Behavior Research Methods, Instruments, & Computers
1990, 22 (2), 176-178

Dynamic graphics in the exploratory analysis of multivariate data

FRANK M. MARCHAK and DAVID A. WHITNEY
The Analytic Sciences Corporation, Reading, Massachusetts

Reported here is a study in which the dynamic depiction of three-dimensional data is compared with traditional static scatterplots, particularly with respect to the ability of observers to extract cluster information from multivariate data sets.

Scientific visualization involves applying computer graphics and image processing techniques to provide insight through visual methods by studying "those mechanisms in humans and computers which allow them in concert to perceive, use and communicate visual information" (McCormick, DeFanti, & Brown, 1987, p. 3). It has emerged as an important tool by allowing interactive, real-time use of computer graphics to view and interpret large data sets in a variety of domains. Applications range from visual modeling of the structure of physical objects such as the brain to models of phenomena such as neuron firing.

The statistical analysis of multivariate data provides an ideal domain for the application of scientific visualization techniques. The use of graphical methods in exploratory data analysis has been advocated by a variety of investigators (e.g., du Toit, Steyn, & Stumpf, 1986; Tukey, 1977). Recent advances in computer technology have brought the ability to produce complex dynamic graphics within the reach of anyone with access to a personal computer. However, the literature on graphical methods provides little guidance on what dynamic techniques are most effective in revealing the structure present in a data set or how they compare with traditional static techniques.

Most work in the area of graphical perception has been done by statisticians rather than psychologists (e.g., Chambers & Kleiner, 1982; Cleveland & McGill, 1984; Grotch, 1983). Cleveland and McGill (1984) investigated the effectiveness of a variety of perceptual features used to extract quantitative information from graphs, including length, direction, area, volume, and color. Using psychophysical methods, they rated the features from most to least accurate, showing that subjects were more accurate at judging length or direction than area and more accurate at judging area than volume or color. Unfortunately, their studies were confined to two-dimensional graphs, and the findings are not readily extendible to multidimensional representations.

Grotch (1983) investigated several techniques to aid in the interpretation of three-dimensional data displays. He found that factors such as depth cues and the connection, projection, and flickering of points in space all aided the perception of any structure present in the data set. In contrast, representations such as contour and surface plots failed to provide a good quantitative feel for the data.

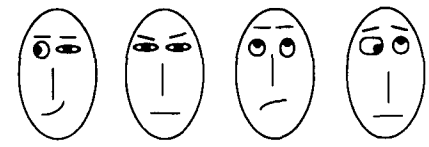
One problem inherent in multidimensional data is the adequate represention of a large number of variables in two, or at most three, dimensions. A variety of techniques have been proposed, including Chernoff faces (Chernoff, 1973), harmonic function plots (Andrews, 1972), and three-dimensional box plots (Hartigan, 1975) (Figure 1). Brown (1985) compared these various techniques and found that they vary in their ability to convey structure in the data. The variation is primarily a function of the familiarity with representation method.

While these methods allow the representation of data sets with up to 20 dimensions, many techniques for analyzing multivariate data define a smaller set of derived variables to focus on certain properties of the original data (Chambers & Kleiner, 1982). Principal components analysis, discriminant analysis, canonical correlation, factor analysis, and multidimensional scaling are often used to reduce the number of variable dimensions while still providing a realistic representation of the data. However, reduction to two dimensions usually preserves only 40% to 70% of the variance of the original data. The addition of a third dimension often increases the variance accounted for to 70% to 90%, and it also provides pattern-recognition information, particularly for cluster analysis (Grotch, 1983).

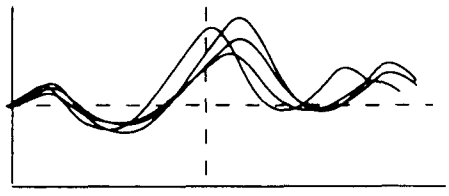
Given that a three-dimensional data plot is sufficient and desirable, the problem remains of trying to fit a basically two-dimensional plot to data with more than two dimensions. One possibility is multiple scatterplots of all pairs of the original variables (Figure 2). This method tends to obscure any structure that might otherwise be obvious in the full three-dimensional rendering. Given the recent advances in scientific visualization techniques, it is now possible to present realistic three-dimensional plots of multivariate data sets in real time on any personal computer.

In the present study, we compare dynamic depiction of three-dimensional data to traditional static scatterplots, particularly with respect to the ability of observers to extract cluster information from multivariate data sets. Fol-

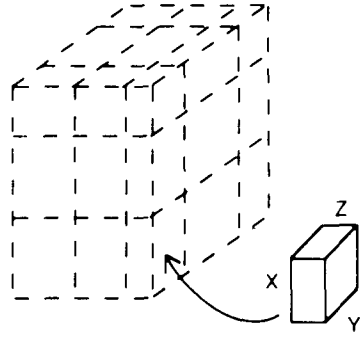
Correspondence may be addressed to Frank M. Marchak, The Analytic Sciences Corporation, 55 Walkers Brook Road, Reading, MA 01867.



Chernoff Faces



Harmonic Plots



Three-Dimensional Box Plot

Figure 1. Multidimensional graphic techniques.

lowing the work of Brown (1985) and Freni-Titulaer and Louv (1984), who examined cluster perception in static graphics, three issues are explored. First, is there a difference between dynamic and static graphics in the ability of observers to recover previously defined data clusters? Second, do the dynamic graphic methods of rotation and animation differ in their effects on cluster perception? Finally, how do differences in cluster distances, as measured with a Euclidean metric, affect cluster perception?

METHOD

Subjects

24 engineers from The Analytic Sciences Corporation participated. All had normal or corrected vision.

Design

A 3 × 3 × 3 factorial design was used: three graphic types (dynamic rotation, dynamic animation, and static), three cluster distances (near, far, and none), and three numbers of clusters (1, 2, and 3). Presentation order of graphic type was balanced across subjects, while cluster distances and cluster numbers were randomly distributed for each subject.

Stimuli

Nine sets of four-dimensional clusters were constructed for the display, using three different cluster groupings and three different cluster distances. In the three-group cluster set, each cluster was composed of 30 points sampled from a normal distribution with specified 4 × 4 correlation matrices with standard deviations of 5. In the two- and one-group cluster sets, clusters were composed of 45 and 90 points, respectively, sampled from the same distribution as above.

Each set of points had a variable mean, which was adjusted to change the distance between clusters. Distance was measured using the standard Euclidean metric, such that in the near condition the distance was 7, in the middle condition the distance was 12, and in the far condition the distance was 17. The fourth dimension in the data sample served as the animation variable for the three-dimensional animation display.

Apparatus

A Macintosh computer was used to display the stimuli. They were presented using MacSpin, a dynamic graphical data-analysis package produced by D2 Software (see Donoho, Donoho, & Gasko, 1988). Two MacSpin features were used: rotation and animation. Rotation involved spinning the data set around each of the three axes. Animation allows choosing a fourth variable that controls a threshold for displaying the other three variables. The animation variable controls which observations are visible, as determined by its range. Motion in the display shows variation in the data due to the animation variable.

Procedure

The nine data sets were entered into MacSpin and were printed as three two-dimensional scatterplots for each data set. In the rotation condition, each data set was rotated

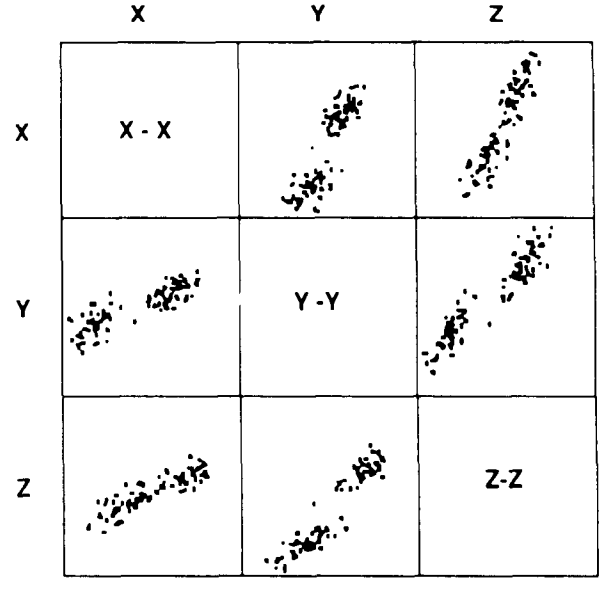


Figure 2. Multidimensional scatterplots.

360° about each of the three axes. In the animation condition, each data set was animated by its fourth variable. As the range of the animation variable increased, the threshold determined which data points were made visible. The static condition involved presenting three scatterplots of the *x*-*y*, *y*-*z*, and *x*-*z* dimensions.

The subjects viewed each data set and were asked to determine whether there were one, two, or three data clusters in each set. Responses were recorded on a response sheet. No time limit was imposed on responding.

RESULTS AND DISCUSSION

Total number of correct responses across presentation method showed no significant difference, with a mean number correct of 5.5 for rotation, 4.5 for animation, and 5.1 for static. Collapsing the data across number of clusters, repeated measures analyses of variance were performed on cluster distance and presentation method. For cluster distance, $F(2,46) = 122.96$, $p = .000$, with means of 1.10, 1.43, and 2.54 for near, middle, and far distances, respectively. Using Newman-Keuls, all pairwise comparisons were significant to $p < .05$. For the presentation methods, $F(2,46) = 4.87$, $p = .012$, with means of 1.85, 1.51, and 1.71 for rotation, animation, and static conditions, respectively. Newman-Keuls showed a significant difference ($p < .05$) between rotation and animation.

The findings suggest that for these data and presentation methods, there is minimal benefit, at least in total number correct, in the dynamic presentation of scatterplots as opposed to the static presentation. Rotation proved to be the best presentation technique, but it was not significantly better than static presentation. The reasons for this might be that the experimental conditions tended to constrain the subjects' ability to interact with the data. In a real analysis situation, one would use rotation and animation jointly, choosing different views for animation and generally examining the data more thoroughly. The fact that animation was limited to only one rotational viewpoint meant that if the clusters were not well separated from that view, animation would serve little benefit. This might explain its poor performance.

An interesting (but not statistically significant) finding was that in the single cluster conditions, subjects tended to be less accurate in determining the number of clusters with rotation than they were with either static or animated presentations. This suggests that rotation might lead to overreporting structure in the data that is not there in reality.

Overall, it seems that given the current data, dynamical presentation techniques do not necessarily aid in identification of structure in scatterplots. As mentioned above, the constrained nature of the tasks may have affected this outcome. Further examination, using more realistic combinations of presentation techniques and allowing subject interaction, may provide more information about the value of these representations.

REFERENCES

ANDREWS, D. F. (1972). Plots of high dimensional data. *Biometrics*, **28**, 125-137.

BROWN, R. L. (1985). Methods for the graphic representation of systems simulated data. *Ergonomics*, **28**, 1439-1454.

CHAMBERS, J. M., & KLEINER, B. (1982). Graphical techniques for multivariate data and for clustering. In P. R. Krishnaiah & L. N. Kanal (Eds.), *Handbook of statistics* (Vol. 2, pp. 209-244). New York: North-Holland.

CHERNOFF, H. (1973). Using faces to represent points in K-dimensional space graphically. *Journal of the American Statistical Association*, **68**, 361-368.

CLEVELAND, W. S., & MCGILL, R. (1984). Graphical perception: Theory, experimentation and application to the development of graphical methods. *Journal of the American Statistical Association*, **79**, 531-554.

DONOHO, A. W., DONOHO, D. L., & GASKO, M. (1988). MacSpin: Dynamic graphics on a desktop computer. In W. S. Cleveland & M. E. McGill (Eds.), *Dynamic graphics for statistics* (pp. 331-351). Monterey, CA: Wadsworth.

DU TOIT, S. H. C., STEYN, A. G. W., & STUMPF, R. H. (1986). *Graphical exploratory data analysis*. New York: Springer-Verlag.

FRENI-TITULAER, L. W. J., & LOUV, W. C. (1984). Comparison of some graphical methods for exploratory multivariate data analysis. *American Statistician*, **38**, 184-188.

GROTCH, S. L. (1983, November). Three-dimensional and stereoscopic graphs for scientific data display and analysis. *IEEE Computer Graphics & Applications*, **3**(11), 31-43.

HARTIGAN, J. A. (1975). *Clustering algorithms*. New York: Wiley.

MCCORMICK, B. H., DEFANTI, T. A., & BROWN, M. D. (Eds.). (1987). Visualization in scientific computing [Special issue]. *Computer Graphics*, **21**(6).

TUKEY, J. W. (1977). *Exploratory data analysis*. Reading, MA: Addison-Wesley.