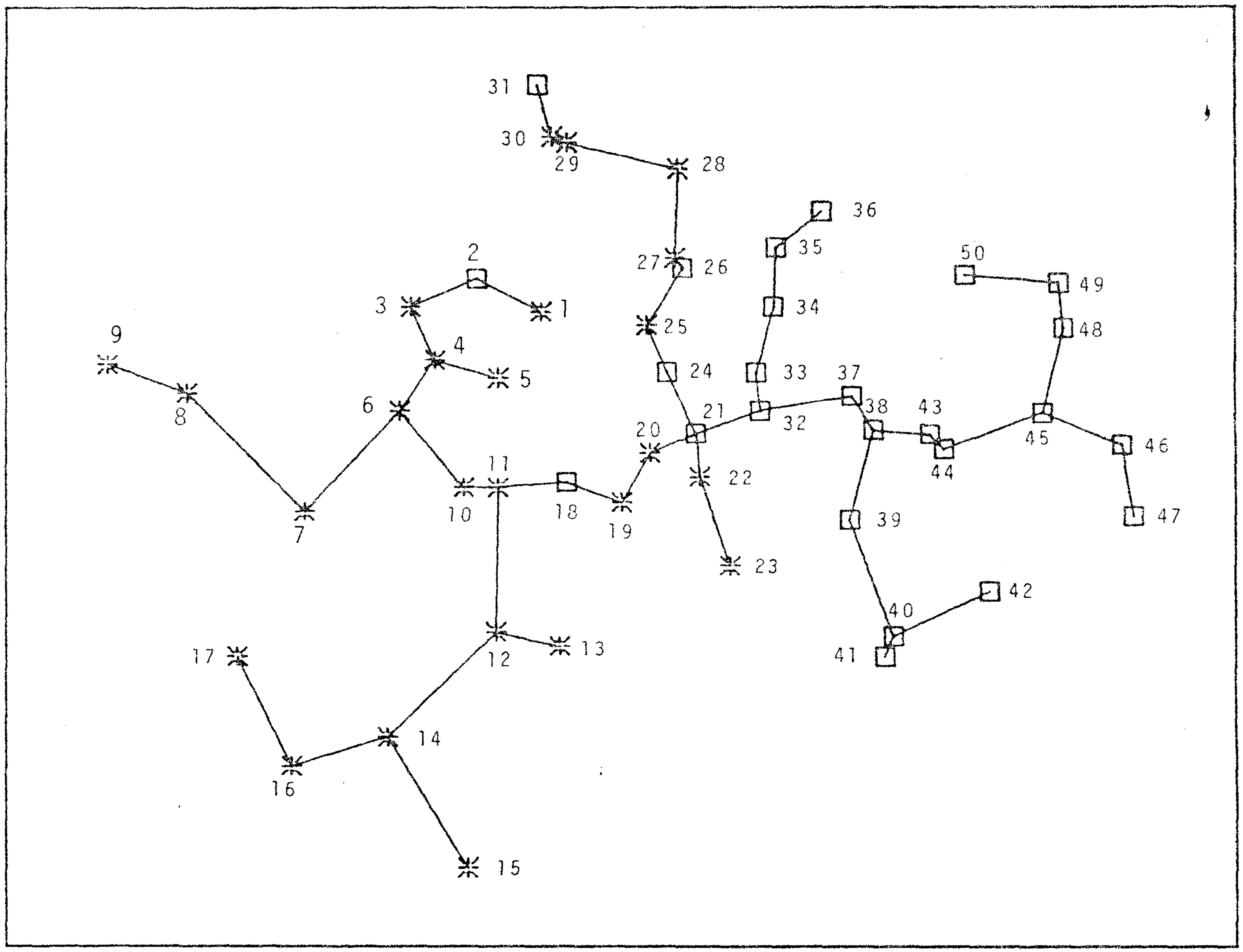

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50

FIGURE 0

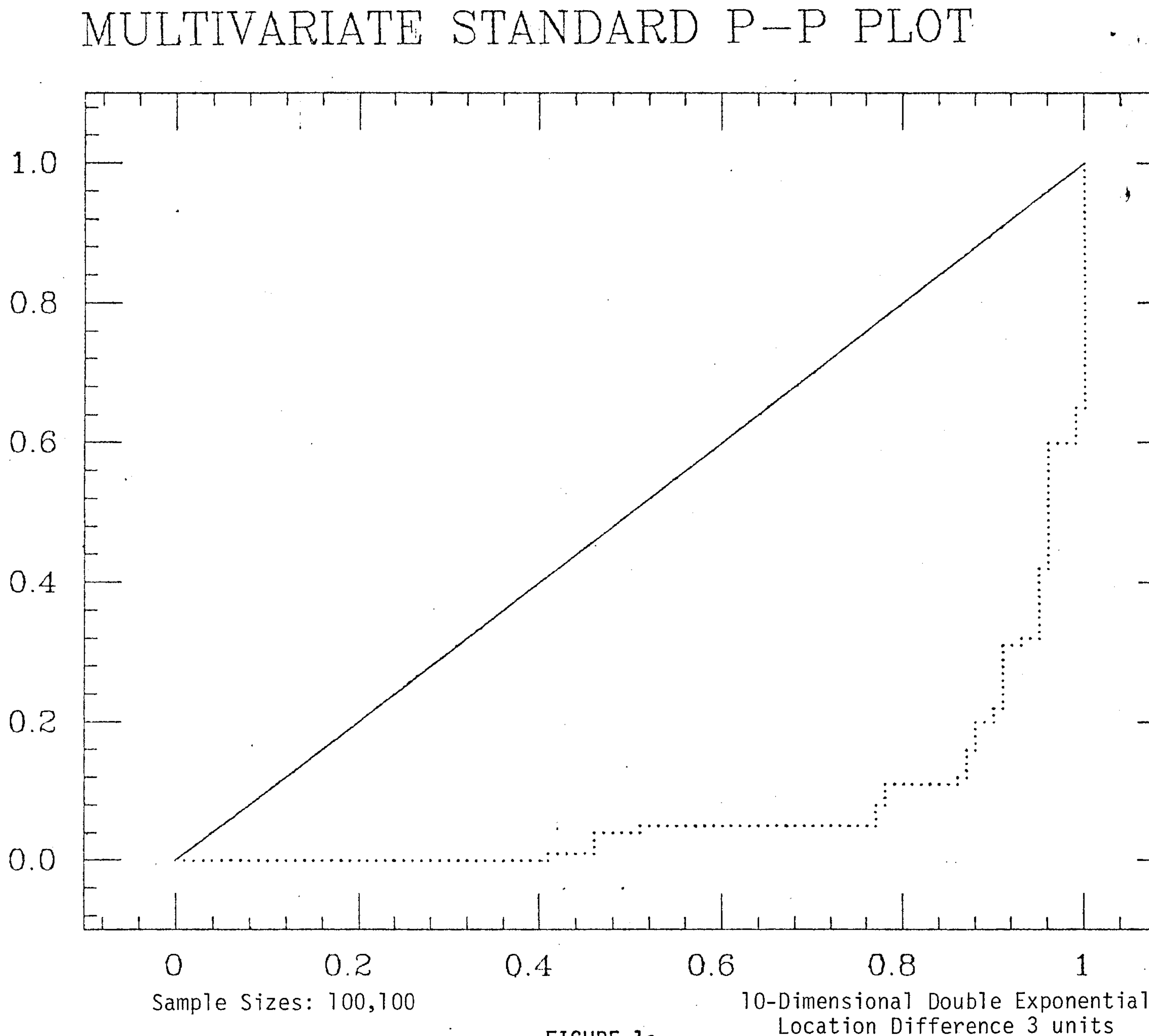
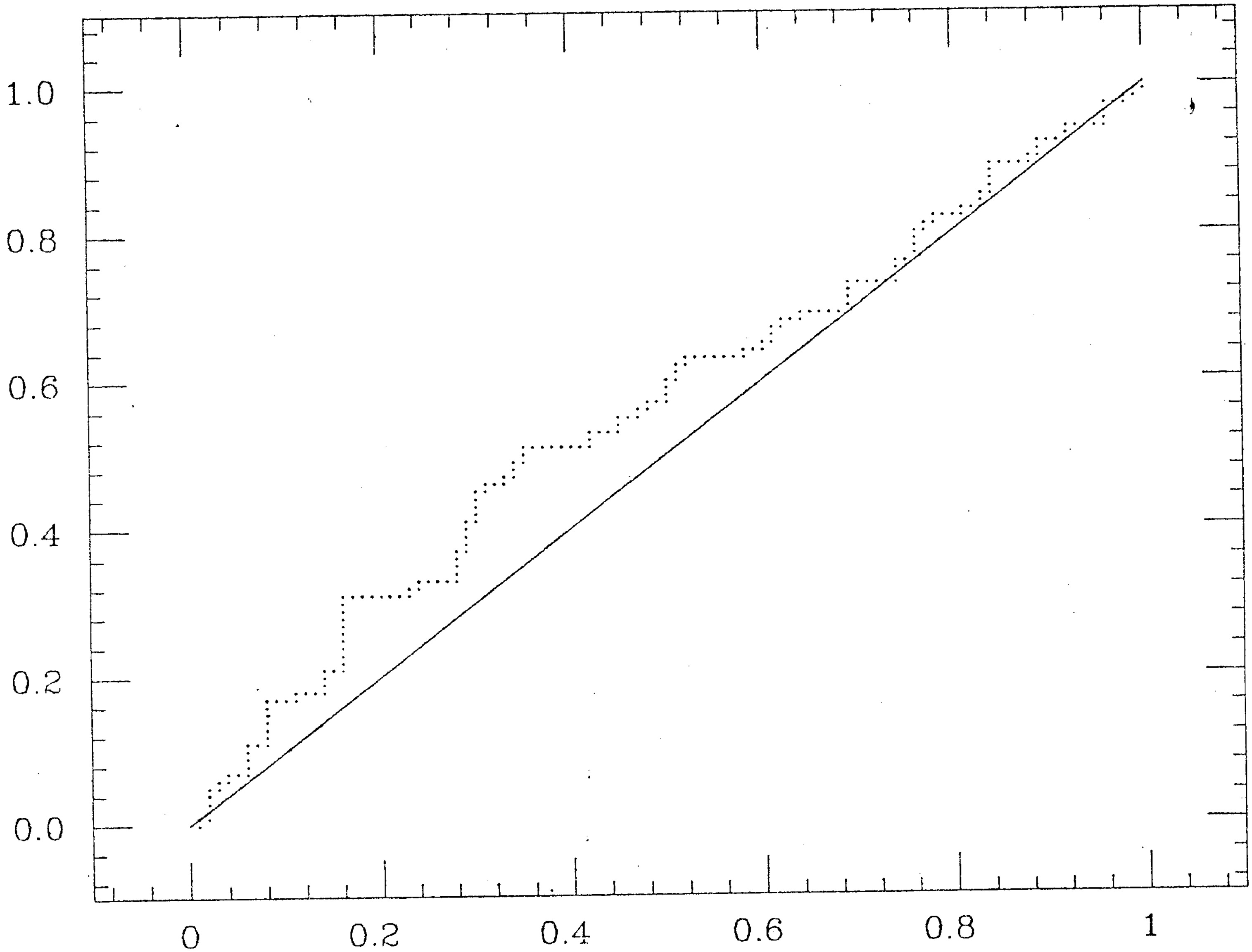

MULTIVARIATE STANDARD P-P PLOT
1.0
0.8
0.6
0.4
0.2
0.0
0
0.2
0.4
0.6
0.8
1
Sample Sizes: 100,100
10-Dimensional Double Exponential
Location Difference 3 units

FIGURE 1a

MULTIVARIATE RADIAL P-P PLOT

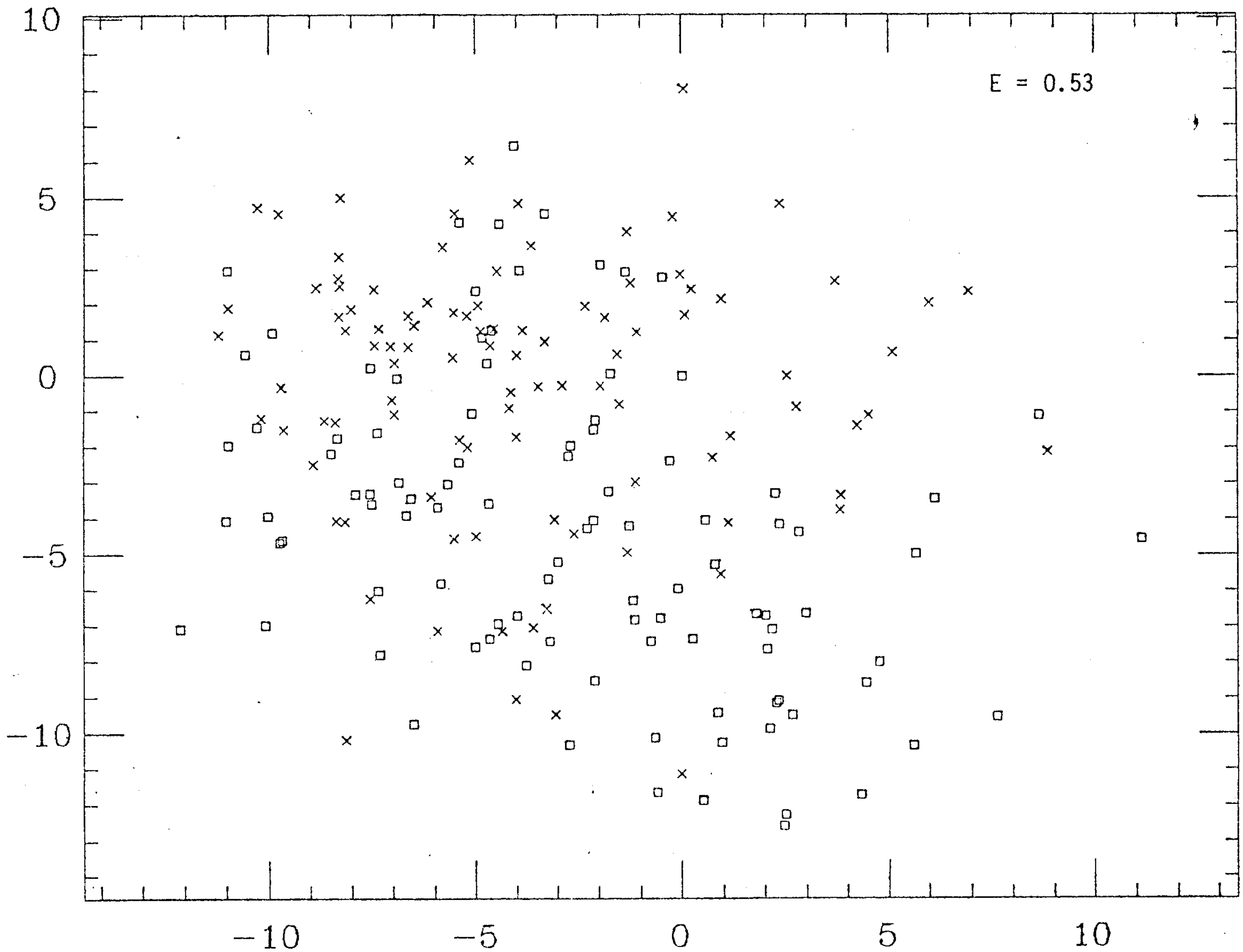


Sample Sizes: 100,100

10-dimensional Double Exponential
Location Difference 3 units

FIGURE 1b

MULTIVARIATE PLANAR REPRESENTATION

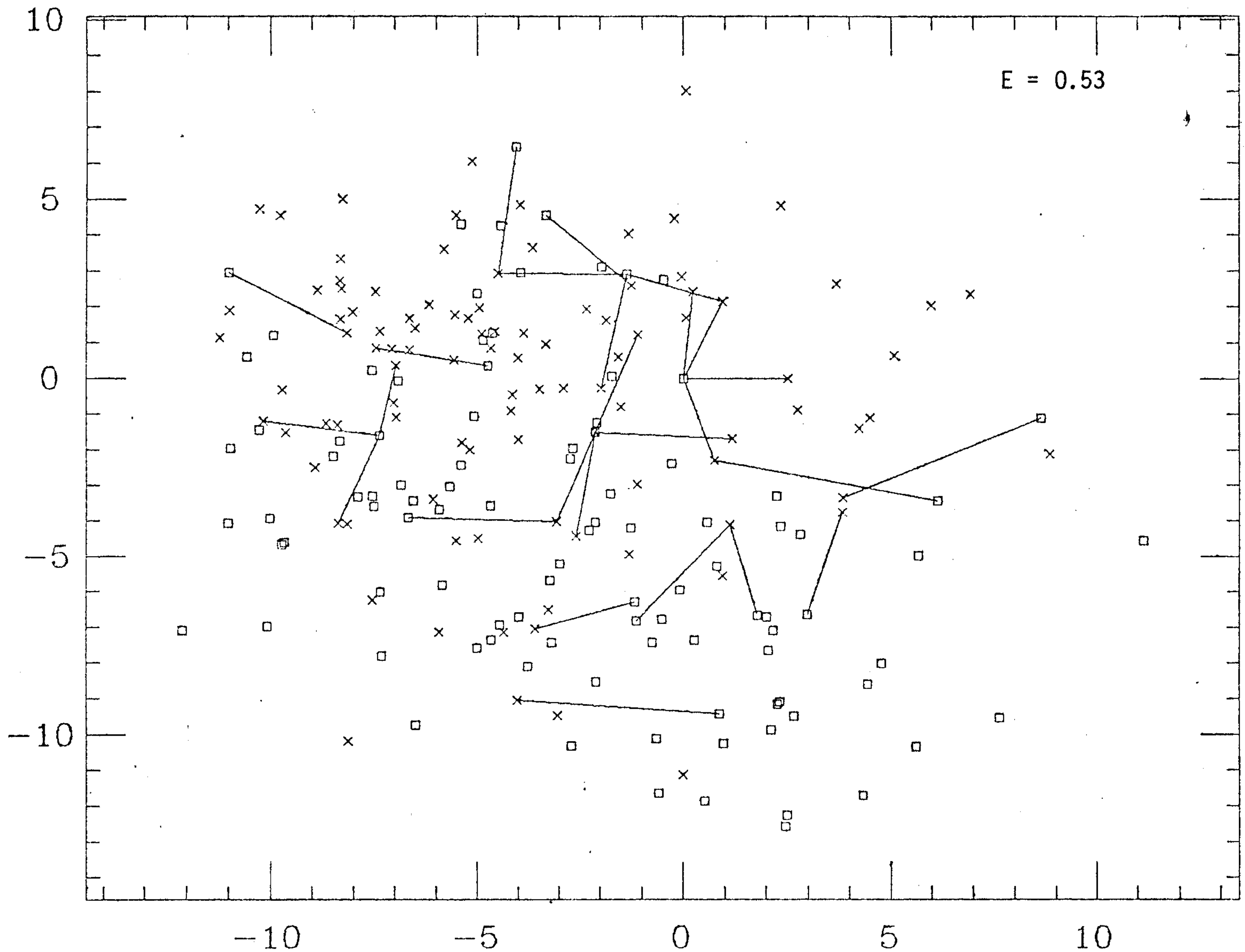


Sample Sizes: 100,100

10-dimensional Double Exponential
Location Difference 3 units

FIGURE 1c

MULTIVARIATE PLANAR REPRESENTATION



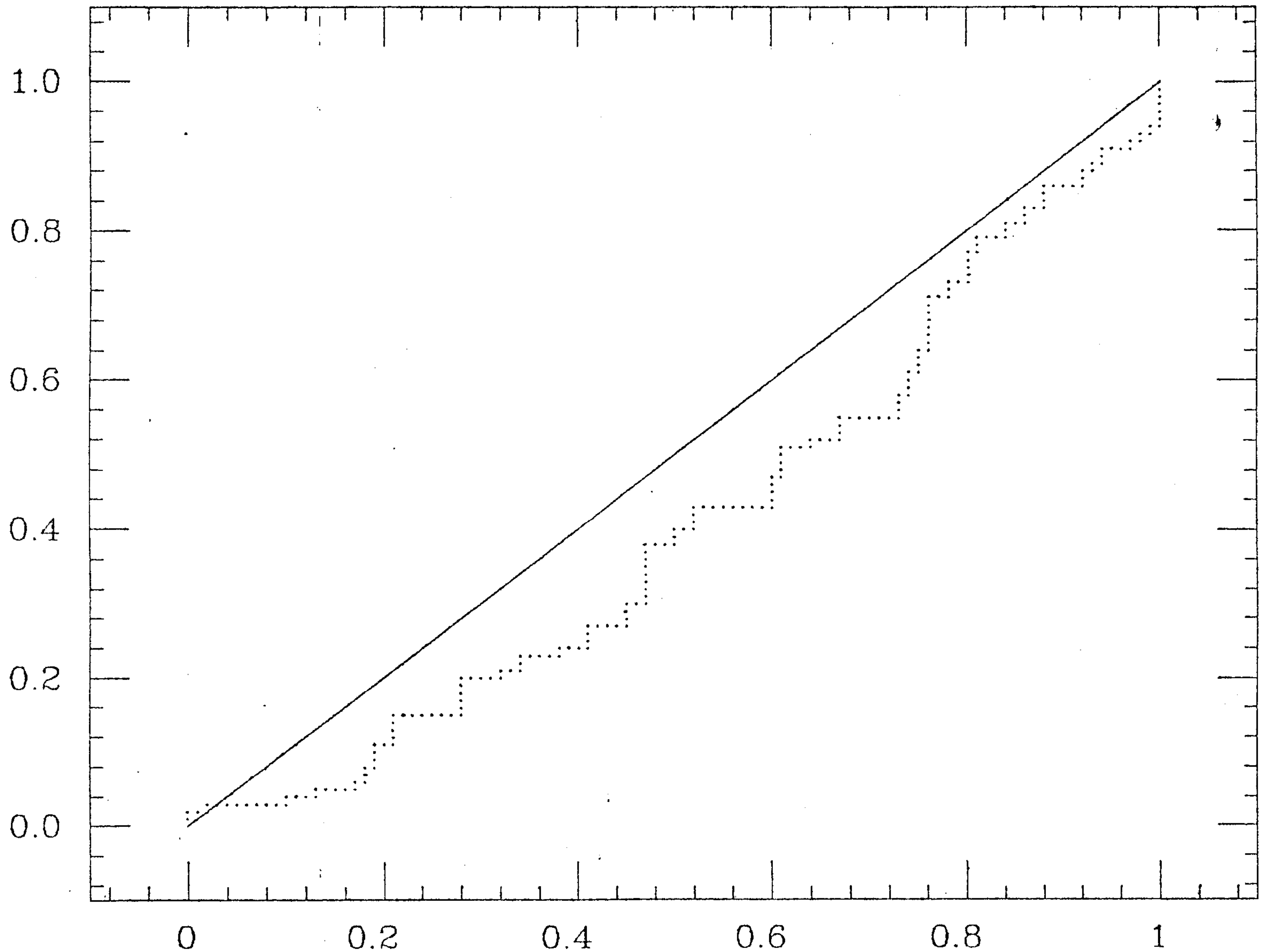
Sample Sizes: 100,100

X - Y MST EDGES

10-dimensional Double Exponential
Location Difference 3 units

FIGURE 1d

MULTIVARIATE STANDARD P-P PLOT

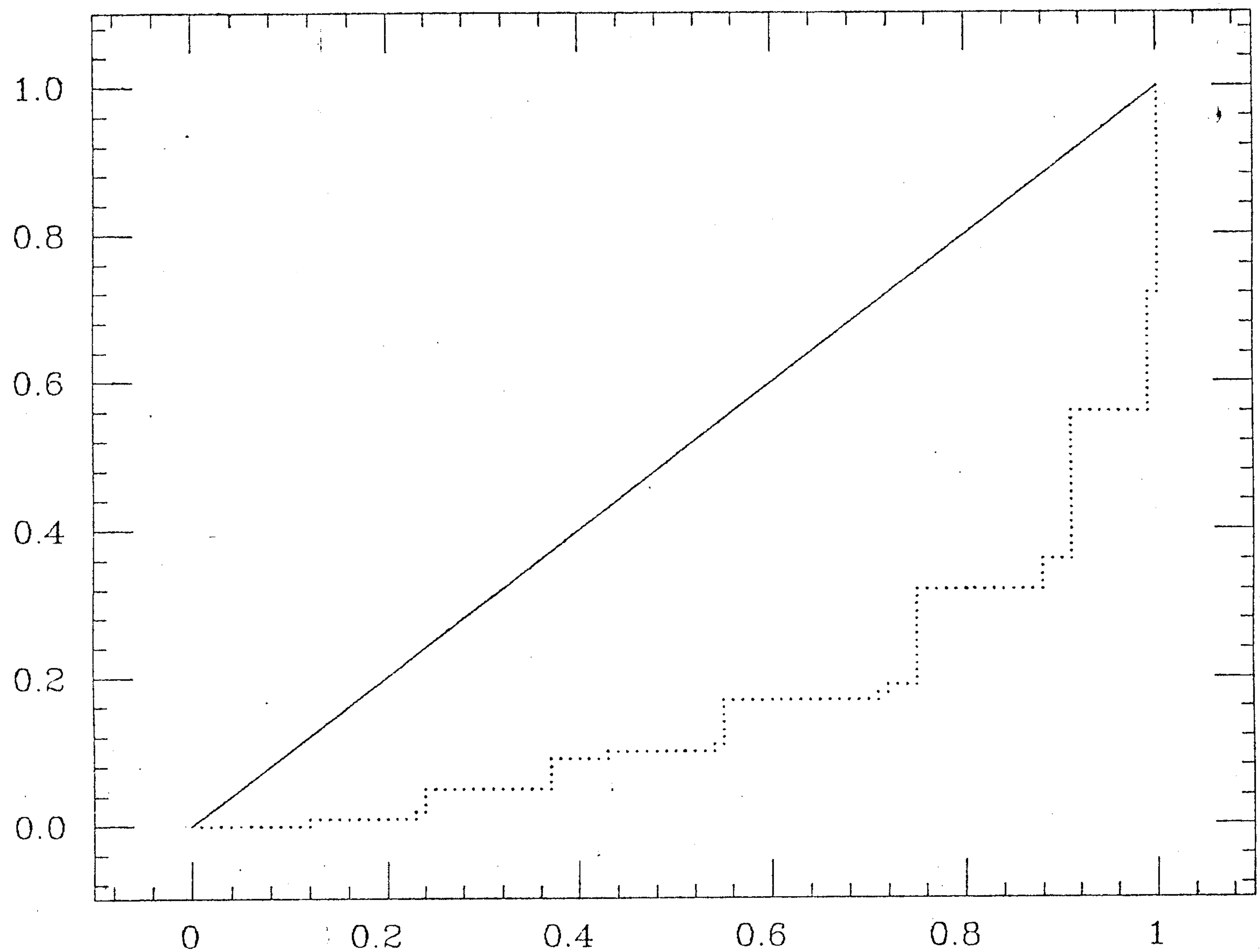


Sample Sizes: 100,100

10-dimensional double exponential
scale difference: 3 units

FIGURE 2a

MULTIVARIATE RADIAL P-P PLOT

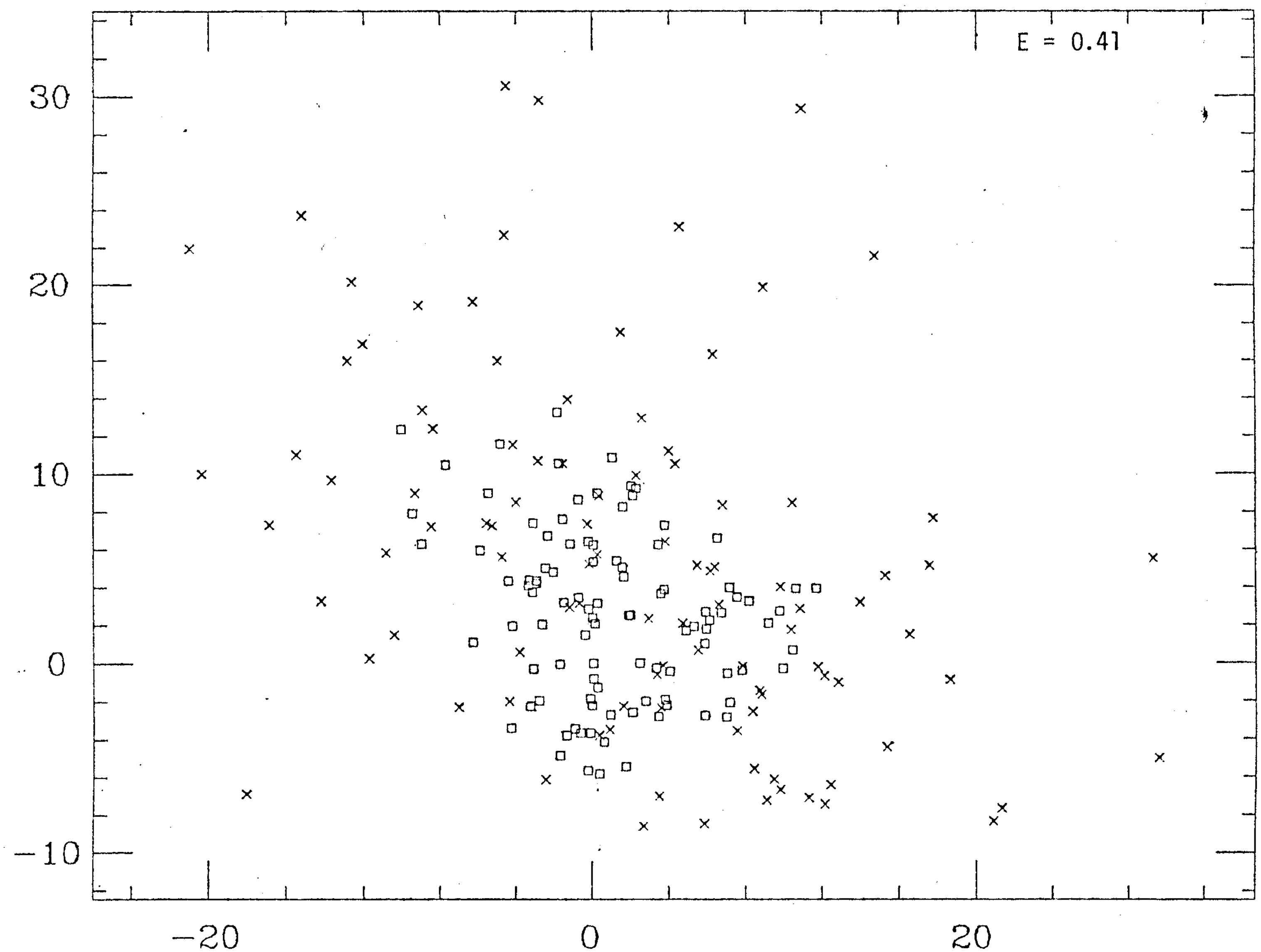


Sample Sizes: 100,100

10-dimensional double exponential
scale difference: 3 units

FIGURE 2b

MULTIVARIATE PLANAR REPRESENTATION

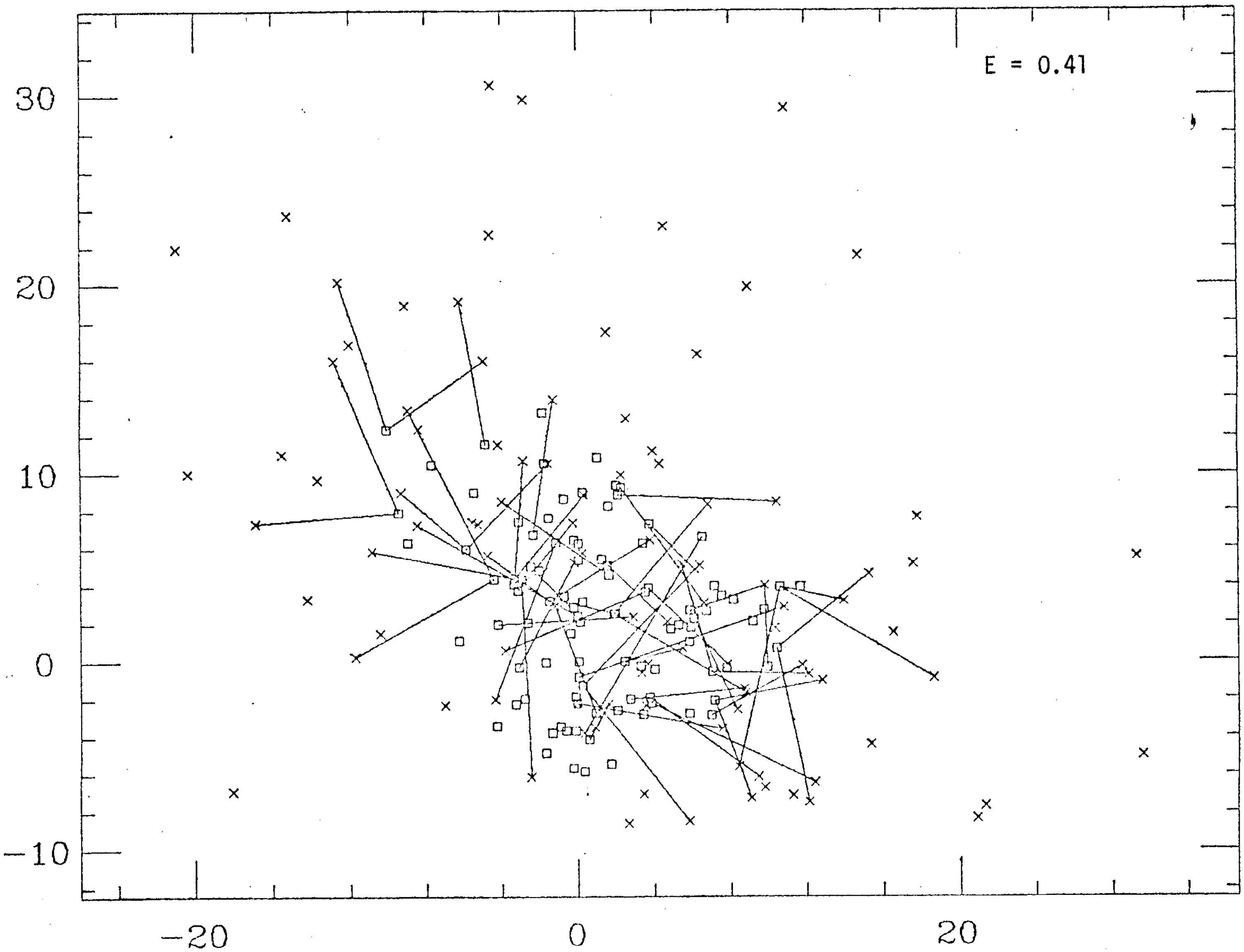


Sample Sizes: 100,100

10-dimensional double exponential
scale difference: 3 units

FIGURE 2c

MULTIVARIATE PLANAR REPRESENTATION



Sample Sizes: 100,100

X - Y MST EDGES

10-dimensional double exponential
scale difference: 3 units

FIGURE 2d

MULTIVARIATE PLANING

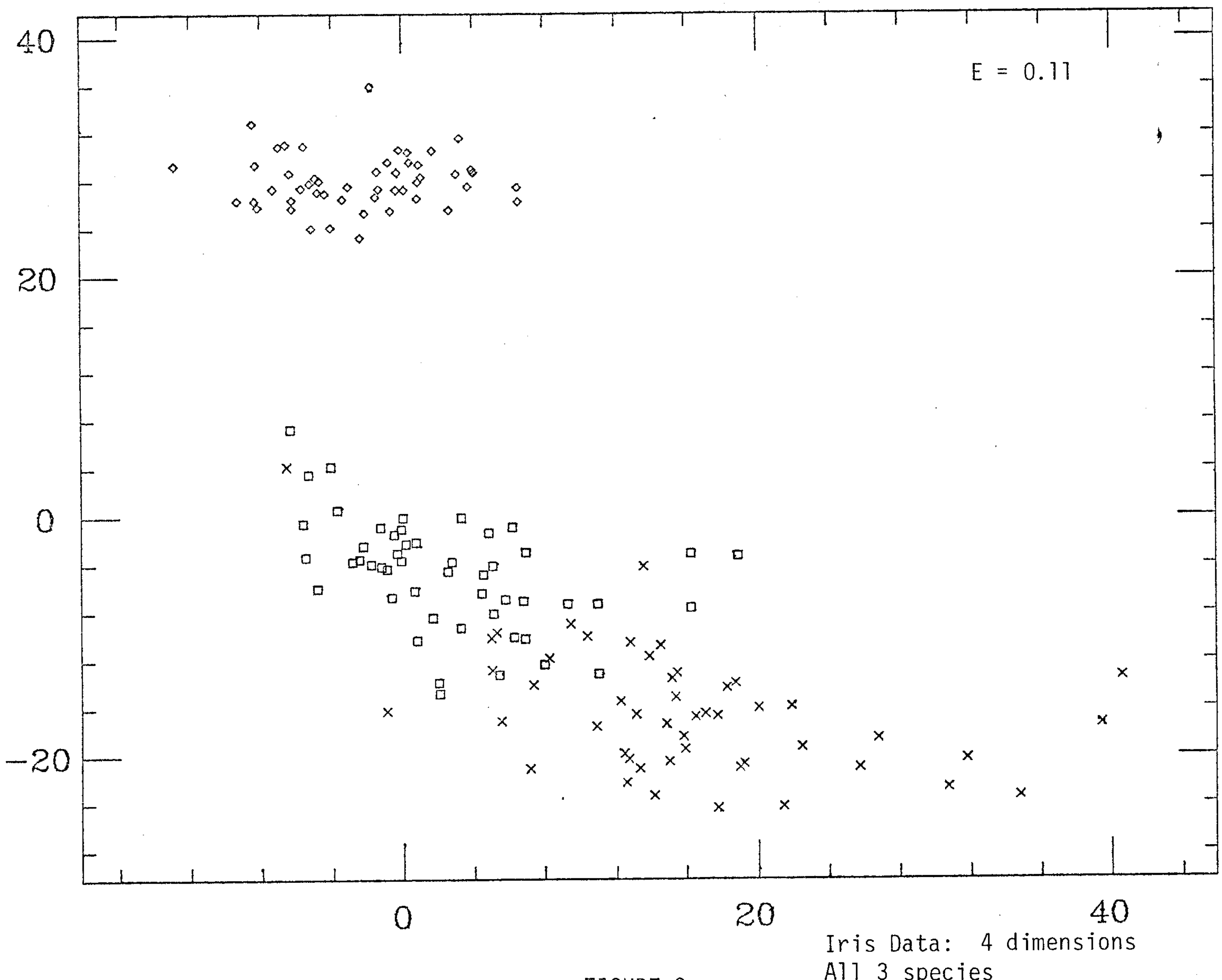
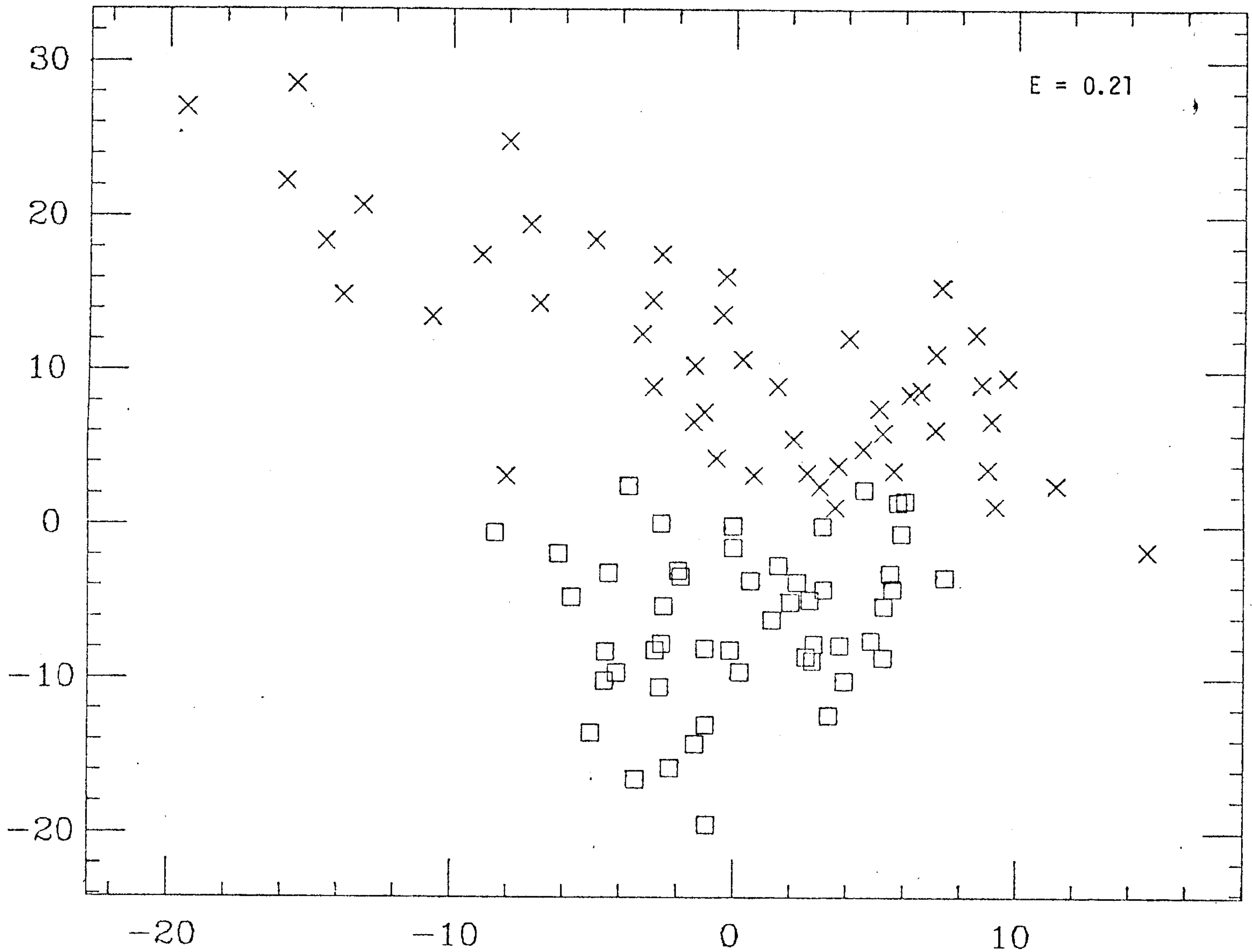


FIGURE 3

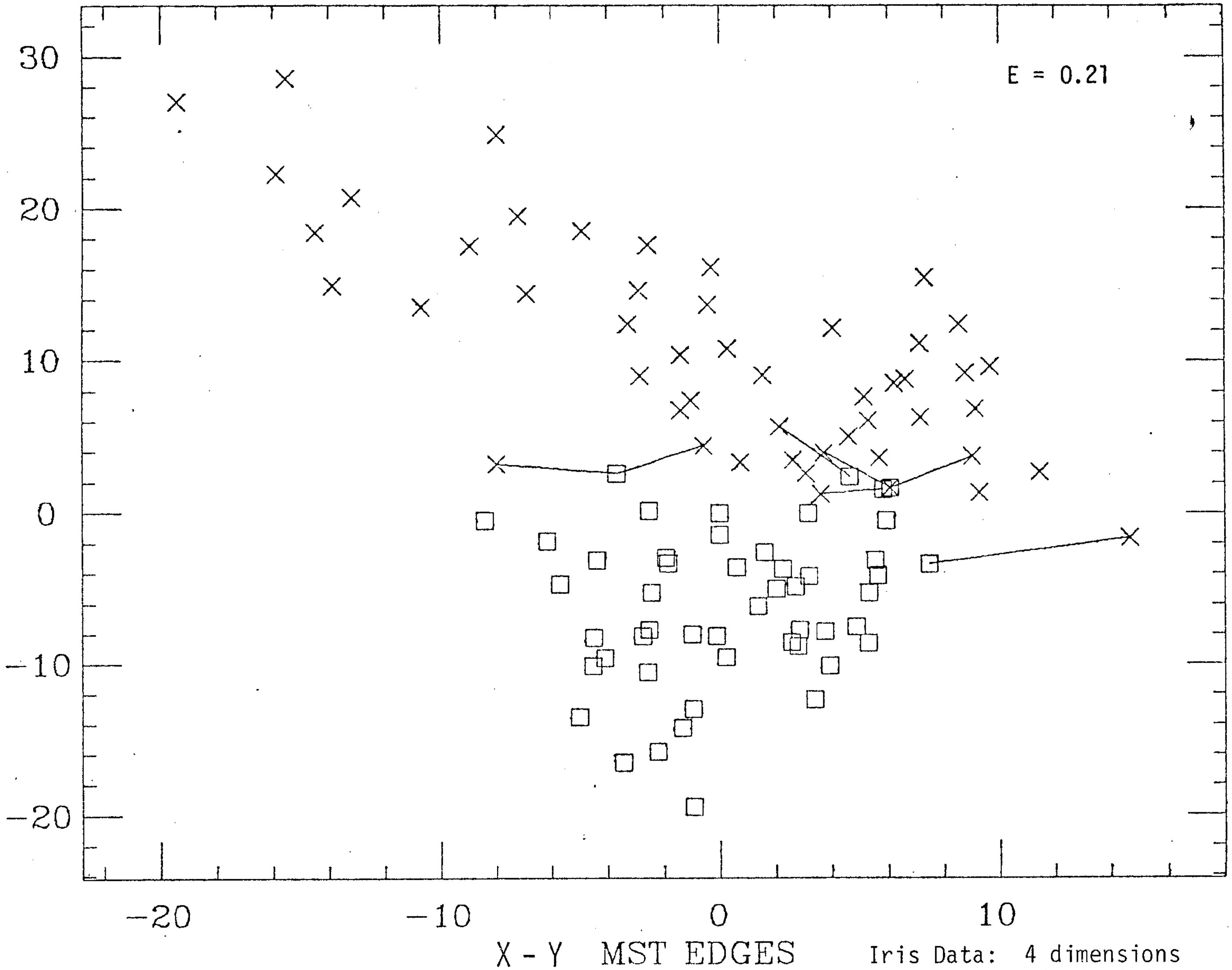
MULTIVARIATE PLANAR REPRESENTATION



Iris Data: 4 dimensions
Versicolor and Virginica

FIGURE 4a

MULTIVARIATE PLANAR REPRESENTATION



Iris Data: 4 dimensions
Versicolor and Virginica

FIGURE 4b

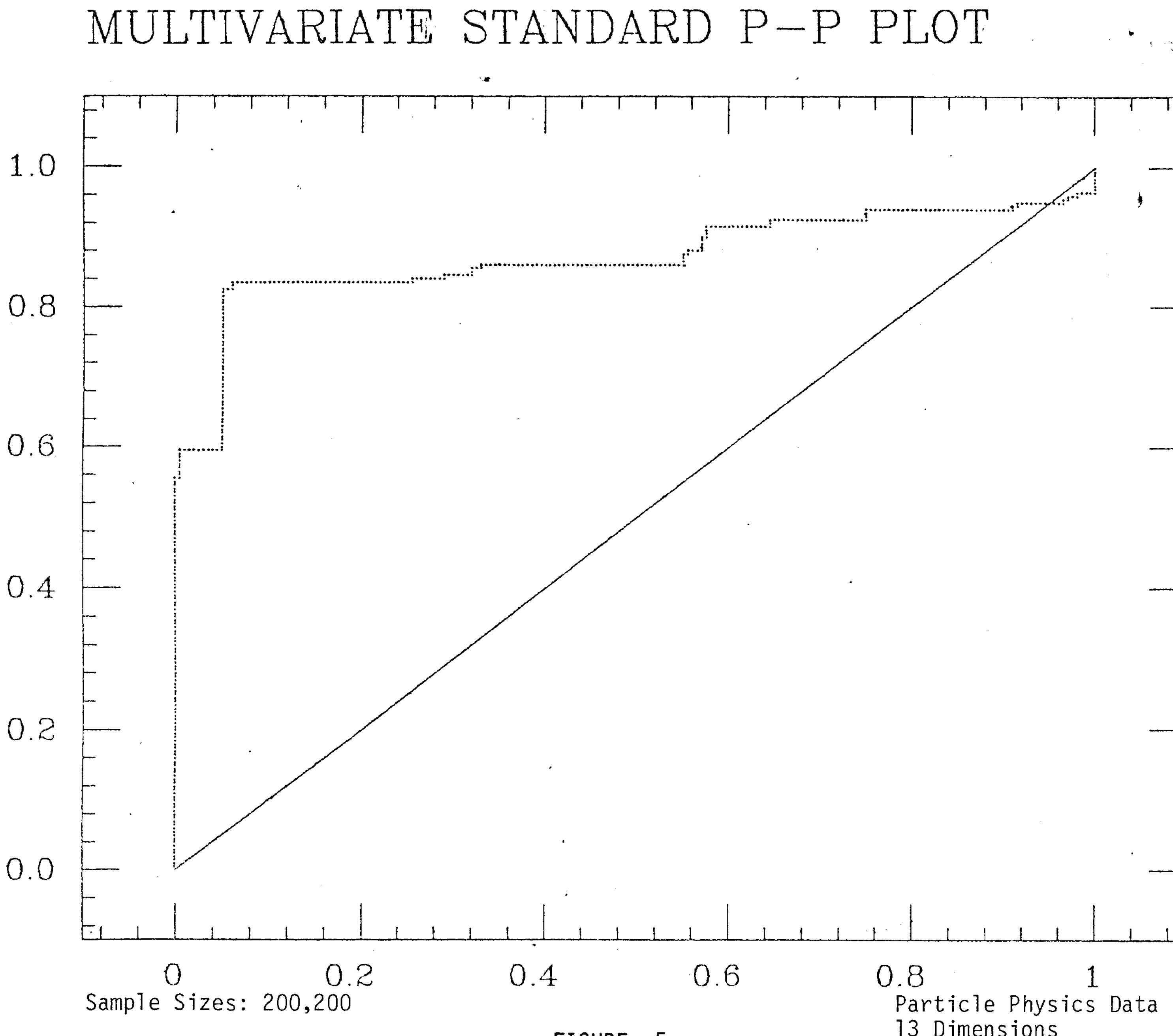

MULTIVARIATE STANDARD P-P PLOT
1.0
0.8
0.6
0.4
0.2
0.0
0
0.2
0.4
0.6
0.8
1
Sample Sizes: 200,200
Particle Physics Data
13 Dimensions

FIGURE 5a

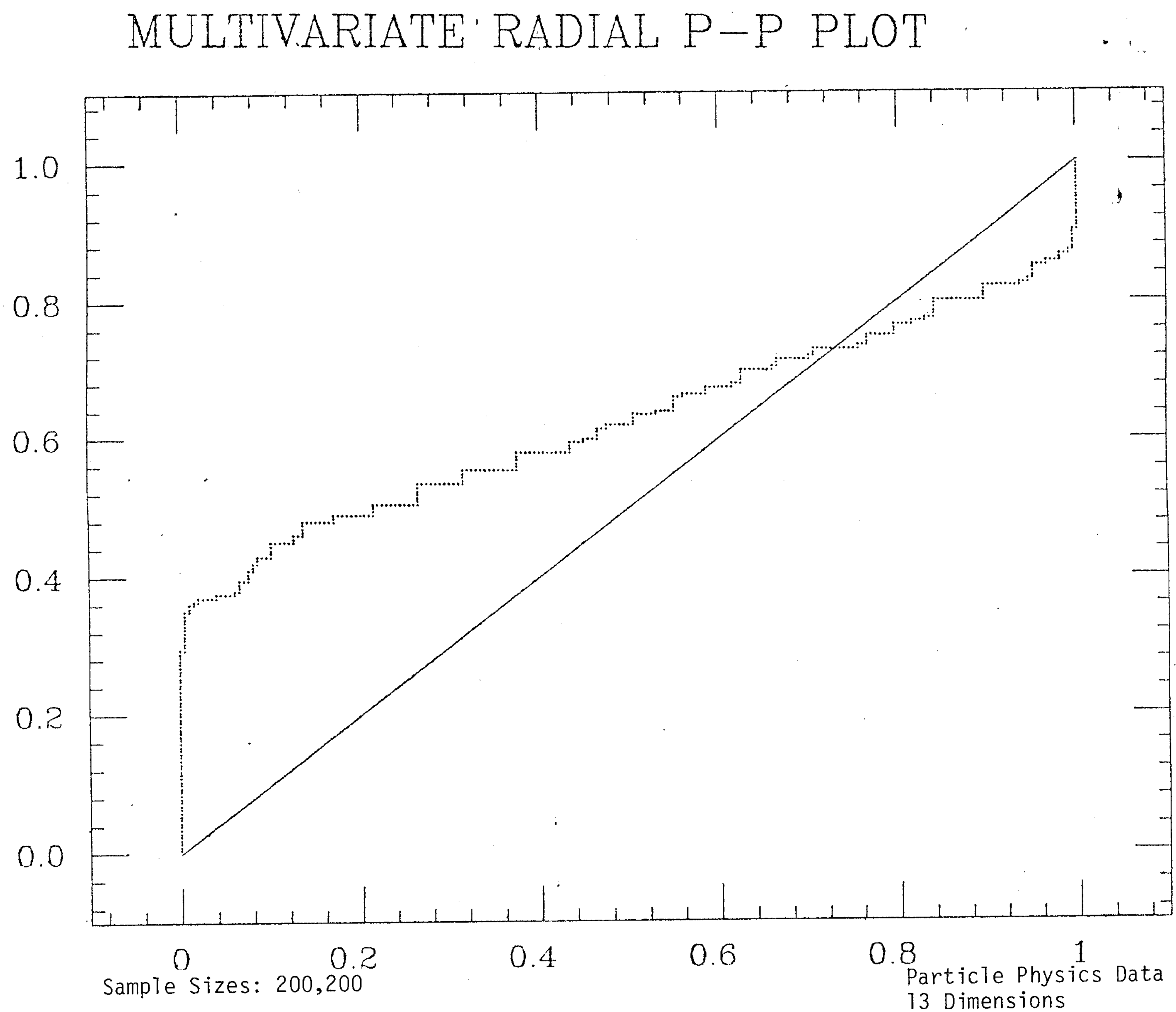
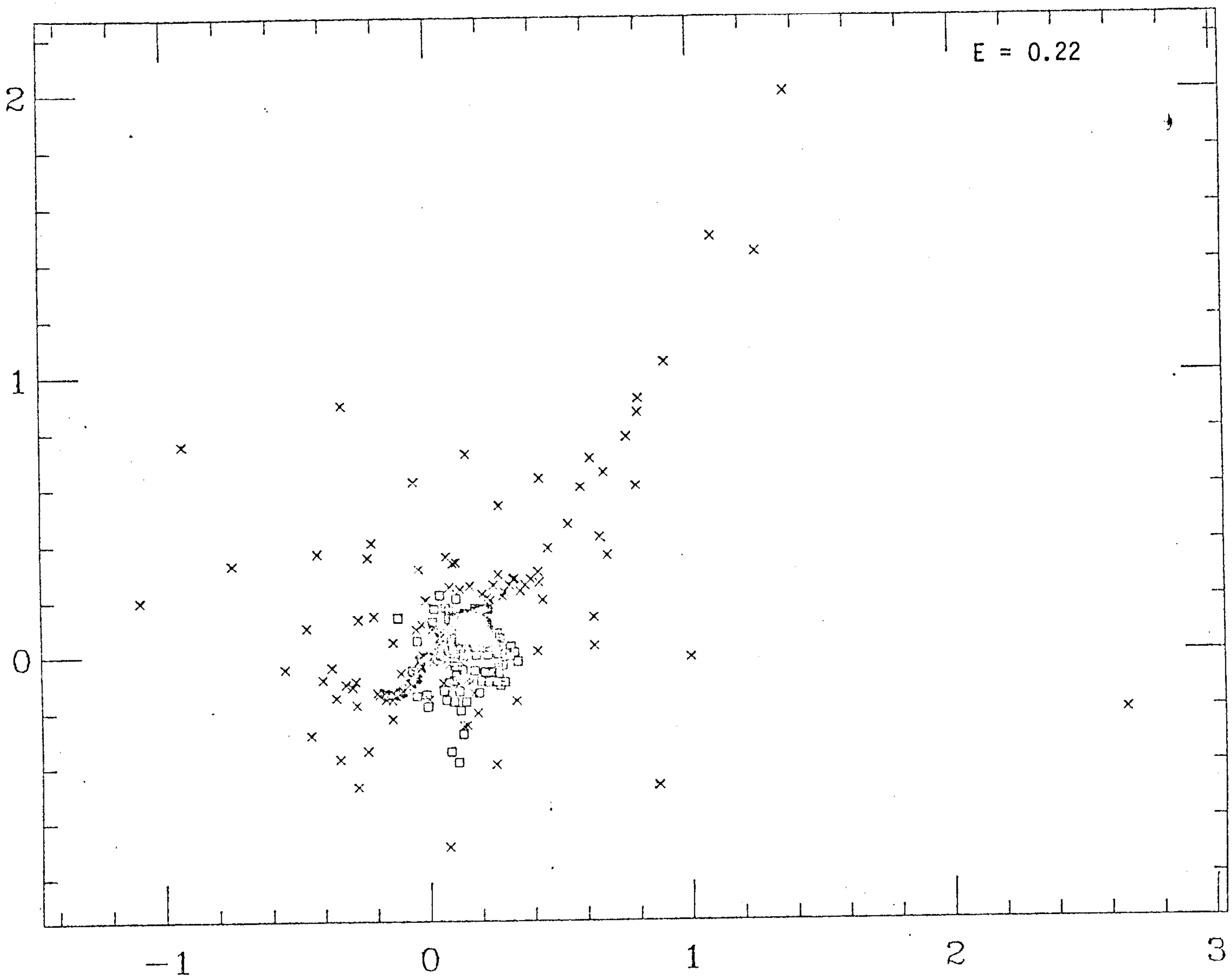

MULTIVARIATE RADIAL P-P PLOT
1.0
0.8
0.6
0.4
0.2
0.0
0
0.2
0.4
0.6
0.8
1
Sample Sizes: 200,200
Particle Physics Data
13 Dimensions

FIGURE 5b

MULTIVARIATE PLANAR REPRESENTATION



Sample Sizes: 200,200

Particle Physics Data
13 Dimensions

FIGURE 5c

MULTIVARIATE PLANAR REPRESENTATION

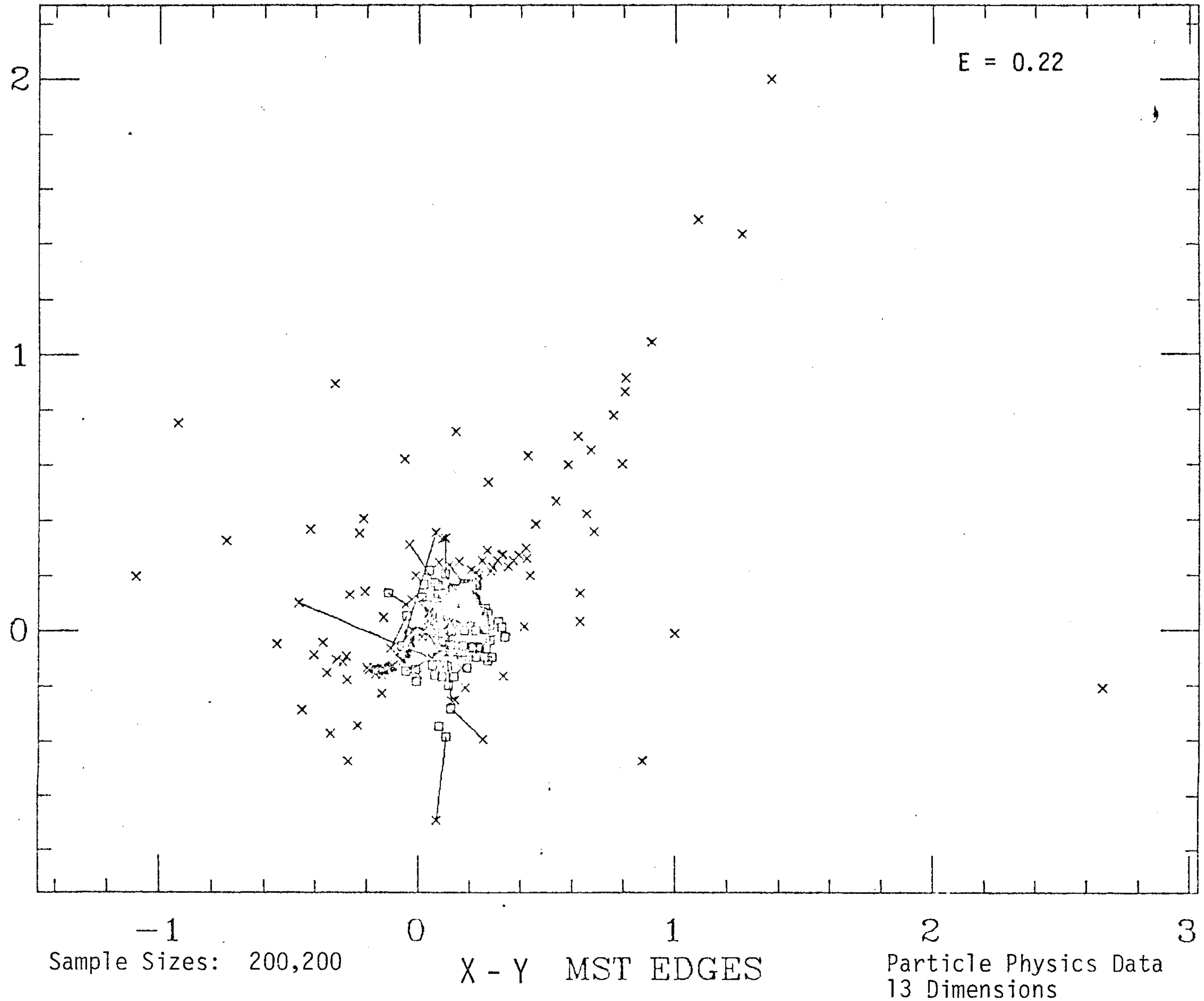


FIGURE 5d

MULTIVARIATE PLANAR REPRESENTATION

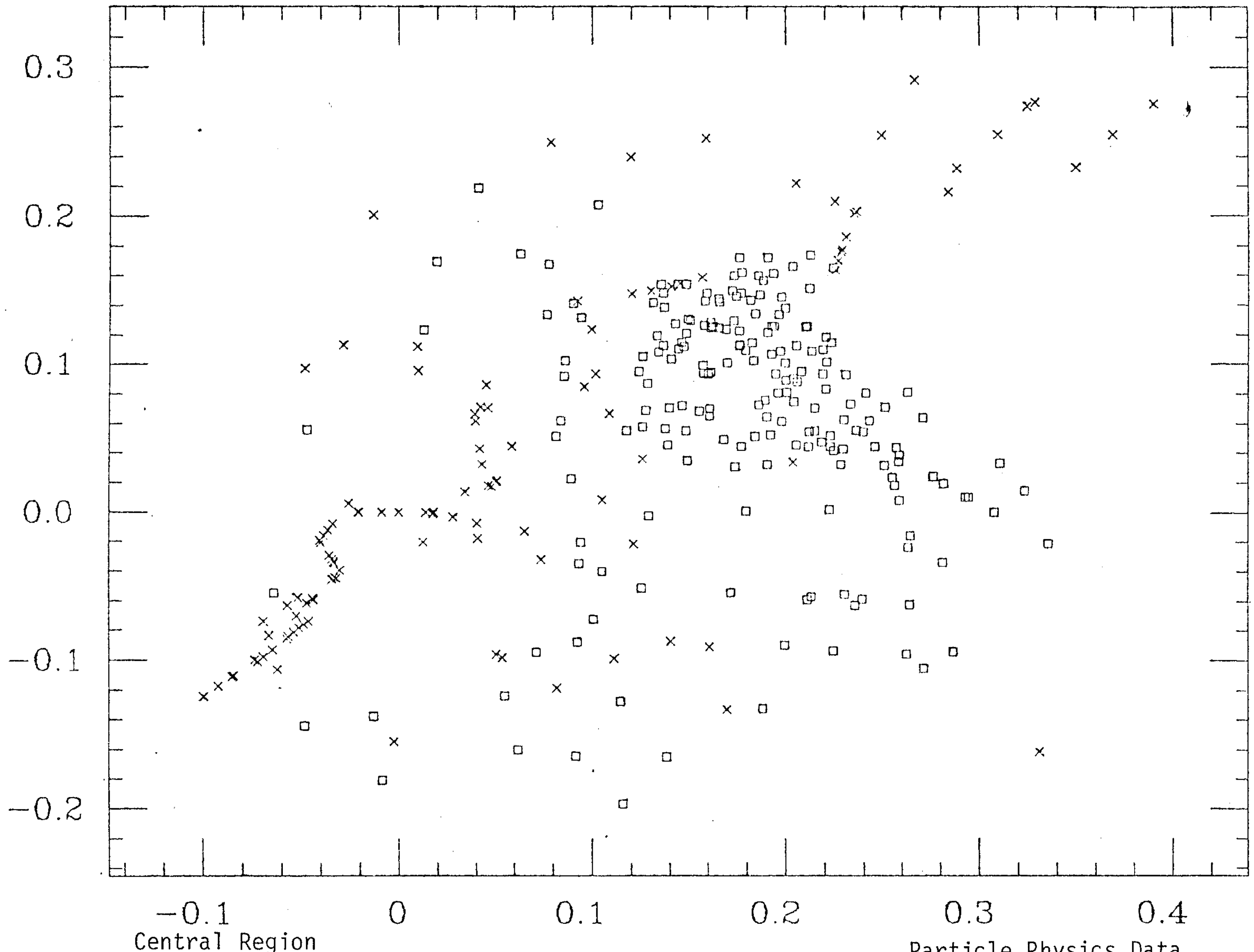


FIGURE 5e

MULTIVARIATE PLANAR REPRESENTATION

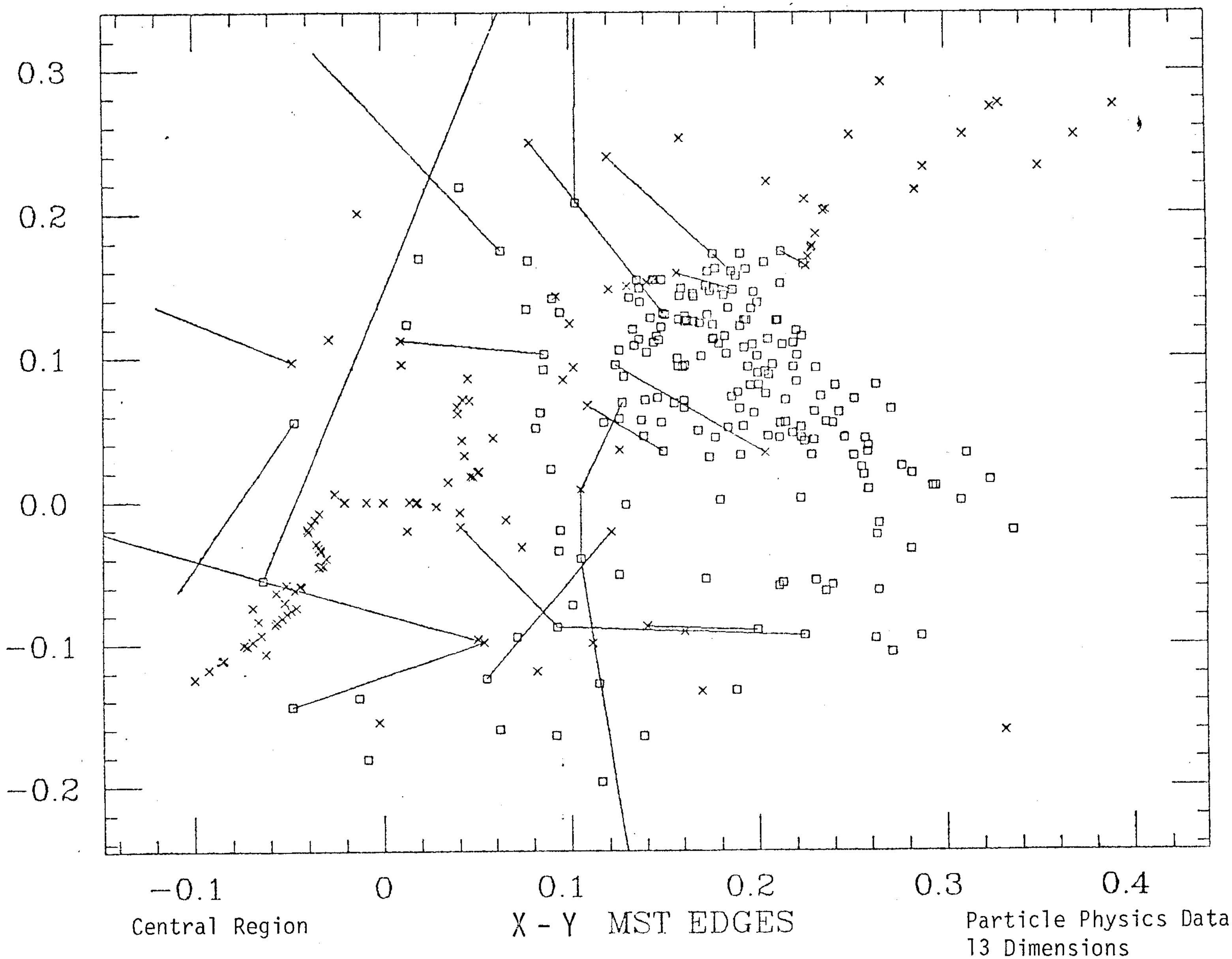


FIGURE 5f