STATISTICAL COMPUTING

This department includes the two sections *New Development in Statistical Computing* and *Statistical Computing Software Reviews*; suitable contents for each of these sections are described under the respective section heading. Articles submitted for the department, outside the two sections, should not be highly technical and should be relevant to the teaching or practice of statistical computing.

A Graphical Display of Large Correlation Matrices

D. J. MURDOCH and E. D. CHOW

Large correlation matrices are hard to look at. In this article we present correlations as elliptical glyphs for a simple intuitive display of large matrices.

KEY WORDS: Correlation matrix; Ellipses; Graphical displays.

1. INTRODUCTION

As Hills (1969) said, "The first and sometimes only impression gained from looking at a large correlation matrix is its largeness." Sometimes this can be improved by suitable rounding (Ehrenberg 1977), but for very large matrices (e.g., 40×40 or larger), even rounding may not be enough. In this paper we present an intuitive display of large correlation matrices that is compact and easily implemented on a laser printer, in which numerical entries are replaced with elliptical glyphs.

Our idea of substituting ellipses for correlation matrix entries does not appear to have been suggested previously, although many authors [e.g., Anderson 1957, Tukey and Tukey 1981, and Cleveland 1985] have used symbols to display numerical values in other contexts. Hills (1969) dealt with the largeness of the correlation matrix by displaying a QQ plot of the entries, rather than attempting to display the matrix itself. Ling (1973) used overtyped characters to encode correlations by gray levels.

Section 2 of this article describes the display; Section 3 describes its implementation using PostScript (Adobe Systems Incorporated, 1990) and S (Becker, Chambers, and Wilks 1988), and we conclude by mentioning some other applications of the techniques.

2. THE ELLIPSE MATRIX

There are a large number of ways to represent quantitative variables graphically, many of which are difficult to decode accurately (Cleveland 1985). Fortunately, in correlation matrices arising as observed correlations among a large number of variables or as correlations of parameter estimates from the fit of a large model, accurate numerical decoding is not as essential as is an impression of whether correlations are physically significant.

In the case of observed data an excellent graphic display for up to 10 or 20 variables is a scatterplot matrix; similarly, a scatterplot matrix of bootstrap parameter estimates may be used to show correlations among a relatively small number of parameters. However, when the number of variables is very large scatterplot matrices are impractical, and a simpler summary is needed. One possibility is a table of highly rounded correlations (Figure 1, left), but we have found that a plot of ellipses is a more effective way to view large matrices of correlations (Figure 1, right).

The ellipses that we plot are shaped to be contours of a bivariate normal distribution with unit variances and correlation ρ, with the contour tangent to a unit square. The ellipse is plotted using the points

$$(x, y) = (\cos(\theta + d/2), \cos(\theta - d/2)) \qquad (1)$$

where $\theta \in [0, 2\pi]$ and $\cos(d) = \rho$.

It is helpful to shade the ellipses; this assists the viewer in finding especially high or low correlations. However, solid black ellipses lead to visual vibrations that are distracting (Tufte 1990); a shade of gray is better. Outlines are necessary when plotted on a black and white laser printer, as the dithering used to simulate gray on these printers tends to distort the shape of small ellipses.

Choosing an appropriate order for the rows and columns of the matrix can be helpful, as Ehrenberg (1977) noted for tables. For example, Figure 2 shows the correlation matrix resulting from a nonlinear fit to an age–period–cohort model involving a total of 42 variables (including an intercept term). The variables have been ordered into four groups: the intercept term, 13 age group contrasts, 7 period contrasts, and 21 cohort contrasts. The contrasts are all Helmert contrasts presented in their natural order.

From this plot it is clear that the terms for the older age groups are very highly correlated, as are the later period terms and the middle cohort terms. There are also various

D. J. Murdoch is Associate Professor and E. D. Chow is Assistant Professor, Department of Mathematics and Statistics, Queen's University, Kingston, Ont., Canada, K7L 3N6. Financial support for this work was provided by the Natural Sciences and Engineering Research Council of Canada. The authors thank Wayne Oldford for his many helpful comments and suggestions.

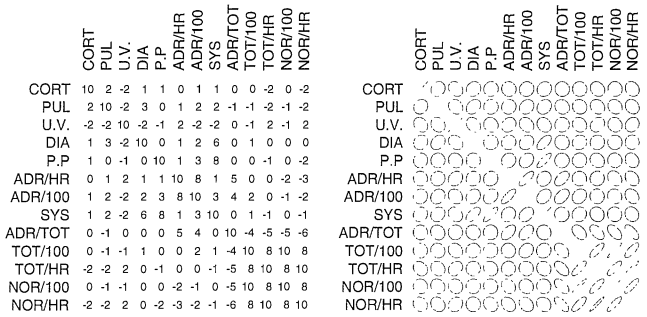

Figure 1. Correlations of 13 Physiological Parameters. On the left are shown the correlations rounded to one decimal place (with the decimal point omitted); on the right is the ellipse matrix. (Data from Hills 1969.)

strong cross-correlations between the groups of variables. The final cohort term is anomalously uncorrelated with everything else; this is likely a numerical problem caused by the overparametrized model.

When the problem does not suggest a natural ordering for the variables, sometimes the matrix itself can be used. For example, Figure 1 presents Hills' (1969) correlation matrix of physiological measurements on medical students with the variables ordered so that the average squared correlation is increasing in each row and column; this presentation helps the viewer to recognize the relatively low correlations between most variables, but the high correlations among the last four. The patterns of correlations are much more striking in the ellipse matrix than in the tabular presentation.

3. IMPLEMENTATION

One approach to implementation of the ellipse matrix is to treat each ellipse as a character in a font of graphical shapes. The ellipse matrix is then built by invoking commands to place text on a plot, instead of a series of calls to graphics primitives. This approach has the following advantages:

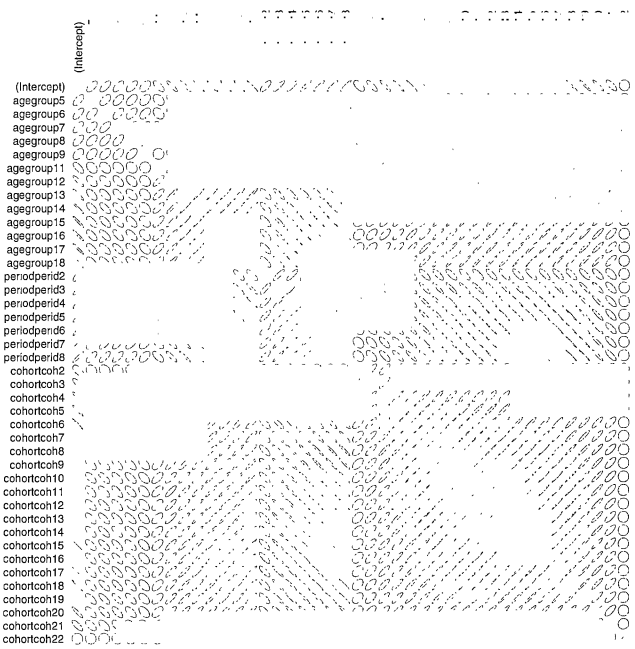

Figure 2. Approximate Correlation Matrix from Fit of 42 Variable Age—Period—Cohort Model.

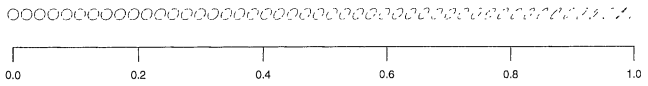

Figure 3. The 48 Glyphs Used for Nonnegative Correlations. The glyphs for negative correlations are reflections of positive ones.

- Characters are easy to display. Displaying a character from a font is a simpler task than connecting a series of points calculated from the parametric description (1). Ellipse-drawing primitives exist for some graphics devices, but do not always allow a slanted major axis.
- Correlation ellipses form a natural font. Only a finite number of correlation ellipses are distinguishable by the human eye, so there is little precision lost in restricting our ellipses to a finite "alphabet."

To construct a font of ellipses we employed PostScript, a page description language ideally suited to font creation and manipulation. The font machinery of PostScript is such that characters are printed very efficiently; as a result large matrices of characters are as quickly rendered as small ones. Ellipses, in particular, are well suited to implementation in a PostScript font—we were able to completely define an Ellipse font in less than 500 bytes of code. The ellipses themselves are efficiently described as scaled and rotated circles.

Our Ellipse font encoded the printable ASCII characters as 95 equally spaced correlations from -1 to 1 (Figure 3).

The figures in this paper were prepared in PostScript, calling on the above font to supply the ellipses. To generate the PostScript code we used the S-PLUS implementation of S; it would suffice to use any statistical package whose interface to PostScript is flexible enough to allow the introduction of a new, user-defined, font.

To display an ellipse matrix on-screen, one can construct ellipses from graphics primitives, for example, via the parametric description (1). Alternatively, one can generate PostScript code to be displayed by a suitable previewer. Both of these options are available in our S-PLUS implementation of the ellipse matrix procedure, available from Statlib. For instructions on how to obtain copies, send email to `statlib@lib.stat.cmu.edu`.

4. CONCLUSIONS

Use of ellipses as glyphs for correlations has the advantage over other symbols of immediate recognizability and connection with the quantity being plotted. They would also be a natural choice for plotting partial correlations. However, it is difficult for most people to decode a numerical correlation from the ellipse shape; correlations less than .5 in absolute value are especially hard to distinguish. For example, the `PUL` by `DIA` correlation in Figure 1 is .28, but it does not appear strikingly different from the .04 `PUL` by `P.P.` correlation next to it. (On the other hand, the similarity of the ellipses may serve to point out how weak a correlation of .28 really is.) If numerical decoding were the aim, then the use of Framed Rectangles (Cleveland 1985) might be more appropriate. The same technique of imple-

mentation as a PostScript font would be possible with those glyphs as well.

Fonts have always had currency in statistical graphics as a source of plotting symbols. With the arrival of powerful page description languages such as PostScript it is a simple matter to construct distinctly nonalphabetic character sets conveying numeric information.

[Received August 1994. Revised September 1995.]

REFERENCES

Adobe Systems Incorporated (1990), *PostScript Language Reference Manual* (2nd ed.), Reading, MA: Addison-Wesley.

Anderson, E. (1957), "A Semigraphical Method for the Analysis of Complex Problems," in *Proceedings of the National Academy of Sciences*, 13, 923–927; reprinted (1960) in *Technometrics*, 2, 387–391.

Becker, R. A., Chambers, J. M., and Wilks, A. R. (1988), *The New S Language*, Pacific Grove, CA: Wadsworth.

Cleveland, W. S. (1985), *The Elements of Graphing Data*, Pacific Grove, CA: Wadsworth & Brooks/Cole.

Ehrenberg, A. S. C. (1977), "Rudiments of Numeracy," *Journal of the Royal Statistical Society*, Ser. A, 140, 277–297.

Hills, M. (1969), "On Looking at Large Correlation Matrices," *Biometrika*, 56, 249–253.

Ling, R. F. (1973), "A Computer Generated Aid for Cluster Analysis," *Communications of the ACM*, 16, 355–361.

Tufte, E. R. (1990), *Envisioning Information*, Cheshire, CT: Graphics Press.

Tukey, P. A., and Tukey, J. W. (1981), "Graphical Display of Data Sets in 3 or More Dimensions," in *Interpreting Multivariate Data*, ed. V. Barnett, Chichester: John Wiley, pp. 245–275.

The Convergence of Efroymson's Stepwise Regression Algorithm

Alan J. MILLER

The stepwise regression algorithm that is widely used is due to Efroymson. He stated that the F-to-remove value had to be not greater than the F-to-enter value, but did not show that the algorithm could not cycle. Until now nobody appears to have shown this. To prove that the algorithm does converge, an objective function is introduced. It is shown that this objective function decreases or can occasionally remain constant at each step in the algorithm, and hence the algorithm cannot cycle provided that Efroymson's condition is satisfied.

KEY WORDS: Stepwise regression; Stopping rule.

1. EFROYMSON'S ALGORITHM

The algorithm that is commonly called just stepwise regression was first presented by Efroymson (1960). It is basically as follows:

1. Enter into the (linear) regression model any variables that are to be "forced in."
2. Find the variable from those not in the model but available for inclusion that has the largest F-to-enter value. If it is at least as great as a prespecified value, F_{in}, then add the variable to the model. Stop if no variable can be added.
3. Find that variable among those in the model, other than those forced in, that has the smallest F-to-remove value. If it is less than a prespecified value, F_{out}, then drop the variable from the model. Repeat this step until no further variables can be dropped; then go to step 2.

Alan J. Miller is Honorary Research Fellow, CSIRO Division of Mathematics and Statistics, Clayton, Vict. 3169, Australia.

Efroymson says that F_{out} must be not greater than F_{in}, and suggests a value of 4.0 for both, but does not comment on whether or not the process terminates. As the algorithm is used thousands of times a day, there can be little doubt that it does converge, but nobody seems to have presented a proof.

In the above process, if RSS_p is the residual sum of squares for a model with p parameters, then the F-to-enter statistic is

$$\frac{\text{RSS}_p - \text{RSS}_{p+1}}{\text{RSS}_{p+1}/(n-p-1)},$$

and similarly the F-to-remove statistic is

$$\frac{\text{RSS}_{p-1} - \text{RSS}_p}{\text{RSS}_p/(n-p)}.$$

2. PROOF OF CONVERGENCE

For any subset of variables, S, consider the objective or Lyapunov function

$$L(S) = \text{RSS}_p \prod_{i=1}^{p} \left(1 + \frac{F}{n-i}\right), \tag{1}$$

where n is the number of cases, and F is any value such that $F_{\text{out}} \le F \le F_{\text{in}}$. In most practical cases the p parameters will consist of an intercept and $(p-1)$ regression coefficients.

As there is only a finite number of different subsets of variables to be considered, the algorithm must either terminate or cycle. The above objective function can be shown to decrease every time that a variable is added or deleted, and therefore the algorithm cannot cycle. Thus the Efroymson stepwise regression algorithm can be viewed as a heuristic algorithm to minimize the objective function $L(S)$. Al-