SLAC PUB-2495
April 1980
(M)

# M and N PLOTS*

Persi Diaconis
Stanford University and Stanford Linear Accelerator Center
Stanford, California 94305

and

Jerome H. Friedman
Stanford Linear Accelerator Center
Stanford, California 94305

## ABSTRACT

We describe a class of plotting procedures which can be used to view data in one through four dimensions. The plots can be made by hand or computer. They can show four-dimensional aspects of higher dimensional data.

*Work partially supported by the Department of Energy under contract number DE-AC03-76SF00515
Work partially supported by National Science Foundation under grant MCS77-16974

(Submitted to Biometrika Journal)

Keywords and Phrases:

scatterplots

multidimensional scaling

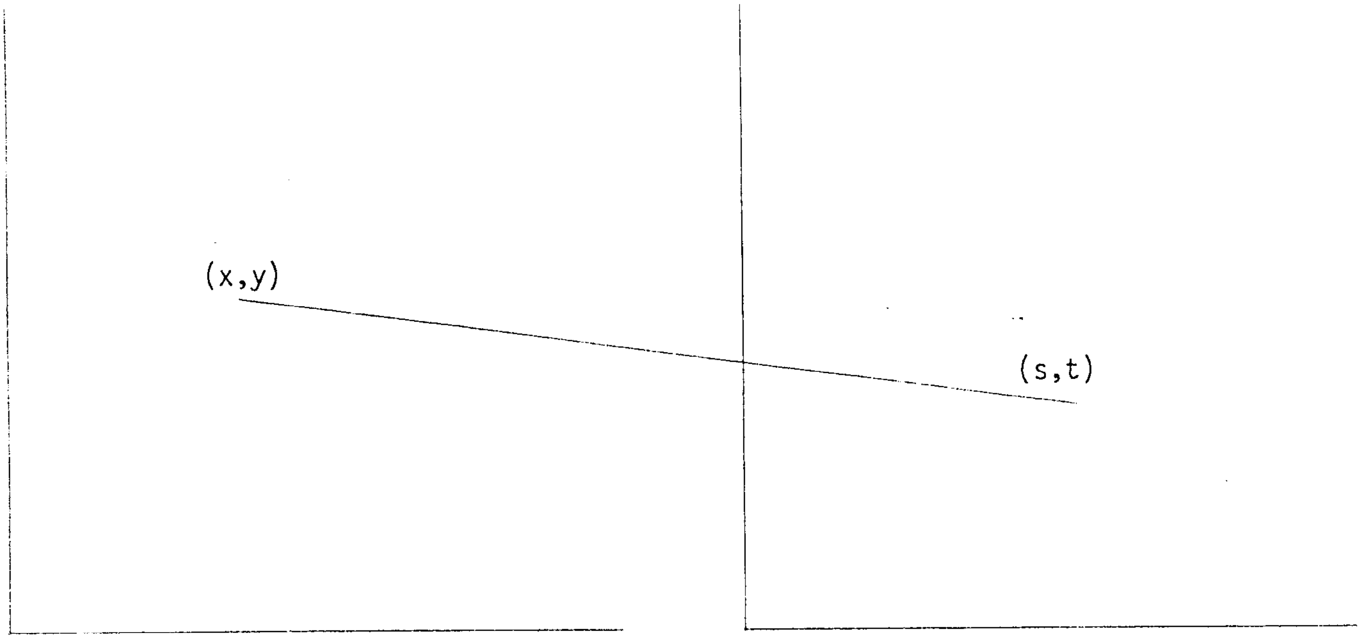
geometry and statistics

minimal spanning trees

## 1. INTRODUCTION

In this paper, we describe a method for scatterplotting four-dimensional data. The basic idea is a variation on a suggestion by Tukey and Tukey (1978): represent a four-dimensional observation p = (x,y,s,t) by a point in each of two coordinate systems.

FIGURE 1



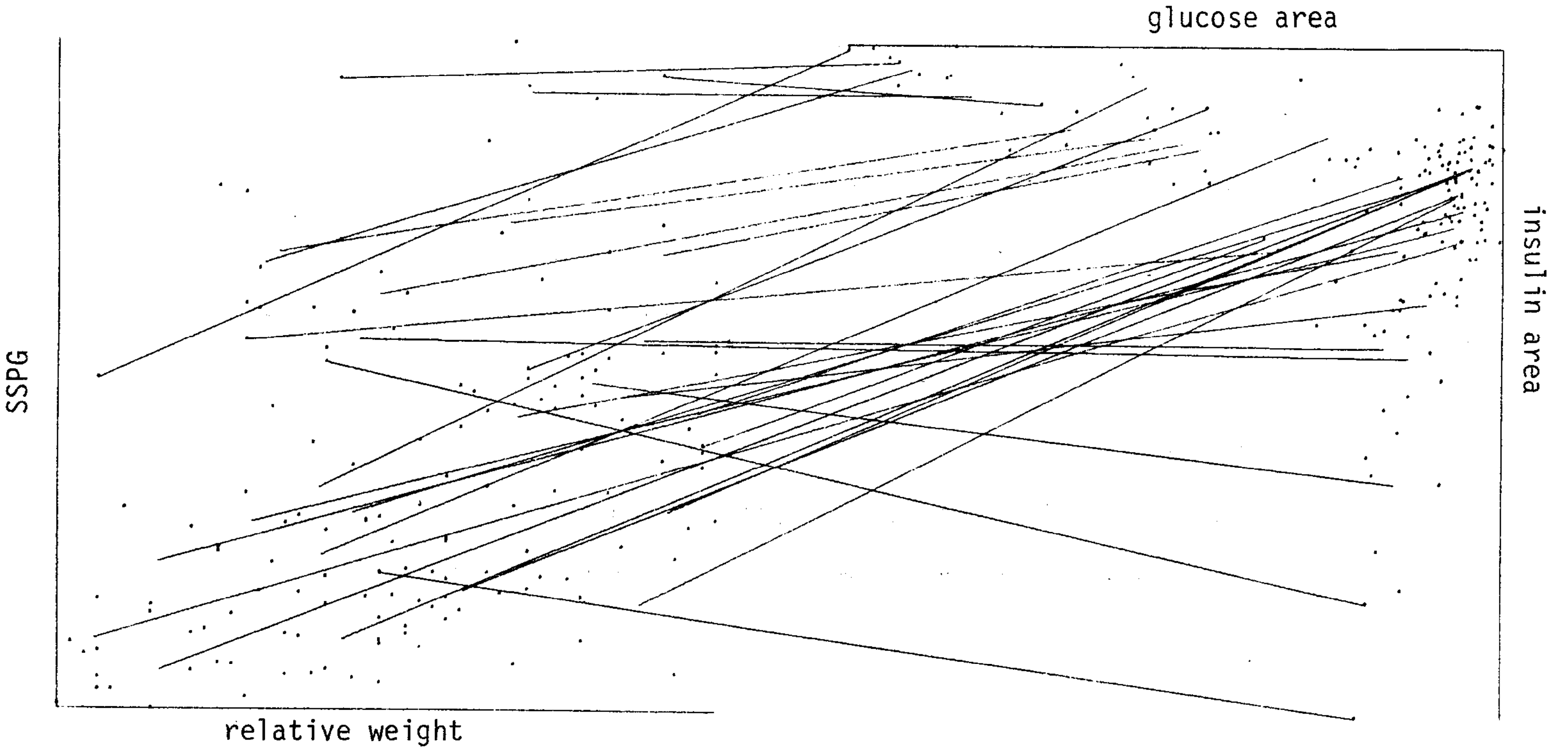
To show that the dots corresponding to (x,y) and (s,t) represent the same point p in four dimensions, connect them with a straight line. When many points are plotted, the large number of lines can become visually confusing. The lines can be "thinned down" (details are given in Section 3) so that the remaining lines give an accurate representation of which points in one plot correspond to those of the other plot.

Before a general discussion of the technique, we illustrate it by considering an example in some detail. These data are from a diabetes study of Reaven and Miller (1979). For each of 145 subjects in the study, five variables were measured. The variables, here somewhat crudely described, are:

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
5) a measure of how glucose and insulin interact - SSPG

Variables two and three, the two measures of glucose tolerance, exhibited a very high degree of linear association (r=0.96) so that only variables one, three, four and five will be considered. Figure 2 shows a 2 and 2 plot of this data. The right-hand part has been rotated 180° to make the picture easier to view (this is discussed in Section 3). We have found it useful to focus on each scatterplot separately, and then see what additional structure can be seen in the lines.

FIGURE 2



- In the left-hand plot, both variables are quite spread-out in their range. There are almost no points in the upper left-hand part of the plot, so thin people tend to have lower SSPG; medium weight and heavier people seem quite spread out in SSPG.
- In the right-hand plot, there is a good deal of structure. glucose area seems tightly clustered around low values with some scattered high values. People with high values of glucose area are generally considered to be diabetics. The insulin area variable is more uniformly spread out with a high density of medium values. Subjects with very low values of insulin area seem to have higher values of glucose area.

We next discuss the lines connecting the two plots. Consider first the lines emanating from the densist region of the right-hand picture. These lines move down to the left. They spread out on relative weight but seem to range over lower values of SSPG. This cluster of points represents "normal subjects" who have low values of glucose area and insulin area and low values of SSPG.

Next, consider the lines emanating from the points in the right-hand picture representing high values of glucose area. These lines fan out over medium values of relative weight and higher values of SSPG.

Finally, lines representing subjects with higher values of insulin area range over medium values of relative weight and medium values of SSPG.

The three groupings suggested above agree qualitatively with the groupings of Reaven and Miller (1979). One difference brought out in our analysis: relative weight seems to be a relevant factor. Further pictures of this data set are in Sections 2 and 5.

In the diabetes example, 2 and 2 plots allow some higher dimensional aspects of the data set to be seen. The idea is easily generalized. Section 2 considers examples of general M and N plots. Section 3 explains an efficient thinning algorithm and ways of drawing M and N plots by hand.

The earliest reference to 2 and 2 plots we know of is Eckhart (1968). There is a good deal of literature on visualization of four dimensions. See Manning (1960) and Brisson (1978) for surveys. The best available reference to graphical methods for high-dimensional data is Gnanadesikan (1976).

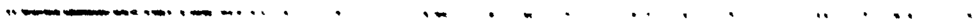
## 2. EXAMPLES OF M AND N PLOTS

Conventional scatterplots use dots on a rectangular coordinate system to graphically represent k two-dimensional vectors. An M and N plot represents (M+N)-dimensional vectors by plotting M coordinates in one coordinate system and N coordinates in a second coordinate system. The two dots representing a point are connected by a straight line segment. Thus, scatterplots are 2 and 0 plots.

### One-dimensional Data

A 1 and 0 plot, the dot plot, is sometimes used to plot one-dimensional data, plotting each point as a dot. For example, Figure 3 shows a 1 and 0 plot of the glucose area variable in the diabetes example. Figure 3 shows the clustering around low values and the spreading at higher values reasonably well. A histogram might be a more informative picture for this data set.

FIGURE 3

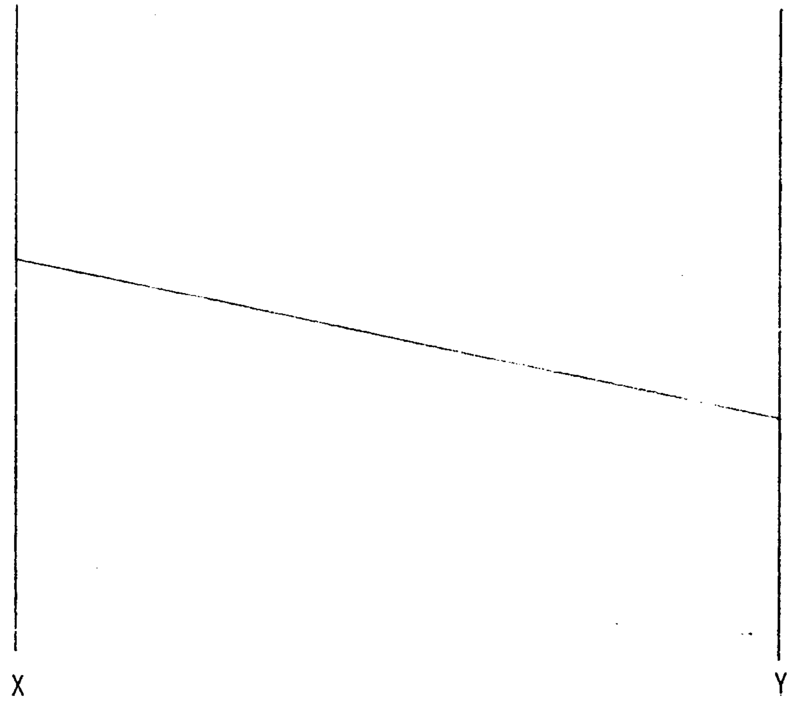


Glucose area

Two-dimensional Data

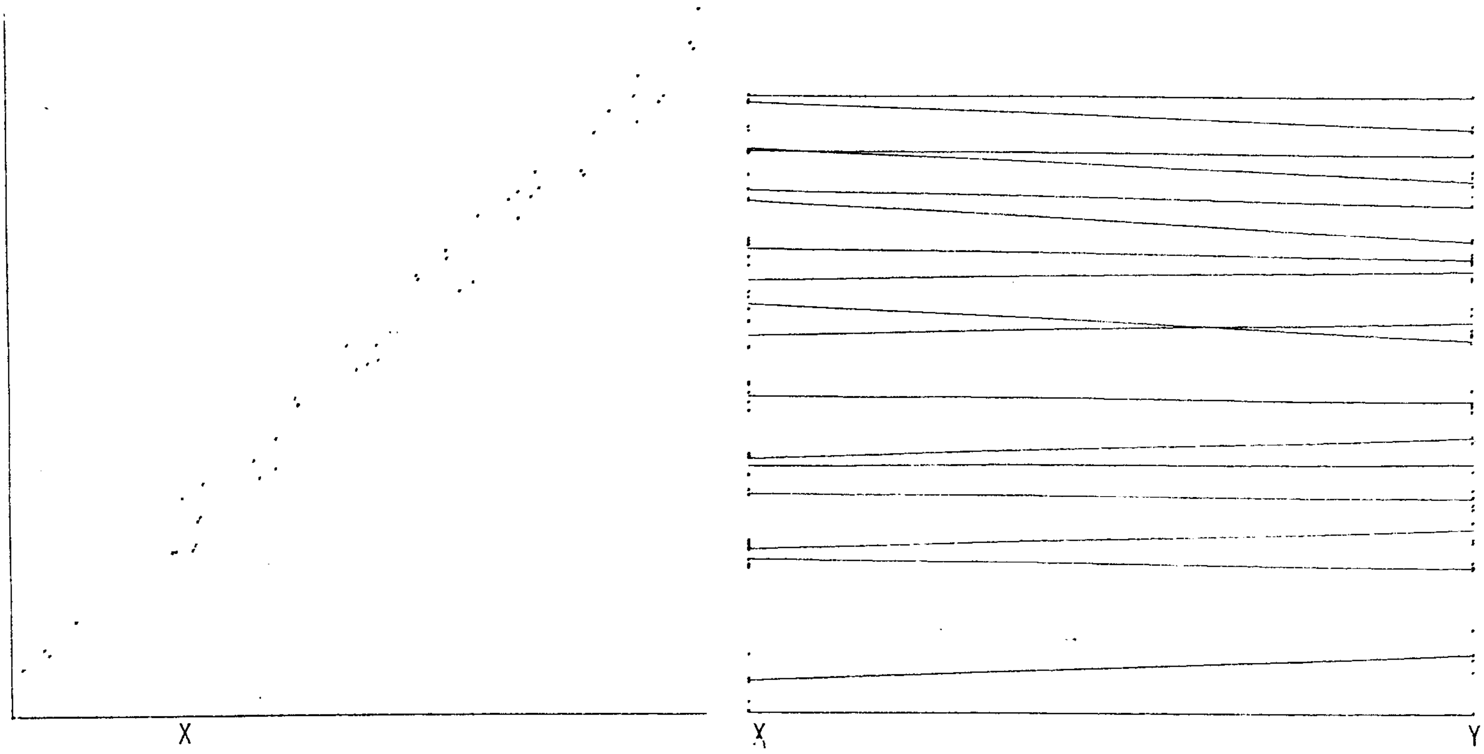
There are two ways to draw M and N plots of two-dimensional data - the conventional scatterplot (a 2 and 0 plot), and the 1 and 1 plot. A 1 and 1 plot represents a two-dimensional point (x,y) by 2 dots and a line segment on a pair of parallel coordinate axis:

FIGURE 4



It is useful to look at 1 and 1 plots of familiar point clouds. For example, Figure 5 shows a cloud of points lying close to a straight line and the corresponding 1 and 1 plot.

FIGURE 5

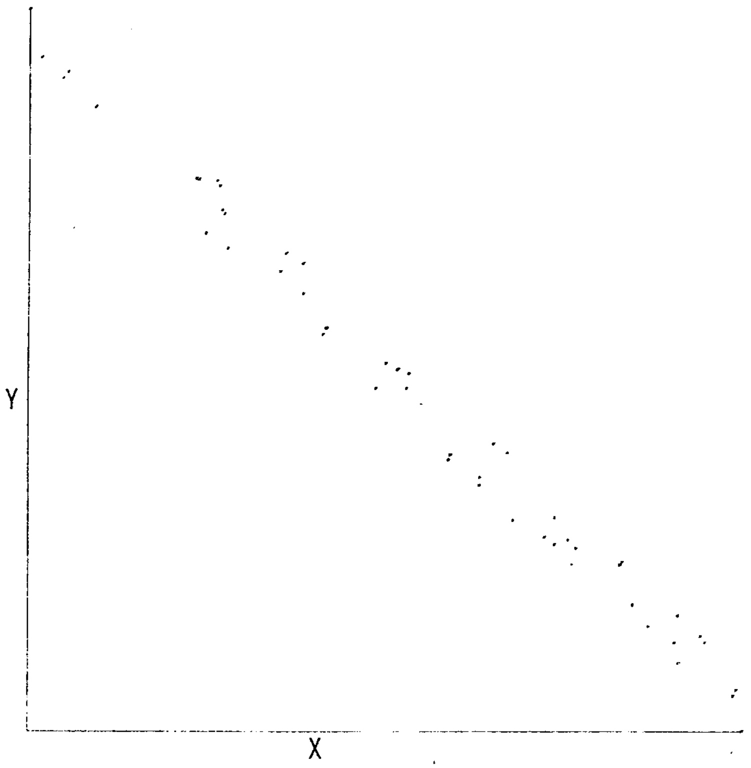


Before considering other examples of l and l plots, we describe the connection between l and l plots and the set of lines in the plane. To avoid confusion, the line segments in l and l plots, such as Figure 5, will be called segments in what follows. Each segment can be thought of as a section of the (infinite) line in the plane that passes through that segment. Thinking of segments as lines helps in understanding the parallel segments in Figure 5.

Consider a collection of points in two dimensions: $(x_1,y_1)$, $(x_2,y_2),\ldots,(x_n,y_n)$. It is easy to see that these points lie on a line if and only if the lines corresponding to these points in a 1 and 1 plot intersect at a point. As usual, parallel lines are thought of as intersecting at infinity. An illustration of this is in Figure 6.

The next group of figures (Figures 7, 8 and 9) are examples of familiar scatterplots. It is useful to learn to recognize the structure of lines in the associated 1 and 1 plot because such structures occur in higher dimensional M and N plots.

FIGURE 6



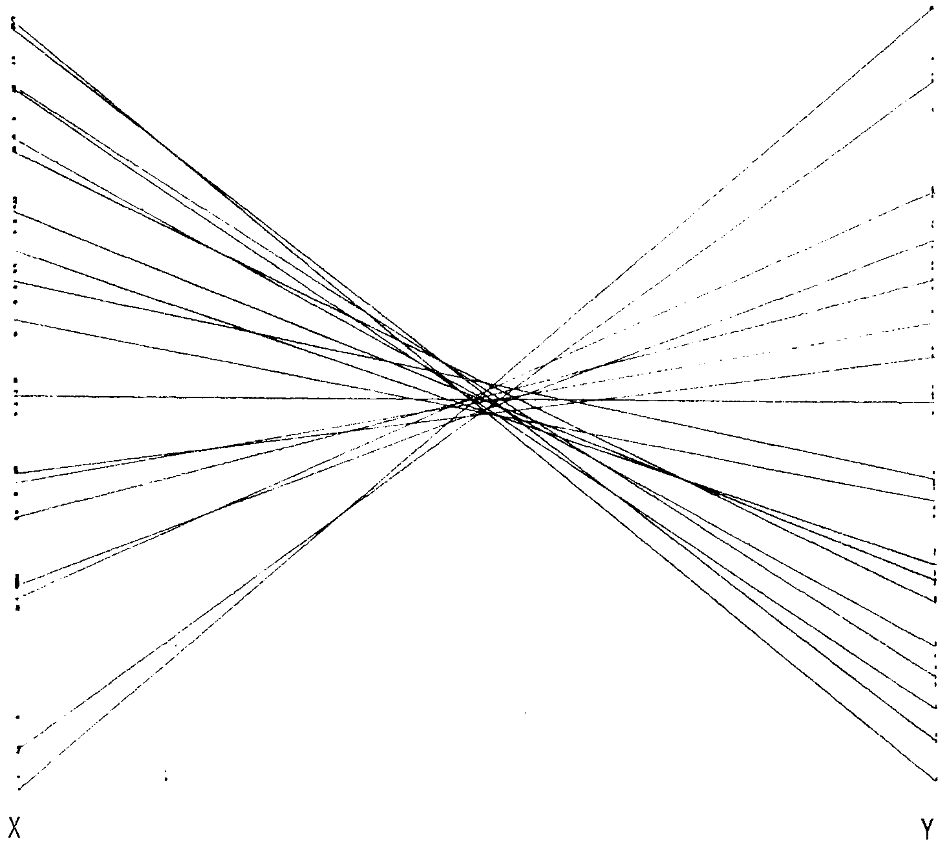


FIGURE 7

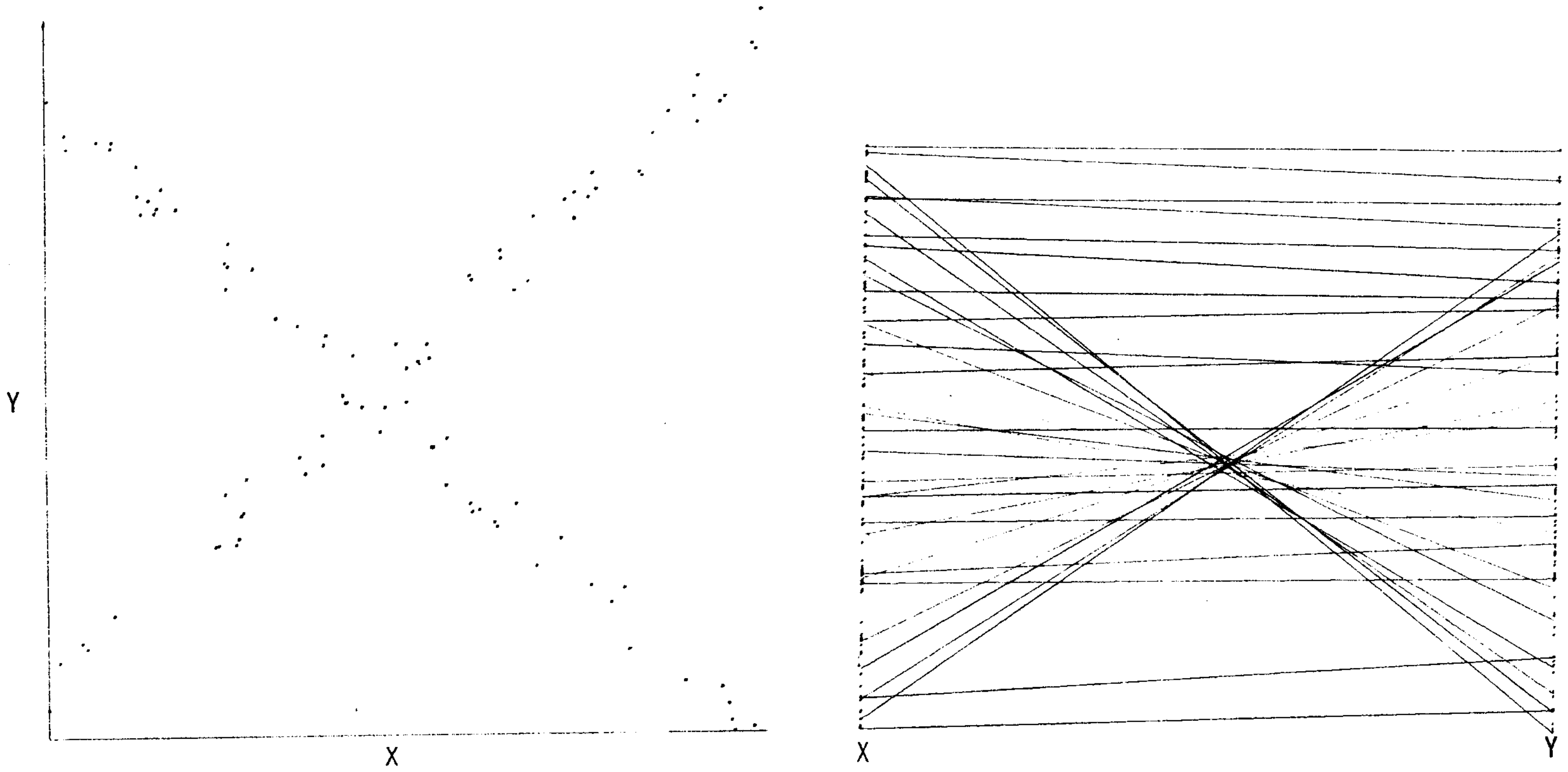


FIGURE 8

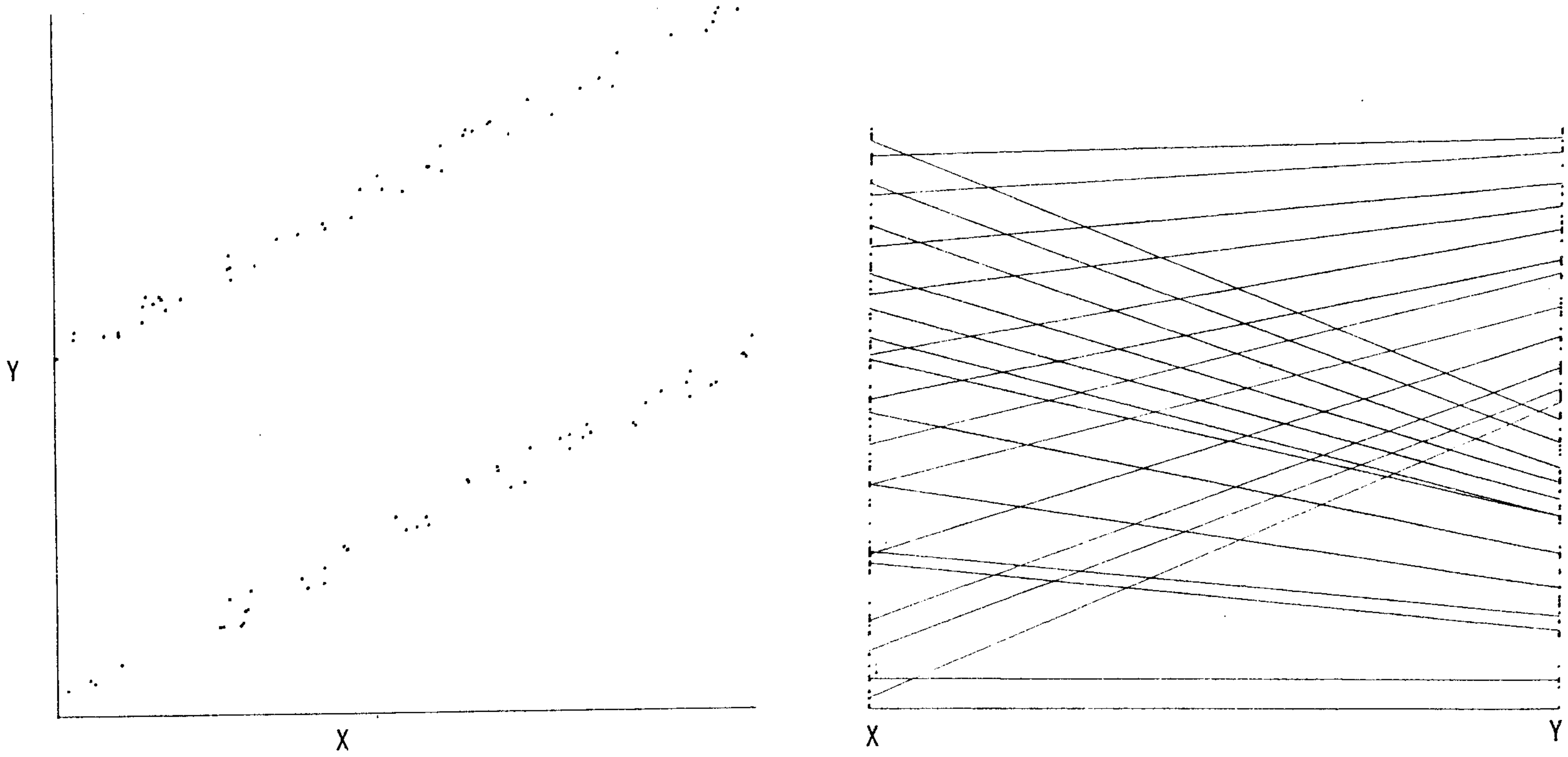
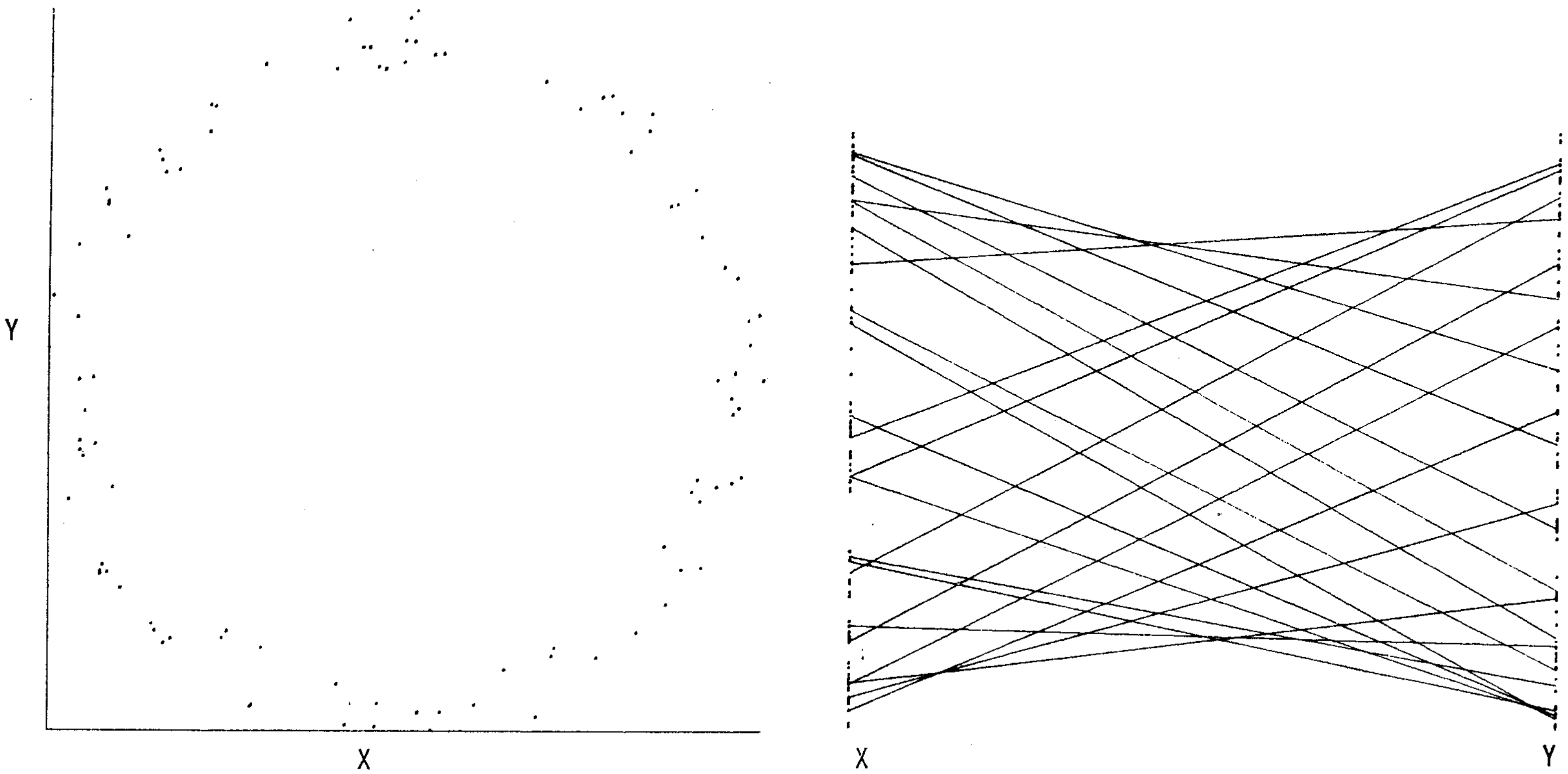


FIGURE 9

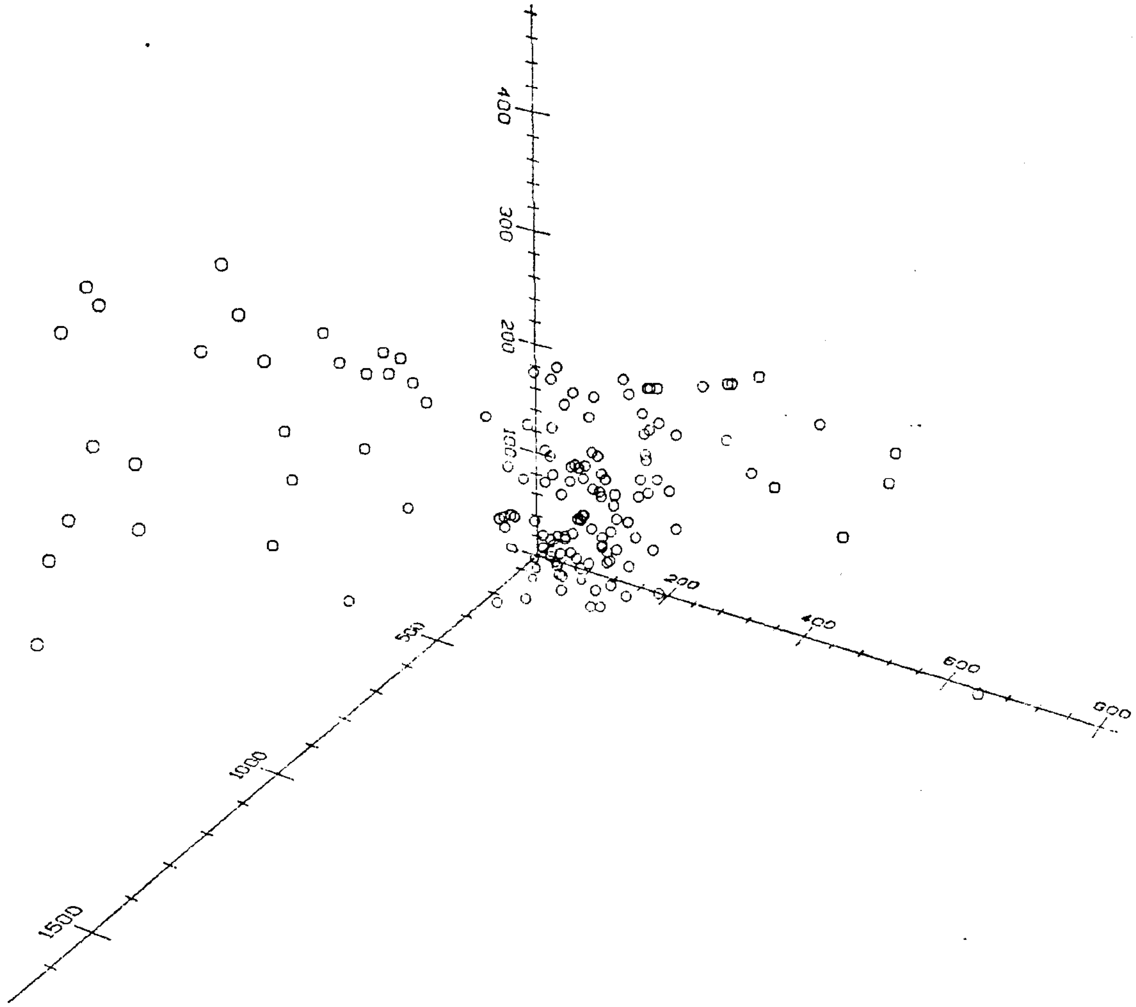


One other aspect of 1 and 1 plots should be noted: the number of times that the segments cross in a 1 and 1 plot is equal to $\frac{1}{2}\binom{n}{2}[1-\tau]$ where $\tau$ is Kendall's measure of association, $\tau$ computed for $(x_1,y_1),\ldots(x_n,y_n)$. Thus, few crossings correspond to $\tau$ close to 1 and many crossings correspond to $\tau$ close to -1. The earliest reference we know for this is Griffin (1958).

Three-dimensional Data

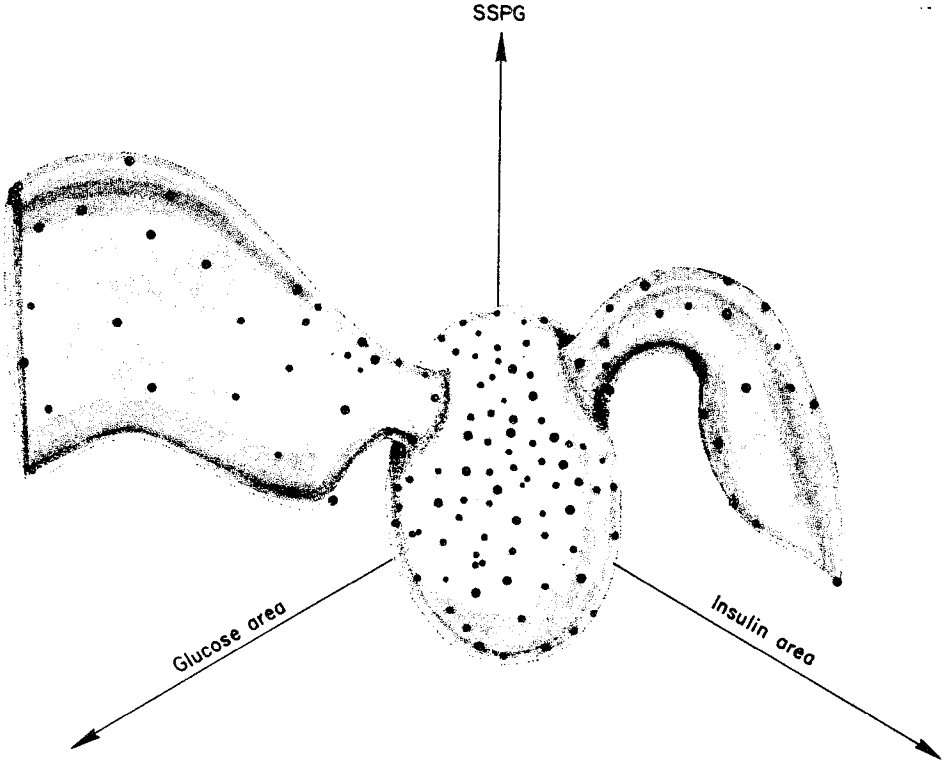
It is not a straightforward task to make a scatterplot of three-dimensional data -- a 3 and 0 plot. Consider Figure 10 which is a scatterplot of three of the variables from the diabetes data set introduced in Section 1.

FIGURE 10



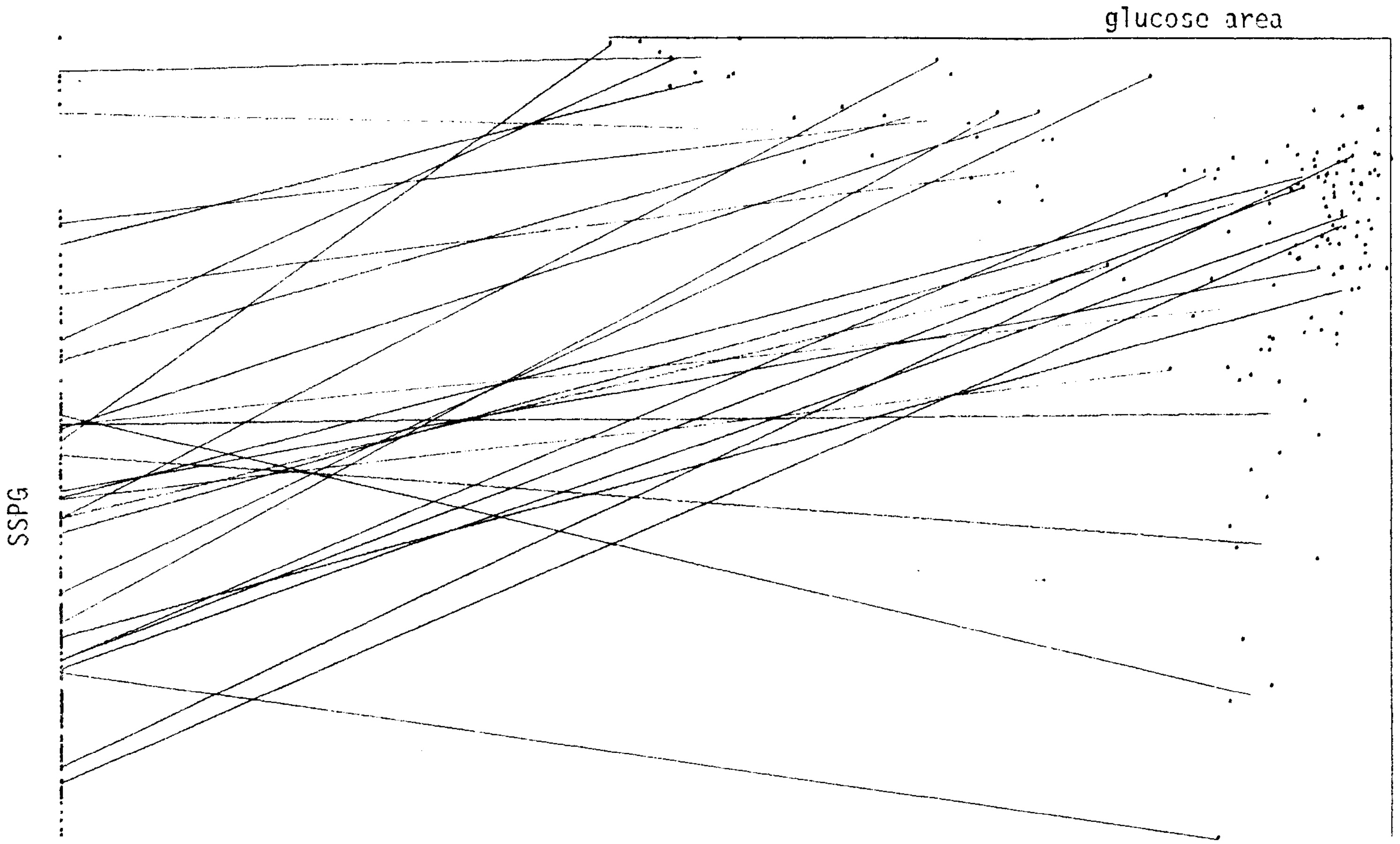
The plot is not easy to make sense of. A version of a three-dimensional scatterplot is at the center of the PRIM-9 plotting program (Fisherkeller, et al (1976)). Briefly, the idea is to show the collection of three-dimensional points as a rotating point cloud using a graphics display terminal. Points closer to the viewer rotate faster and parallax fools the eye into seeing the points as a three-dimensional point cloud. An artist's rendering of this display for the data of Figure 10 gives a very useful picture of this data set. Figure 11 is reproduced from Reaven and Miller (1979).

FIGURE 11



Artist's rendition of data as seen in three dimensions. View is approximately along 45° line as seen through Prim 9 program on the computer; coordinate axes are in the background

FIGURE 12



A 1 and 2 plot of the same three variables from the diabetes data set is shown in Figure 12. The right-hand scatterplot was discussed in Section 1. The 1 and 0 plot of SSPG shows a high density of lower values. Looking at the lines, we see that points from the central cluster of "normal" patients have low  values of SSPG. As one looks down in the right-hand plot, the values of SSPG increases, always remaining below the middle of the range. Looking left from the main cluster in the right-hand plot, the values of SSPG increase somewhat slowly. At the very top, there appears to be an interesting reversal -- the highest values of glucose area are connected to central values of SSPG.

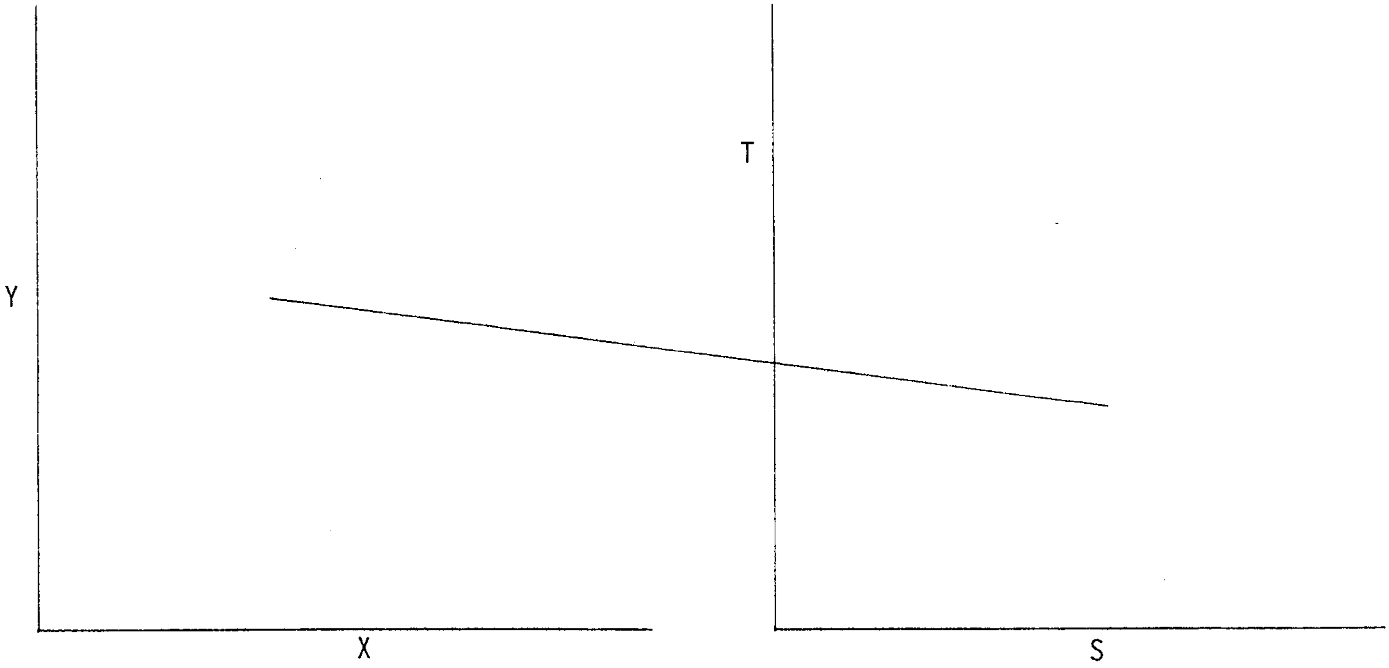
The line segments in a 1 and 2 plot can also be viewed as a plot of lines in three dimensions by picturing the one-dimensional axis as a line lying above and parallel to the plane of the two-dimensional plot.

Another way to plot three-dimensional data is to make a (1,1,1) plot -- three distinct coordinate axes, each showing one dimension and a pair of line segments connecting each dot to the other two representing dots.

Four-dimensional Data

We have already given an example of a 2 and 2 plot in Section 1. We here discuss an interpretation of such plots as a picture of lines in three dimensions. The basic ingredients of a 2 and 2 plot are a pair of coordinate axes and a segment:

FIGURE 13



Imagine the (S,T) plane as a plane parallel to but above the (X,Y) plane. A segment can then be pictured as part of a line running "down" from the (S,T) plane through the (X,Y) plane. Only that part of the line which lies between the planes is visible. In this way, the segments in a 2 and 2 plot can be thought of as lines in three-dimensional space. It is not hard to see that the set of all lines in three dimensions is a four-dimensional space. One argument considers a pair of parallel planes (like the (X,Y) and (S,T) planes described above). "Almost all" lines in the three-dimensional space pass through both planes and uniquely determine four coordinates. This omits lines parallel to the planes, but these form a lower dimensional surface. Hence, the lines in three-dimensional space form a four-dimensional space and can be used to picture four-dimensional data. A similar argument shows that the dimension of the set of lines in n-dimensional space is 2(n-1).

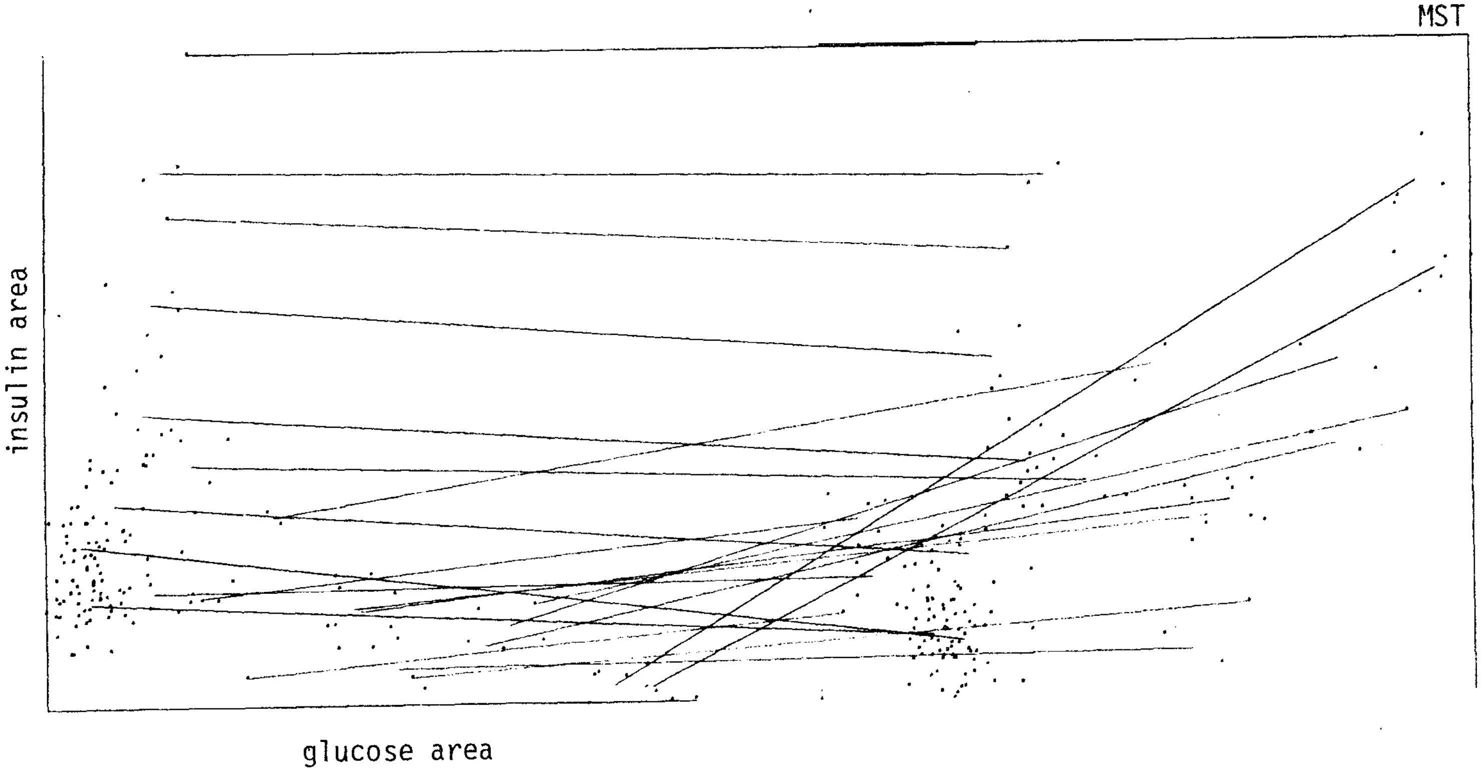
It is natural to try to find a set of coordinates for the lines in three dimensions which do not have the problem of omitting a low-dimensional set of lines. Several approaches are described in Chapter 1 of Jessop (1969). The most widely discussed coordinates - Plückers coordinates -- are not particularly suited to working with statistical data. While natural mathematically, Plückers coordinates use five coordinates to describe the four-dimensional set of lines.

We mention, in passing, that four-dimensional data can also be viewed using a (2,1,1) plot, a (1,1,1,1) plot or a (3,1) plot (a 3 and 1 plot would require a PRIM 9-like graphics device).

## Higher-dimensional Data

With a PRIM 9-like graphical device, or an artist's rendering as in Figure 11, it is possible to draw 3 and 3 plots of six-dimensional data. Higher dimensional data can also be pictured by connecting together lower dimensional M and N plots. A more practical thought is to link the best two-dimensional projection with a plot of the output of one of the many nonlinear mapping or scaling algorithms. This gives a way of labeling the points of the resulting output. An example is given in Figure 14. This is a picture of the Reaven and Miller (1979) diabetes data. The right-hand plot is nonlinear mapping of the original five-dimensional data into the plane. The nonlinear mapping algorithm described in Friedman and Rafsky (1979) was used. This algorithm preserves 2n-1 of the $\binom{n}{2}$ interpoint distances. In this picture, the n-1 distances in the minimal spanning tree of the n points are preserved along with selected other distances. The right-hand plot has a dense area in the lower left-hand part and two "wings". The lines connect this plot to the ordinary projection of the data onto coordinates 2 and 4. The lines indicate that the dense part of the right-hand picture corresponds to "normal" patients and the two "wings" correspond to the other two groups described previously. Looking more closely at the right-hand picture, it seems that the group of "normal" subjects may split into two groups - a dense lower group and a less dense upper group. An outlying observation shows up clearly in the line at the top of the plots.

FIGURE 14



## 3. SOME PRACTICAL DETAILS

In this section, we discuss thinning -- by hand or computer -- rotating, and some ideas for interactive implementation of M and N plots.

### Thinning

To understand the need and results of thinning, consider the data set introduced in Figure 1. Figures 15-25 show the results of different amounts of thinning.

FIGURE 15

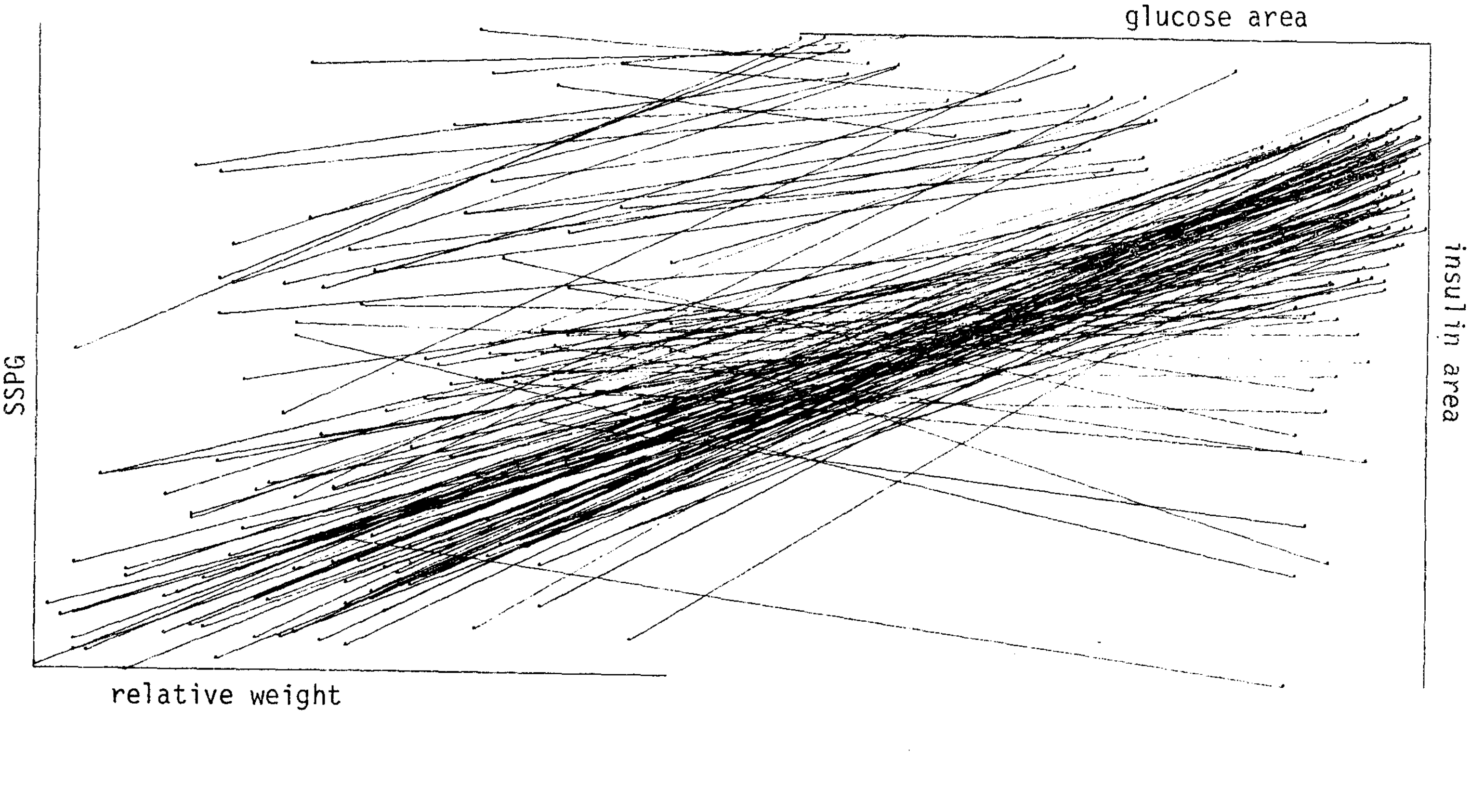


FIGURE 16

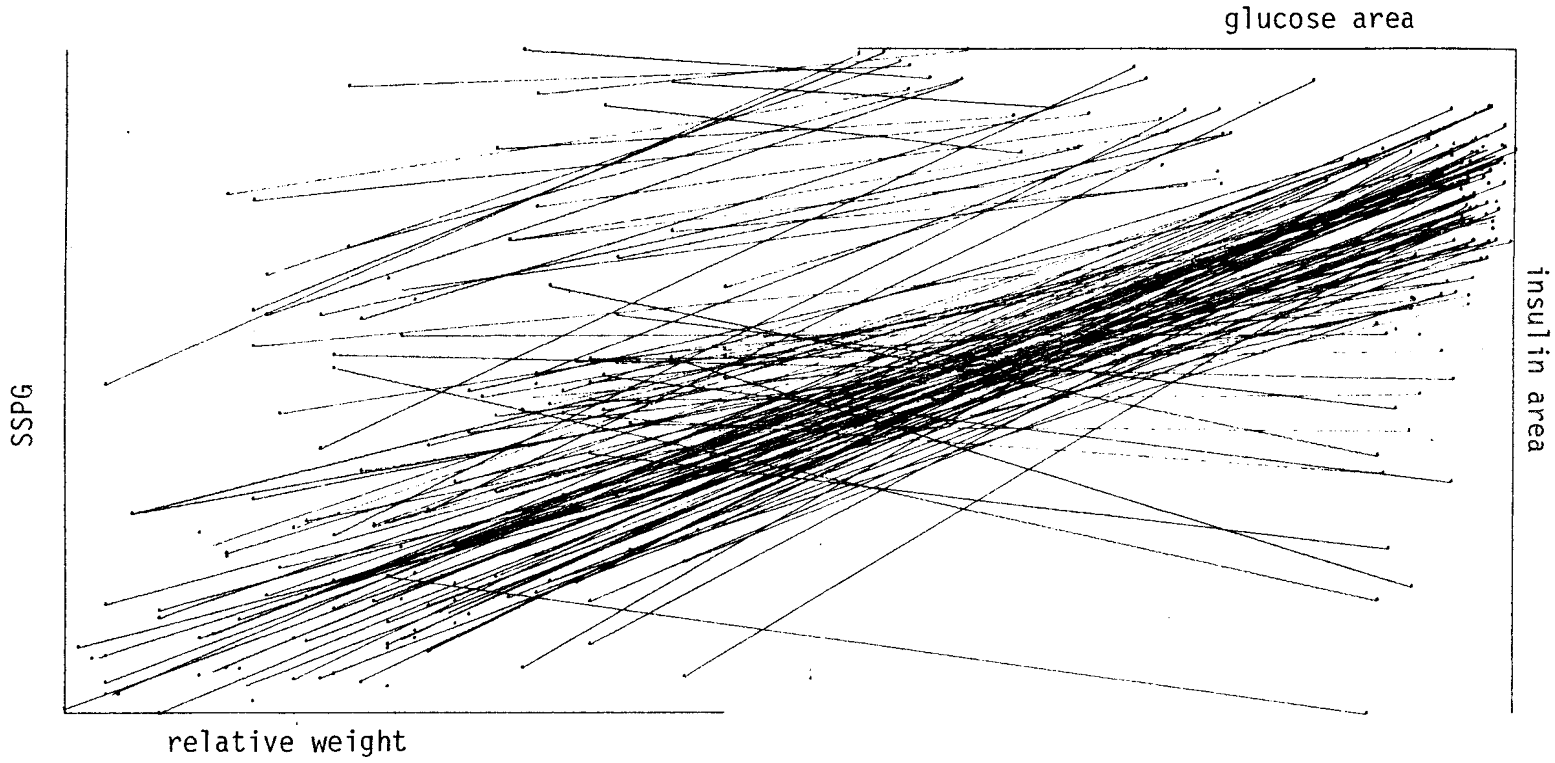


FIGURE 17

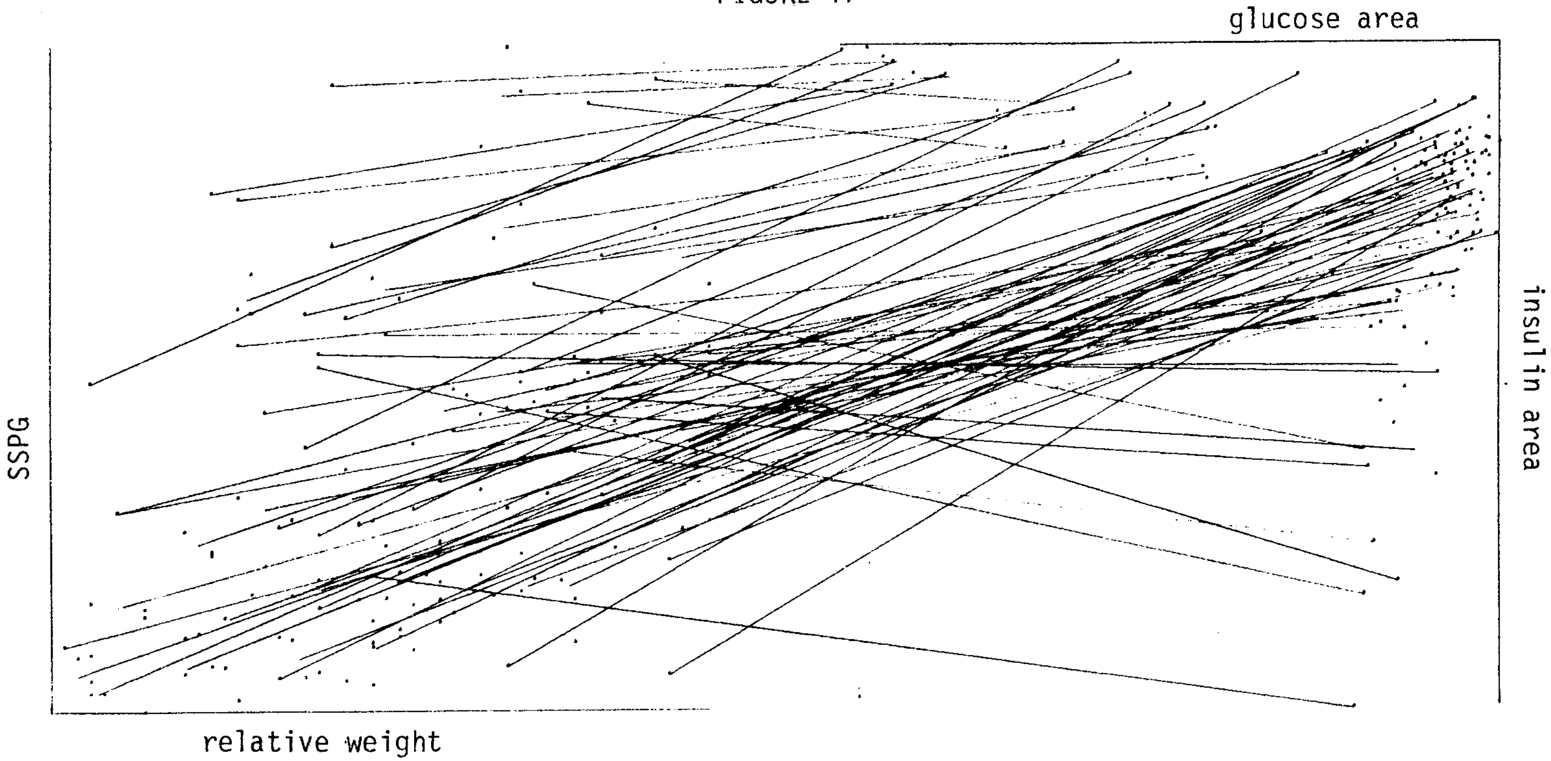


FIGURE 18

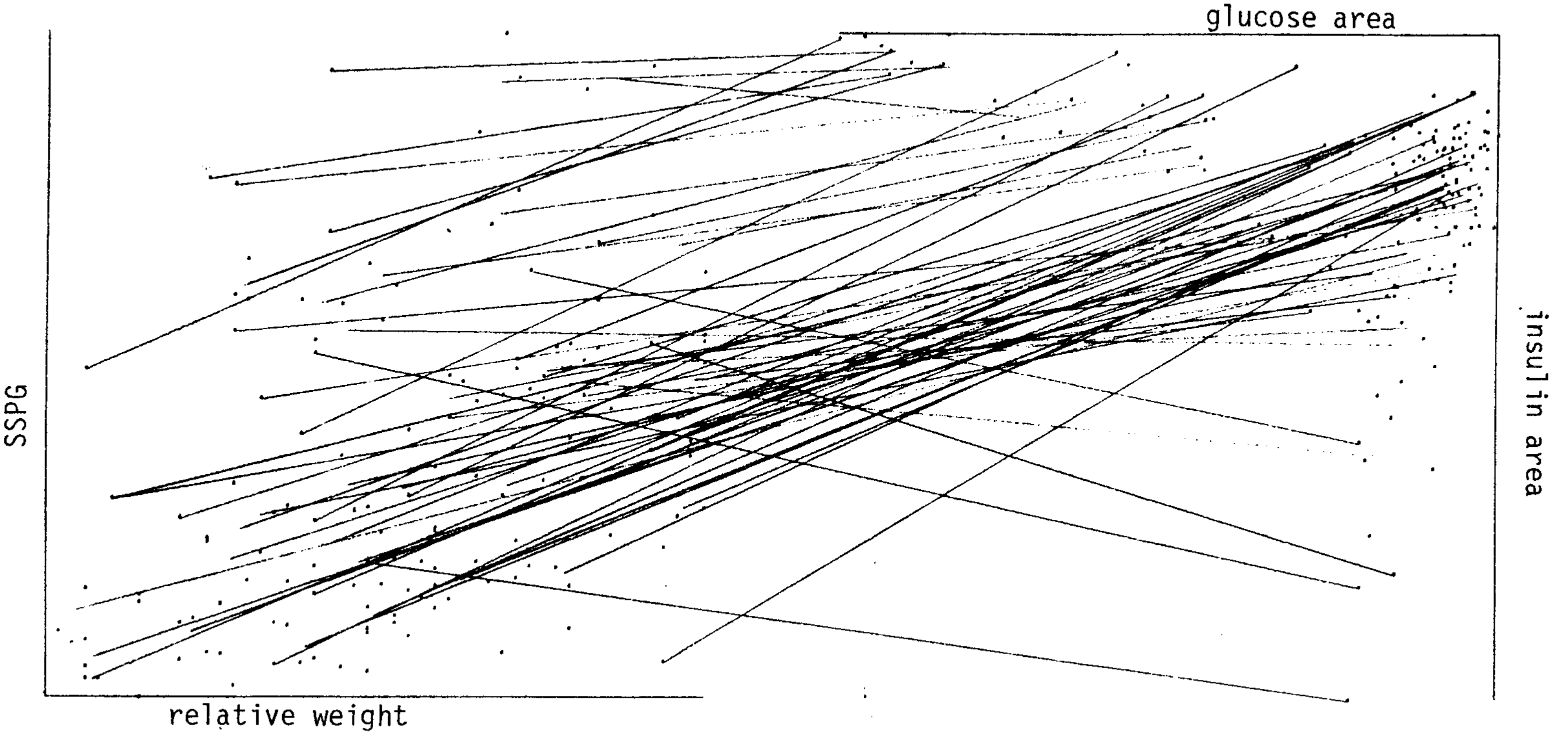


FIGURE 19

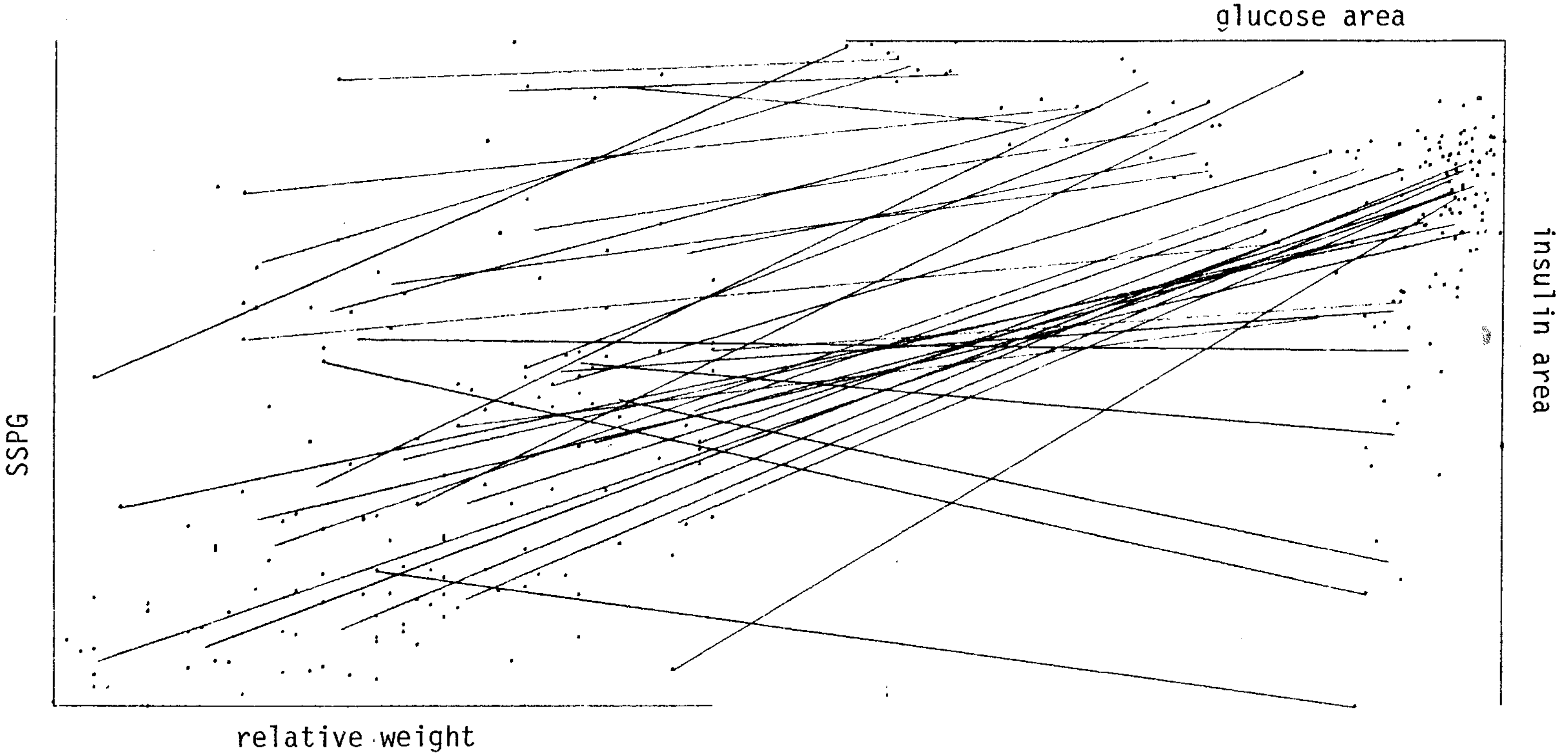


FIGURE 20

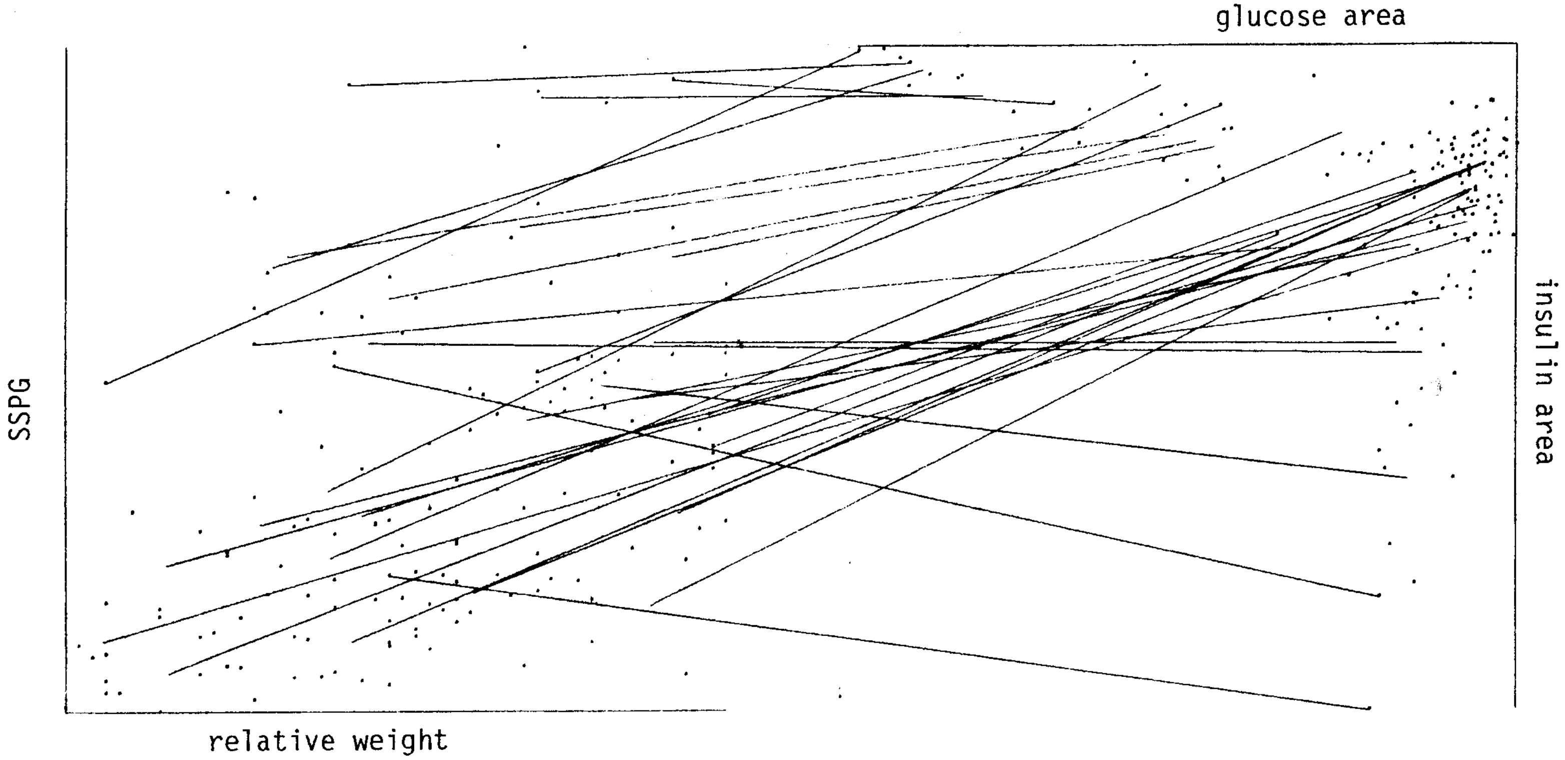


FIGURE 21

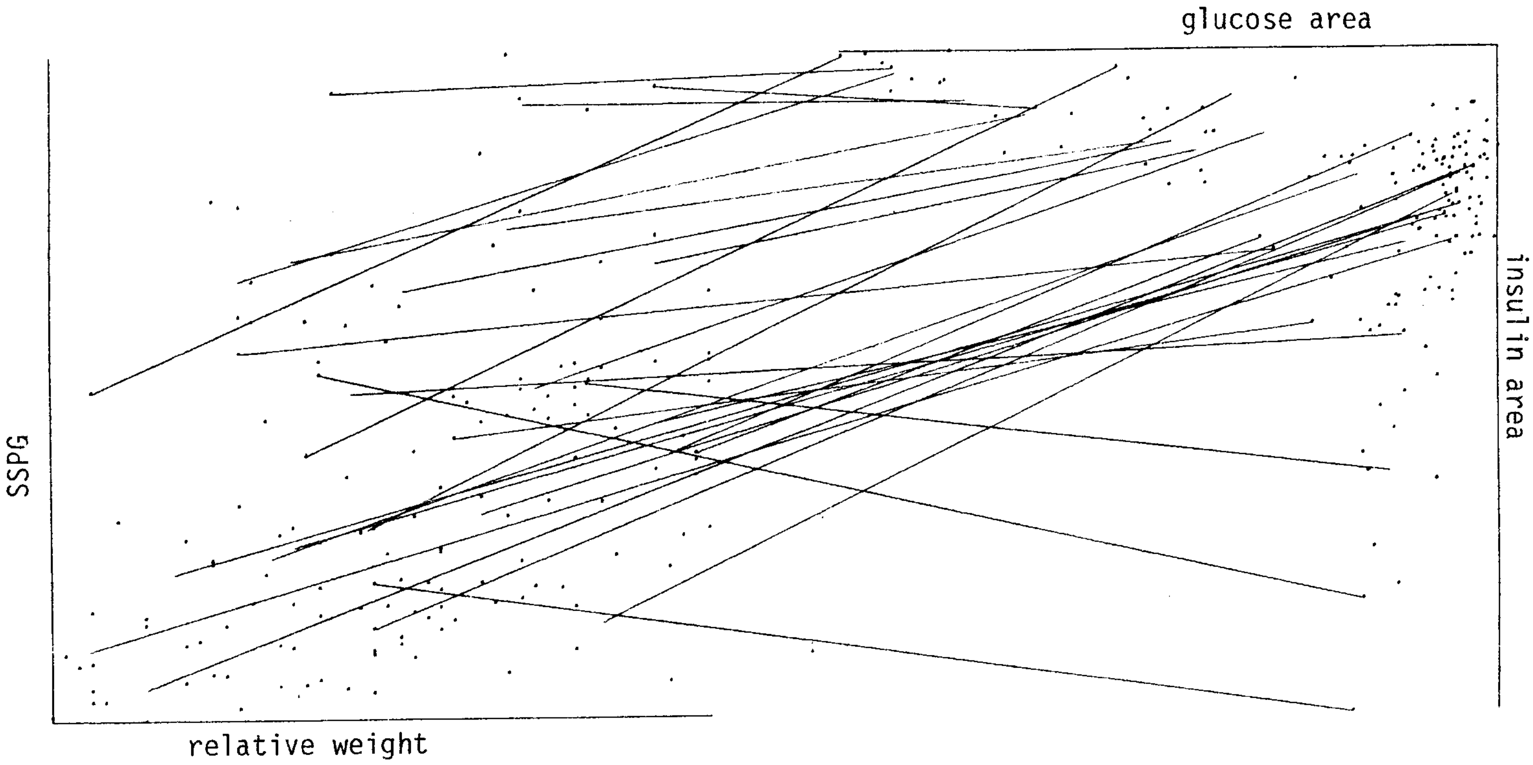


FIGURE 22

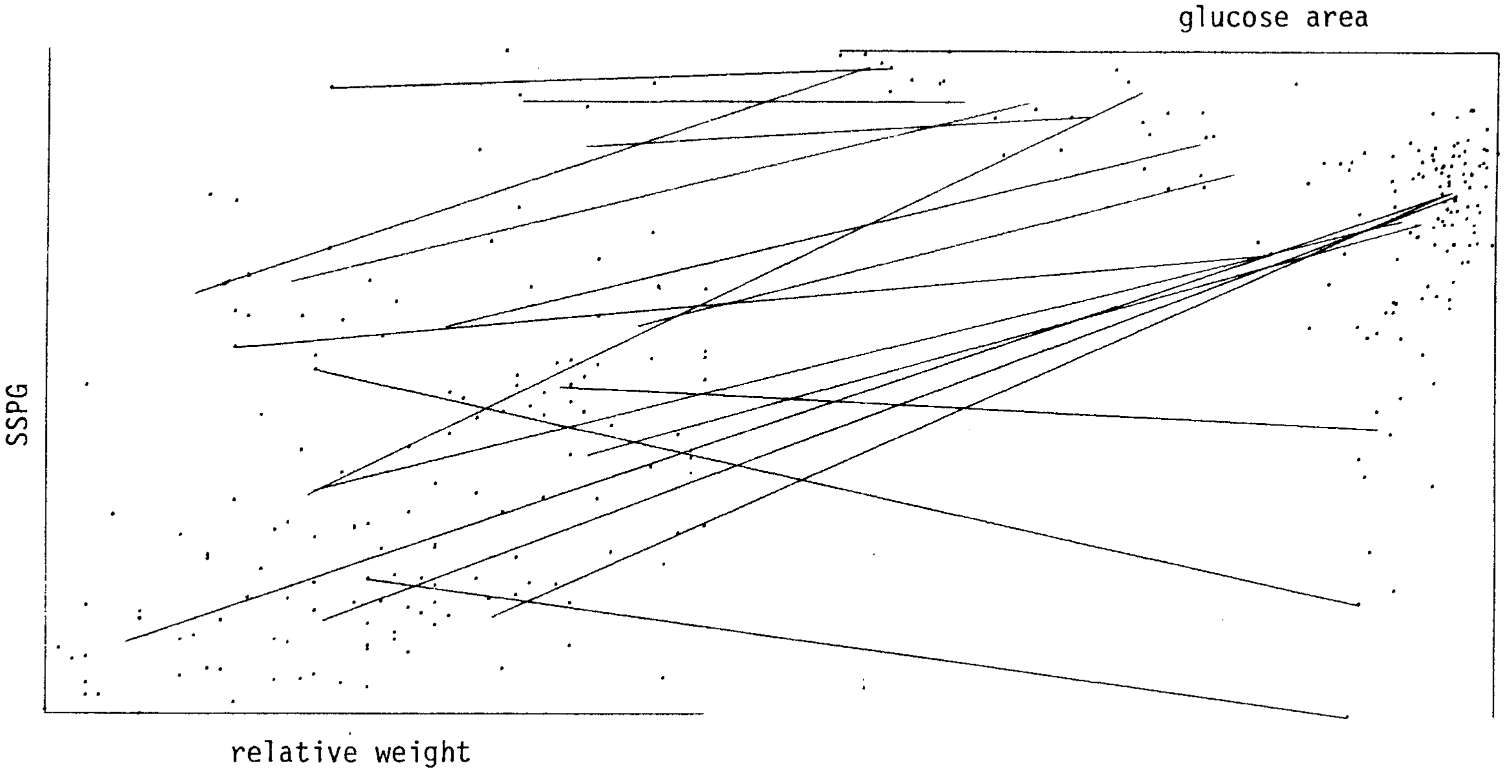


FIGURE 23

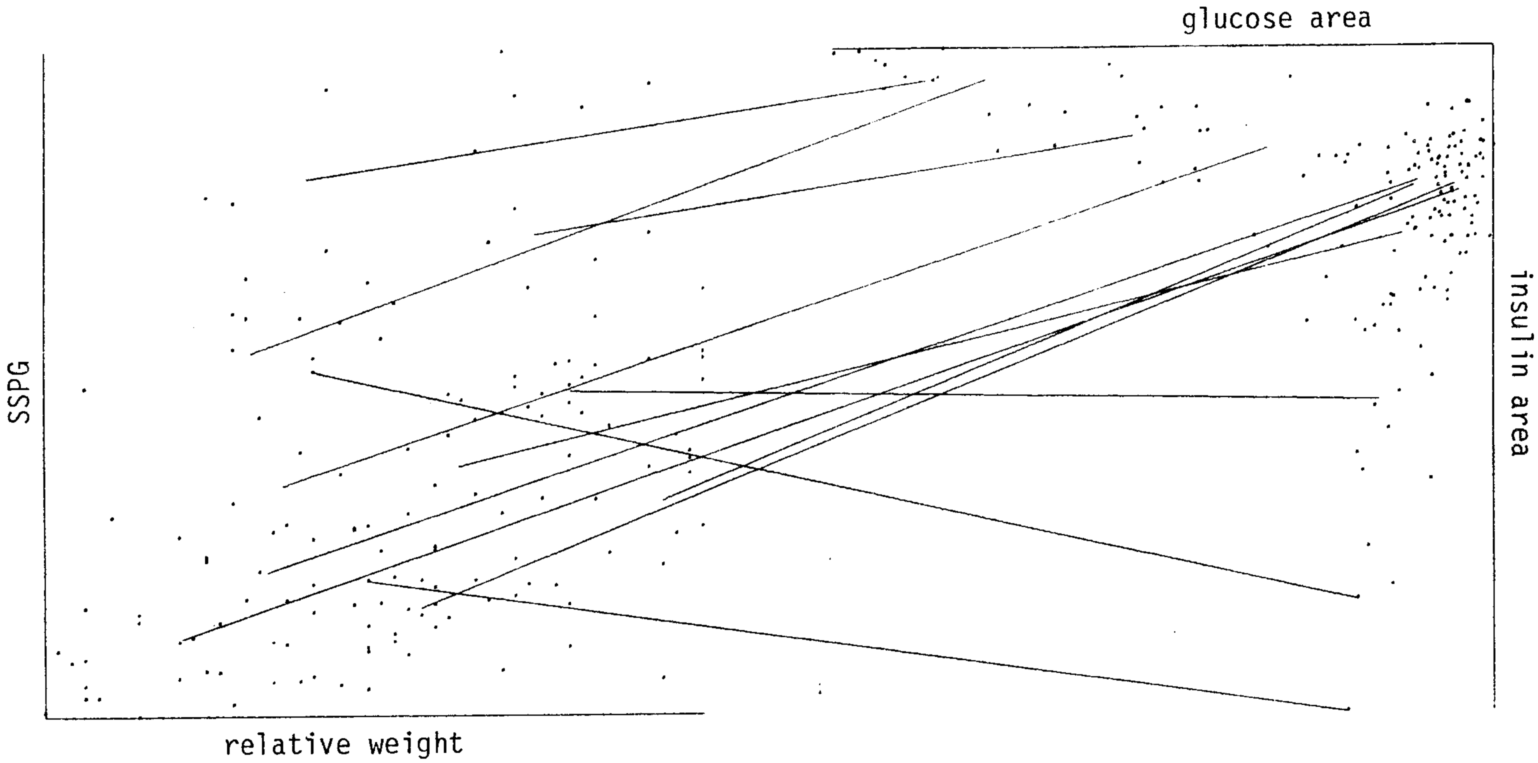


FIGURE 24

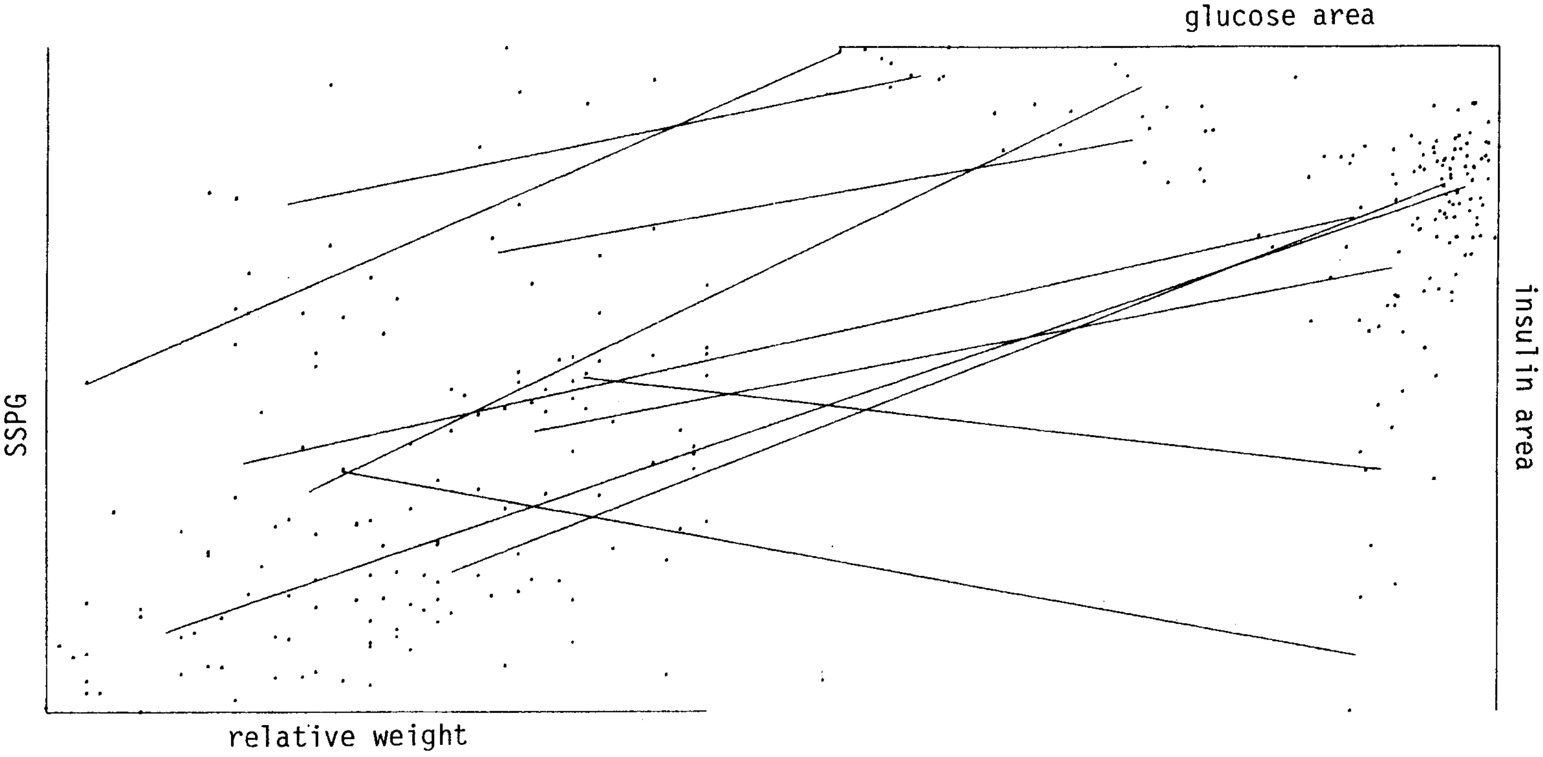
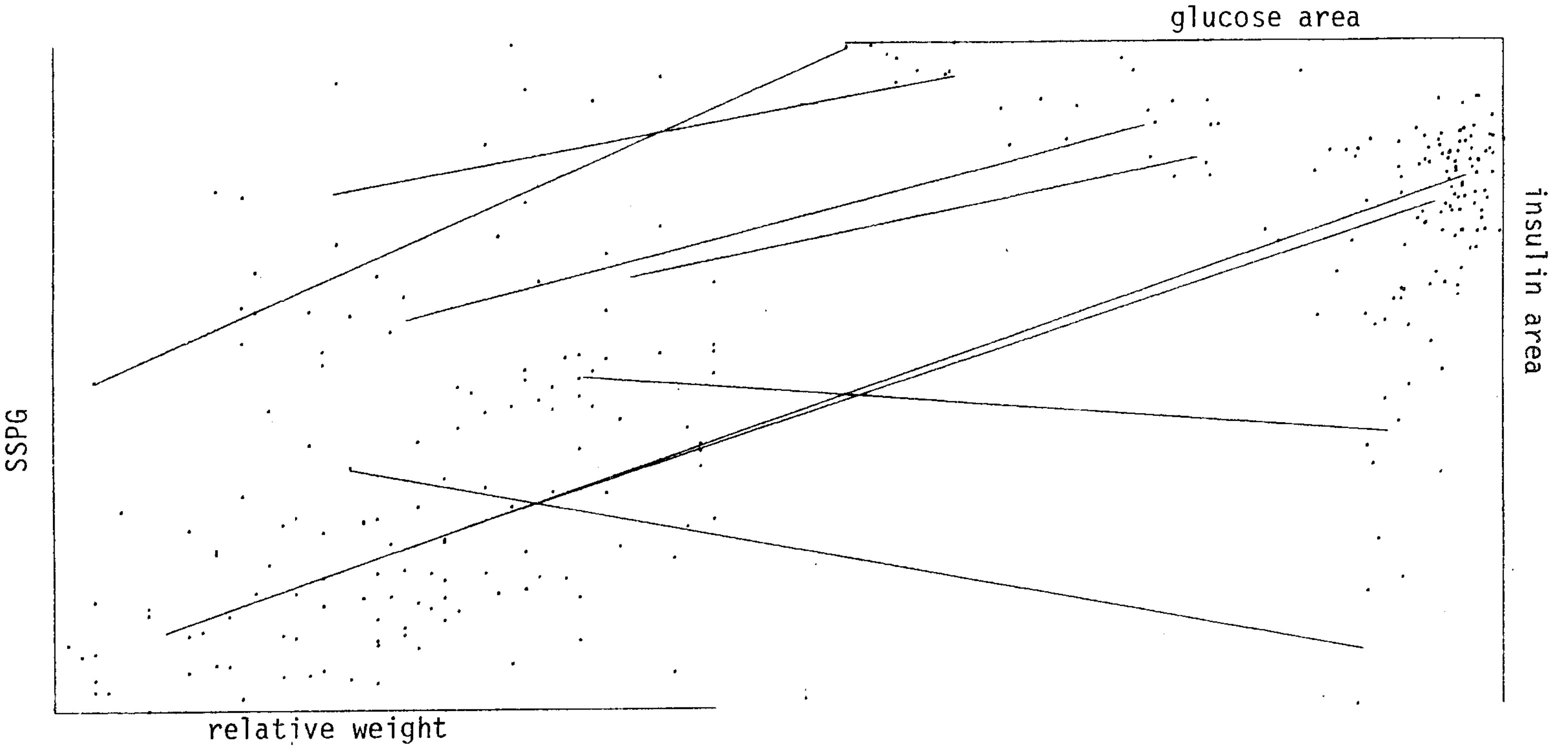


FIGURE 25



It seems clear that thinning allows a better view of the individual scatterplots. The thinning in Figures 17 through 22 seems acceptable. With more thinning, the lines tend to lose their usefulness.

It is straightforward to make an M and N plot by hand. Simply draw a pair of coordinate axis and start plotting points. Lines can be drawn either at random (say, with probability 1 in 5) or more systematically. One systematic approach is to break the (X,Y) and (S,T) planes into boxes. The first time a point is plotted in two boxes, the connecting segment is drawn. Only one segment is drawn for each pair of boxes. Thinning at random causes the density of lines to be proportional to the density of points. Systematic thinning causes the density of lines to be proportional to volume. In either case, we recommend starting with a low density of lines and adding lines systematically as they seem useful.

We next describe an efficient thinning algorithm for use on a computer. The algorithm we prefer (based on some experimentation) causes the density of lines to be proportional to volume (not density). Here is a rough description: suppose we want to make a 2 and 2 plot of n points. Divide the four-dimensional space up into "boxes". A single connecting line is drawn for each non-empty box. The line for each box may be taken to be the line representing the (four-dimensional) average point for that box.

More generally, suppose that the data consists of p-dimensional vectors (or points) $\underset{\sim}{x}_1,\underset{\sim}{x}_2,\ldots,\underset{\sim}{x}_n$. For expository purposes, suppose that the scale is chosen so that the data lies between zero and one in each coordinate. Suppose that the p-dimensional unit cube is divided into "boxes" of side h in each dimension. This makes $J = (1/h)^p$ boxes in all.

Each box can be indexed by a p-truple of integers. We will refer to the $\vec{\underset{\sim}{i}}$-th box, and use lexographic order on the box labels. To determine what box a point $\underset{\sim}{x}$ is in requires checking p inequalities. With this notation, we now present a semiformal description of our recommended algorithm. The main phase is to form a list containing

- the label of each non-empty box
- the sum of the vectors in that box
- the number of points in that box.

To begin, determine the box containing $\underset{\sim}{x}_1$. Call this $\underset{\sim}{i}_1$. The list begins

$$\underset{\sim}{i}_1, \underset{\sim}{x}_1, 1$$

the last component being a counter to indicate how many points are in box $\underset{\sim}{i}_1$. Next, consider $\underset{\sim}{x}_2$. If it is in box $\underset{\sim}{i}_1$, the first entry in the list is changed to

$$\underset{\sim}{i}_1, \underset{\sim}{x}_1 + \underset{\sim}{x}_2, 2$$

If $\underset{\sim}{x}_2$ is in box $\underset{\sim}{i}_2 \neq \underset{\sim}{i}_1$, the list contains

$$\underset{\sim}{i}_1, \underset{\sim}{x}_1, 1$$

$$\underset{\sim}{i}_2, \underset{\sim}{x}_2, 1 .$$

Continue, for each point $\underset{\sim}{x}_j$ determine what box $\underset{\sim}{i}_j$ contains $\underset{\sim}{x}_j$. If the label $\underset{\sim}{i}_j$ appears in the list, add $\underset{\sim}{x}_j$ to the second component of that list entry and increment the counter in the third component of the entry by 1. If $\underset{\sim}{i}_j$ does not appear in the list, insert $\underset{\sim}{i}_j$ in the list by binary insertion (using lexographic order on the labels). After processing all the points, the list contains the labels of the non-empty boxes, the sum of the points in each box, and the number of points in each box.

To draw an M and N plot using this list, make a single pass through the list computing the k-dimensional average for each box, and draw the line corresponding to this average point. If the number of non-empty boxes is B, the algorithm may be seen to run in

$$O\{pn(1+\log B)\} \text{ "operations".}$$

## Rotation

Some of the earlier plots have been rotated to make them less confusing. The next example is a 1 and 2 plot to illustrate the usefulness of rotation. Figure 26 is a plot of 100 triples (x,y,z) where y and z were chosen independently and uniformly in [0,1] and x = y+z. Look at the outside edge of the right-hand plot. Notice how the lines move down in x as y or z decrease. Next, look from top to bottom on the right-hand picture along any fixed line $\ell$ corresponding to $y \doteq$ constant. The lines in a neighborhood of $\ell$ have approximately constant slope. Recalling the discussion of 1 and 1 plots in Section 2, these observations suggest a linear relation between (y,z) and x.

Figure 26 is somewhat confusing to view because the lines change in slope and cross each other. Figure 27 shows a rotated 1 and 2 plot of the same data. Now the linear relation is striking.

FIGURE 26

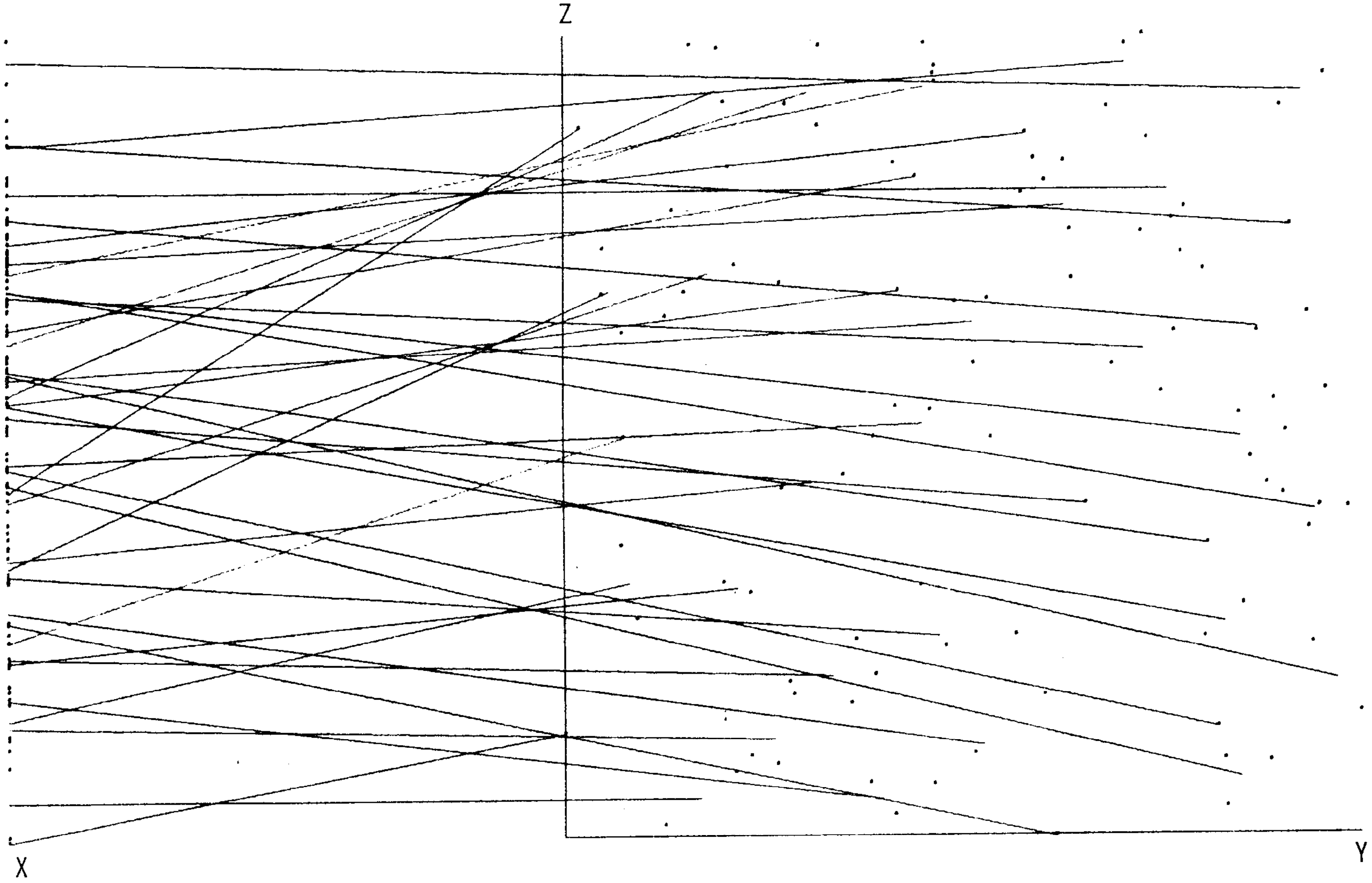
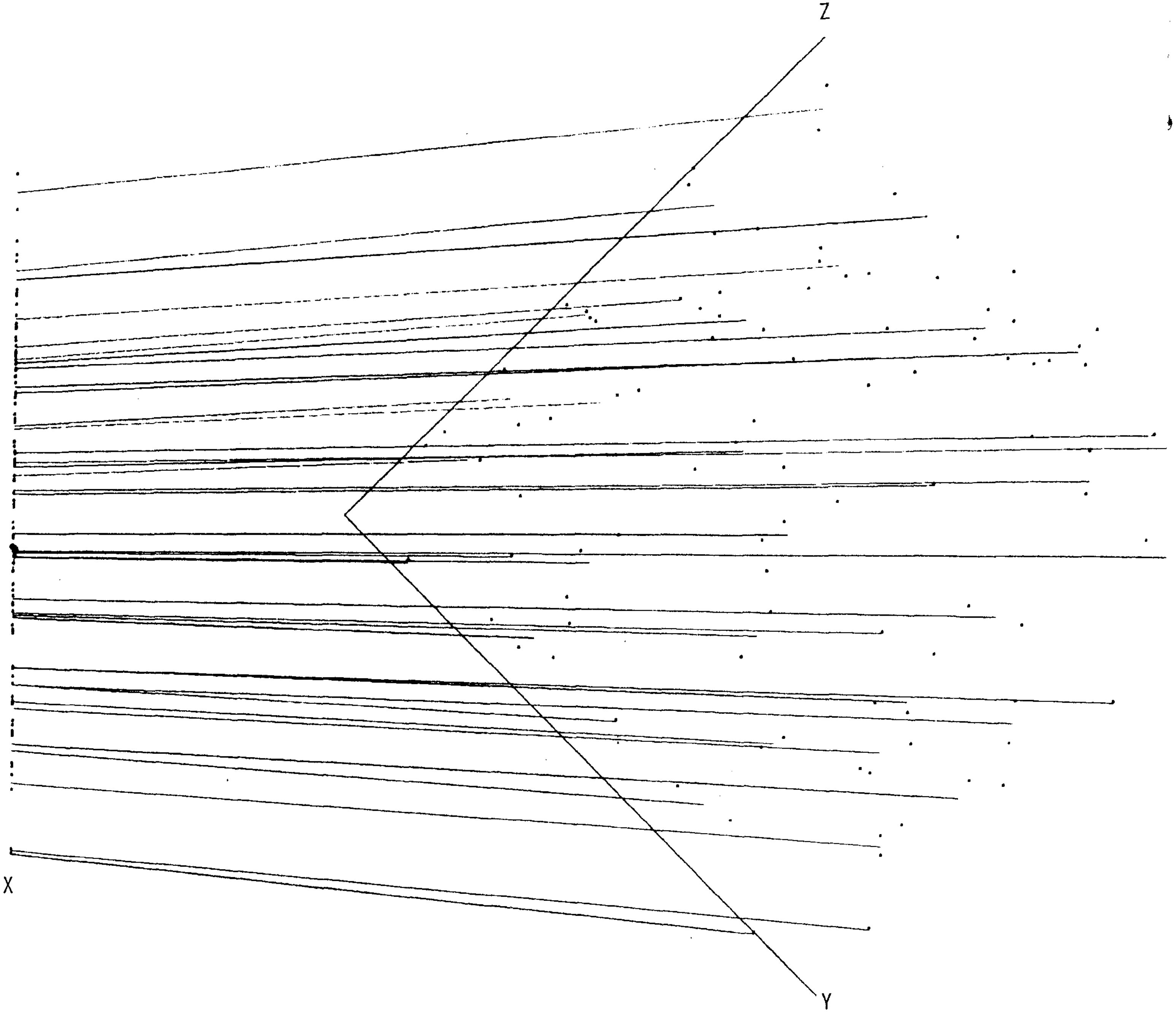


FIGURE 27



## Interactive Usage

When M and N plots are viewed on an interactive graphics device, the decisions about thinning and rotation are made by trial and error. A program making this easy to do was written at Stanford Linear Accelerator Center by Roger Chaffee. The program incorporates some other features which may be of interest.

- All points in the same box as a given point can be made brighter
- Lines can be easily added and deleted
- Lines can be made thicker to show another variable or the density of points in a box

## ACKNOWLEDGMENT

We thank Roger Chaffee for extensive programming and helpful suggestions. We thank Rupert Miller, Colin Mallows, and Aiden Moran for their help. We are grateful to the editors and publishers of Diabetologia for permission to reproduce Figure 11.

# REFERENCES

Brisson, D.W. (1978). Hypergraphics. Visualizing Complex Relationships in Art, Science, and Technology. Westview Press, Boulder, Colorado.

Eckhart, L. (1968). Four-Dimensional Space. (Translation by A.L. Bigelow and S.M. Slaby), Indiana University Press, Bloomington, Indiana.

Fisherkeller, M.A., Friedman, J.H., and Tukey, J.W. (1974). PRIM-9, An Interactive Multidimensional Data Display System. Stanford Linear Accelerator Pub-1408.

Friedman, J.H. and Rafsky, L.C. (1980). Graphics for the Multivariate Two-sample Problem. Submitted to Journal American Statistical Assn.

Gnanandesikan, R. (1977). Methods for Statistical Data Analysis of Multivariate Observations, Wiley, New York.

Griffin, H.D. (1958). Graphic Computation of Tau as a Coefficient of Disarray. Journal of the American Statistical Assn., 53, pp. 441-447.

Jessop, C.M. (1964). A Treatise on the Line Complex, Chelsea, New York.

Manning, H.P. (1960). The Fourth Dimension Simply Explained, Dover, N.Y.

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

Tukey, P.A. and Tukey, J.W.T. (1977). Handout for talk at Statistical Meetings, London, Ontario, November 1977.