GRAPHICAL DISPLAY OF SCATTER DATA USING THE STANDARD DEVIATION ELLIPSE

Charles E. Shipp and Charles G. Margolin
Northrop Corporation

ABSTRACT

A new procedure based on classical equations is used to graphically characterize large volumes of coordinate points without drawing them. The construct, called the Standard Deviation Ellipse, is described and derived, and examples show how it is easier to recognize subtle differences in the data and how it makes presentation of the data more esthetically pleasing.

A SAS procedure is described that can be used with SAS Proc PLOT or with SAS/GRAPH and two examples are given: one from CADAM statistics and one from professional sports showing height versus weight for various sports.

INTRODUCTION

The Standard Deviation Ellipse was conceived while working on employee-resource projects that typically would contrast several variables, two at a time, for different subgroups of the company population. In the past Proc FREQ, Proc MEANS, and Proc CORR have been used to make preliminary investigations. Proc STEPWISE was then used to refine a conceptual model and obtain regression coefficients.

During the process of the work, Proc PLOT was used to get scatter plots of the data, but with large amounts of data, overlaying is not possible, and running separate graphs for each case is voluminous and often not precise in showing differences. It also has the problem that outlyers distort the graph scaling.

It was thought at that time that if an ellipse could be drawn around the majority of the points in a widely accepted manner, then it would simplify visually extracting information from the data. In the derivation that follows, the most straight-forward method was sought and is thought to be found.

DERIVATION

Treat each class (subgroup of the population) separately, selecting the same two variables, and later superimposing the results. For each class the method finds (1) the center of mass for the (x,y) coordinate points, (2) the least-squares fit (LSF) line and a line perpendicular to it (crossing at the center of mass), (3) the four points one standard deviation away from the center of mass after projecting the points onto those lines, and (4) an ellipse that passes through those four points.

The equations, with illustrations, are given on the next page.

ADVANTAGES

Advantages of substituting a curve for the scatter points include (1) the curve is theoretically independent of the number of observations as the number increases, (2) with the resulting simplicity, several classes can easily be overlayed in the same graph, and (3) outlyers are used in the calculations but do not give problems to automatic scaling routines. (4) Most importantly, the Standard Deviation Ellipse provides a new standard of comparison of data showing dispersion and correlation.

EXTENSIONS

When developed, the new procedure will have additional optional keywords, such as provision for drawing a circle rather than an ellipse, supressing print of the output table of summary statistics, limiting the number of input points and specifying the selection rate, specifying the number of output points on the curve, and allowing multiple sets of coordinate data to be processed together, such as X*Y, S*T, U*V, ...

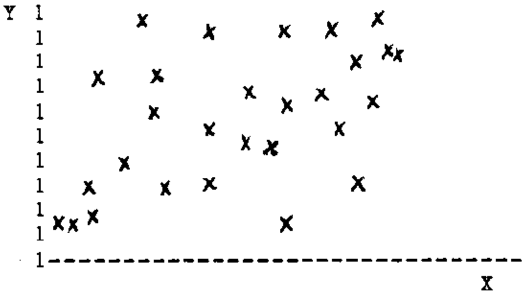
Extending the equations to three dimensions allows surface plotting (for hardware system with that capability) of ellipsoids, spheres, or n-point defined surfaces or patches. Extending to higher dimensions may be useful for tabulated values but cannot be easily plotted. (Color on the surface of an ellipsoid would, for example, provide 4-dimensional plots.)

SYNTAX

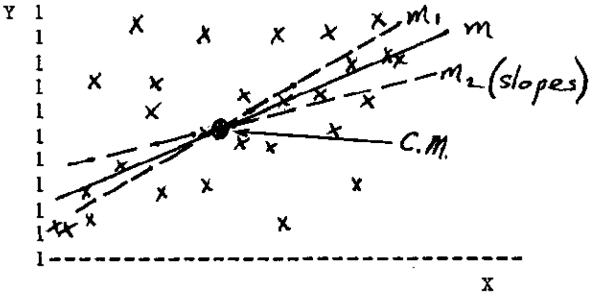
```
DATA;
  ...
PROC ELLIPSE;
  USE X Y;
PROC PLOT;
  PLOT Y*X;
```

PROGRESSIVE PICTURES

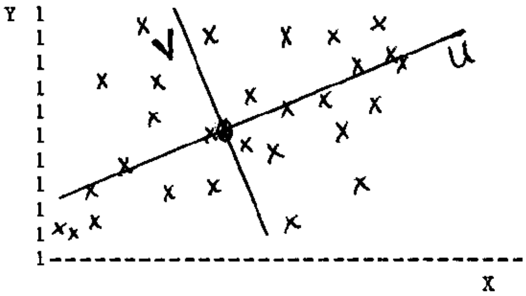
(1) X,Y Coordinate Points



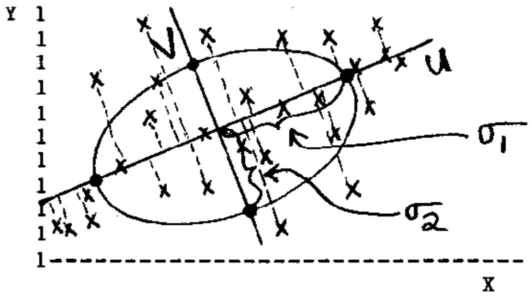
(2) LSF Line and Center of Mass



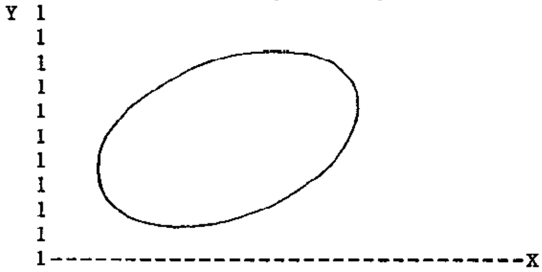
(3) New U,V Coordinate System



(4) Ellipse Production



(5) Ellipse Only



EQUATIONS

(1) The coordinate points of the population are (x_i, y_i), $i=1,n$, with optional weighting factors, (w_i), $i=1,n$.

(2) The center of mass

$$(\bar{x}, \bar{y}) = \left(\frac{\sum w_i x_i}{\sum w_i}, \frac{\sum w_i y_i}{\sum w_i}\right).$$

LSF (least squares fit) line: $y = mx + b$

where b is not used and

$$m_1 = \frac{n \sum x_i y_i - \sum x_i \sum y_i}{n \sum x_i^2 - \sum x_i \sum x_i}$$

when $w_i = 1$, $i=1,n$, and if not then $x_i \rightarrow w_i x_i$, $y_i \rightarrow w_i y_i$, and $n \rightarrow \sum w_i$, and $m_2(x,y) = m_1(y,x)$ for $m = [m_1 m_2]^{1/2}$.

(3) A new coordinate system U*V is defined via translation and the rotation to coincide with the CM and LSF line.

$$\begin{bmatrix} u_i \\ v_i \end{bmatrix} = \begin{bmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{bmatrix} \begin{bmatrix} x_i - \bar{x} \\ y_i - \bar{y} \end{bmatrix}$$

where $\theta = \mathrm{Tan}^{-1}(m)$.

(4) By projecting the points onto the U-axis and then onto the V-axis the illustrated ellipse can be defined as a standard deviation distance measured along each of those U*V axes from the center of mass

For the equations below let
σ = sigma = standard deviation

$$\sigma_1 = \left[\frac{\sum u_i^2}{n-1}\right]^{\frac{1}{2}}, \qquad \sigma_2 = \left[\frac{\sum v_i^2}{n-1}\right]^{\frac{1}{2}}.$$

For plotting purposes, points are now chosen on the ellipse as

$$-\sigma_1 < u_K < +\sigma_1$$

$$\text{and } v_K = \pm \sigma_2 \left[1 - \left(u_K/\sigma_1\right)^2\right]^{1/2}.$$

(5) Points to plot in the X*Y coordinate system are then found as

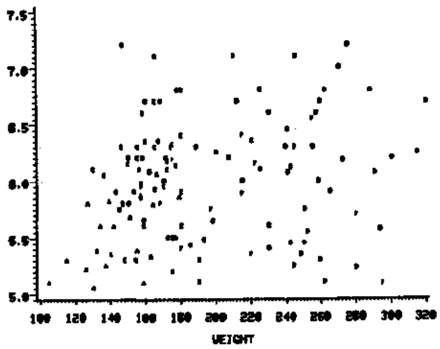
$$\begin{bmatrix} x_K \\ y_K \end{bmatrix} = \begin{bmatrix} \bar{x} \\ \bar{y} \end{bmatrix} + \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix} \begin{bmatrix} u_K \\ v_K \end{bmatrix}.$$

<u>EXAMPLES</u>

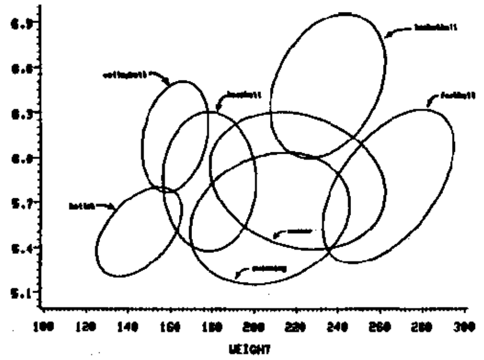
The following two examples show (x,y) coordinate information before and after using Proc ELLIPSE. The "before" plot in each case is a scatter plot using Proc GPLOT and the "after" plot in each case is a plot of the ellipse values generated by Proc ELLIPSE and plotted by Proc GPLOT. In each case the data is representative dummy data, and the plots are plotted on a Tektronix 4014 terminal.

<u>(1) Sports Example</u>

"BEFORE"



"AFTER"

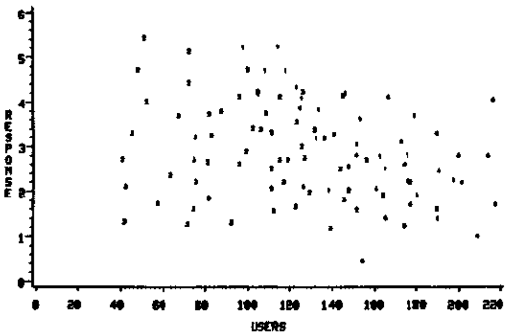


Key: A = ballet, B = basketball, C = baseball, D = football,
E = volleyball, F = soccer, G = swimming

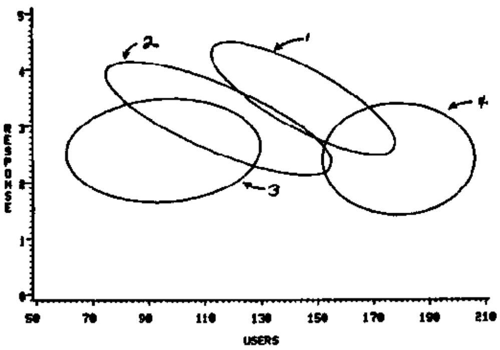
As can easily be seen in the right display, there is strong correlation between height and weight, most notably in cases B and F. The center of mass vs height is about as expected, and the outlyers are not show.

<u>(2) CADAM Example</u>

"BEFORE"



"AFTER"



CADAM response is seen to be strongly dependent upon the number of users with negative correlation, that is, the more users, the longer the response time for each of them. The usefulness of the graph is in noting that different computer system configurations may be in effect for the different weeks, (key: 1 = week 1, 2 = week 2, ..., 4 = week 4.)

<u>CONCLUSIONS</u>

Proc ELLIPSE, based on drawing the Standard Deviation Ellipse, for either one or several classes, allows the SAS user to process more concisely large volumes of scatter data. In the case of several classes, the procedure allows the user to superimpose the results on the same graph, and compare subtle differences which would otherwise be hard to detect and graphically display.

As statisticians and data processing personnel become familiar with this technique, the procedure will provide a standard of comparison for people working in the same field.

APPENDIX 1 MAIN SUBROUTINE

Input values (XD(I),YD(I)), I=1,ND represent the scatter points and (XP(I),YP(I)), I=1,NP are on the ellipse.

```
      SUBROUTINE SDE(XD,YD,ND,XP,YP,NP)
CALCULATES NP POINTS ON THE STANDARD DEVIATION ELLIPSE.
      DIMENSION XD(ND),YD(ND),XP(NP),YP(NP)
CALCULATE SUMS AND CROSS-PRODUCTS, AFTER INITIALIZATION.
      SUMX=0.
      SUMXX=0.
      SUMY=0.
      SUMXY=0.
      DO 10  I=1,ND
        SUMX=SUMX+XD(I)
        SUMY=SUMY+YD(I)
        SUMXX=SUMXX+XD(I)*XD(I)
        SUMXY=SUMXY+XD(I)*YD(I)
   10   CONTINUE
CALCULATE DENOMINATOR FOR CRAMER'S RULE, AFTER STORING C.M.
      XC=SUMX/FLOAT(ND)
      YC=SUMY/FLOAT(ND)
      DENOM=FLOAT(ND)*SUMXX-SUMX*SUMX
COEFFICIENTS OF LSF LINE CAN NOW BE CALCULATED.
      SLOPE=(FLOAT(ND)*SUMXY-SUMX*SUMY)/DENOM
      B=(SUMY*SUMXX-SUMX*SUMXY)/DENOM
CALCULATE MAJOR AND MINOR AXES OF S.D. ELLIPSE.
   40 T=ATAN(SLOPE)
      C=COS(T)
      S=SIN(T)
      DO 50  I=1,ND
C          TRANSLATE:
        XT=XD(I)-XC
        YT=YD(I)-YC
C          ROTATE:
        XE=+C*XT+S*YT
        YE=-S*XT+C*YT
C         ACCUMULATE:
        S1=S1+XE*XE
        S2=S2+YE*YE
   50   CONTINUE
      S1=SQRT(S1/FLOAT(ND-1))
      S2=SQRT(S2/FLOAT(ND-1))
      A=S1
      B=S2
CALCULATE NP POINTS ON THE ELLIPSE.
   70 AXISX=2.*A
      NPD2=NP/2
      DX=AXISX/FLOAT(NPD2)
      X=-A-DX
      DO 80  I=1,NPD2
        X=X+DX
        Y=+B*SQRT(1.0-(X/A)**2)
        XP(I)=XC+C*X-S*Y
        YP(I)=YC+S*X+C*Y
   80   CONTINUE
      X=A+DX
      NPD2P1=NPD2+1
      DO 90  I=NPD2P1,NP
        X=X-DX
        Y=-B*SQRT(1.0-(X/A)**2)
        XP(I)=XC+C*X-S*Y
        YP(I)=YC+S*X+C*Y
   90   CONTINUE
      RETURN
      END
```

APPENDIX 2 TEXT EDITING
(for this paper)

The reader may benefit from the methods used in preparing the manuscript for this paper. The first thing the authors did was experimentally find the equivalent column spacing of the photo master sheets. The tentative guidelines were entered on a full-screen terminal, (IBM 3278), listed on a typewriter terminal, (Agile A-1), and repeated until correct. The resulting guidelines are given below. The vertical column of periods is where to start typing, and so in full-screen edit they are erased when each respective line is begun. The vertical column of ones are not to be typed over. Additional comments are given below on creating and editing other pages, after which the guidelines are removed and the final version listed on the typewriter terminal. It was very interesting and useful to note that the 12-pitch button on the typewriter terminal reduced the 14 x 11 page down to an 8.5 x 11 page, which was convenient for distribution and review.

```
                                                                                                     1
         1         2         3         4         5         6         7         8         9
1234567890123456789012345678901234567890123456789012345678901234567890123456789012345678901234567890
-------------------+--------------------    +   -------------------+--------------------
```

FIRST DRAFT

The form displayed here was stored as PAGE0.DATA and when another page was started it was first obtained as

copy page0.data page3.data

and then words added using full-screen edit. This is often faster and more productive than writing on paper.

EDITING

Revisions are easily carried out in full-screen edit as long as no insertions or line deletions are required in either the left or right column, but not the other. If this occurs then one of the utilities listed next proved helpful.

.
.
.

(Fill in to here)

TEXT EDITING UTILITIES

SAS programs were written to perform the following utility applications for the specific format given here. They are as follows:

1) Move the left column to a separate data set

2) Move the right column to a separate data set

3) (After editing,) move two column data sets back together

4) Move a computer listing onto the page outline form

5) Another application that was not programmed is to take a long column of text and move it onto as many two-column pages as is required.

FINAL DRAFT

Final drafts were typed on the Agile terminal after the guidelines were removed.

Note that about 112 records of logical record length equal to 100 comprise this data set, of which 70 are used for text. The first line of text begins on the eighth record. By entering a '1' in the last column of the first and second record, the writer can perform both vertical and horizontal alignment for the paper cutter and the guidelines will fall right on the blue guidelines of the manuscript photo copy sheets. (Cut horizontally between the two 1's, and cut vertically so the bottom one is also just cut off.)

(This is the last line.)

GRAPHICS

Analysis Using Graphics in Communication