Comment

J. C. Gower

There are several levels on which this paper could be discussed—the development of graphical methods, interactive data analysis, the OMEGA software, the particular analysis presented or substantive issues of dyestuff manufacture. One hardly knows where to begin, so my comments will refer to all of the above, except the dyestuffs, of which I know nothing.

We have only to recall Fisher's well-known statement "I have learned most of my statistics at the machine" to realize that exploratory data analysis is no new thing. Of course, in precomputer days, computationally extensive methods of exploring data were out of the question; it is said that before 1955 *Biometrika* had never published a paper with a multiple regression containing more than five independent variables, and perhaps this was not a bad thing. The original batch mode of running computers did not encourage exploratory analyses; neither did statistical packages. For at least 25 years some workers, notably John Tukey and his associates at AT&T Bell Laboratories, have developed various brands of EDA, but it is only with the recent availability of cheap powerful workstations with high-quality graphics that these methods are beginning to be used routinely. Once again, statisticians can work closely with their data, but now with vastly increased computing power associated with excellent graphics.

We now have the computational technology, but do we have the software? I believe not. Weihs and Schmidli have made a brave attempt with their OMEGA pipeline. However, the statistical facilities it contains seem limited and do not contain many recent advances; the same applies to the software design. Thus statistically I would expect an interactive system to have convenient methods for adding/deleting variables/samples, and OMEGA does not seem to have these. Also I would hope to be able to handle more structured samples. Perhaps it is a little early to expect much of the work developed by the Gifi group in Leiden to be included, but surely much of the Multiple Correspondence Analysis (MCA)/Homogeneity Analysis/Fisher's Method of Optimal Scores should be available. MCA allows categorical data to be handled and, indeed, is effectively a categorical variables parallel of the biplot technique for quantitative variables; recent work allows quantitative and categorical variables to be analyzed simultaneously. Weihs and Schmidli rightly draw attention to the importance of scales of measurement in multivariate methods. MCA allows quantitative variables to be categorized and then scored on new quantitative scales, from which nonlinear transformations may be constructed; if one wishes a smooth transformation, then spline functions may be fitted. Similar types of information can be found from the monotone transformations of nonmetric scaling. Surely these approaches to seeking simplicity through dimension-reducing transformations are preferable to the ad hoc trial of standard transformations, especially when these seem to be applied en bloc to all variables.

Turning to the software design of OMEGA, its structure as given in their Figure 1 seems less flexible than desirable. One would like to repeat analyses after dynamically removing samples, or transforming variables, as guided by informative plots. At least this requires a considerable element of feedback capability; but, more appropriately, it demands a control process that can pick out the next step required at the user's will rather than the strongly ordered structure suggested by Figure 1. Perhaps I am wrong in interpreting the figure in this way as some of the statements in the text suggest rather more flexibility than I give credit for. I doubt whether much is to be gained from developing special-purpose software for exploring multivariate data. There are so many things that may be needed for all types of statistical analysis that the additional overheads on good general-purpose statistical software are not great. Most of the processes described in this paper are already easy to do in Genstat and must also be possible in other command-based systems. What would be beneficial are a few additional basic tools that facilitate interactive feedback and a good computing environment that allows easy linking of different programs.

Another disappointment in OMEGA is the seeming lack of the dynamic "animated control" graphics discussed in Section 4. Perhaps I am frustrated in the same way as the authors are, and dynamic facilities are available in OMEGA but cannot be demonstrated on two-dimensional sheets of paper. Nevertheless, the discussion of the example in Section 5 does not seem to appeal to dynamic graphics unless one includes the

J. C. Gower recently retired as Head of the Biomathematics Division and of the Statistics Department of the AFRC Institute of Arable Crops Research, Rothamsted Experimental Station. He is Visiting Professor of Statistics at City University. His mailing address is Vakgroep Datatheorie, Faculteit der Sociale Wetenschappen, Rijksuniversiteit te Leiden, Wassenaarseweg 52, Postbus 9555, 2300 RB Leiden, The Netherlands.

useful brushing technique. There is not much I can say about the example except that the data seems to have an exceedingly simple and well-defined structure. The authors were indeed fortunate in finding such strong linear structure which did not require the transformation of even one variable. Given the importance of measurement scales, it would have been nice if the authors had published the complete data set. That TOTORG and SUMDYE dominate the analysis is not very surprising as these seem to be totals over variables 1–14. (I think TOTORG is the sum of variables 1–14, but what SUMDYE is, is not clear to me.) What is clear is that the data have strong linear features, and that some of this linearity is inbuilt. How would these linear methods have fared if the samples had occupied a nonlinear manifold in 29-dimensional space? The detection of manifolds is one of the fundamental problems of multivariate data analysis. Projection pursuit is one attempt to help here, but I believe that transformations are likely to have more to offer, especially in nonlinear cases. Years ago, when Prim 9 was new, I asked the following question. Suppose I have a sample of, say, 1000 3×3 orthogonal matrices. These each give nine observations, so their space may be explored by Prim 9. Because sums-of-squares of all rows and all columns are unity, the points will lie on six three-dimensional spheres embedded in the nine-dimensional space. Further, sums-of-products of rows and columns vanish, so the points also lie on three-dimensional hyperboloids. Two-dimensional cross-sections will show circles and hyperbolae and as the cutting-planes move dynamically, the circles will grow larger, then smaller and finally vanish; similarly for the hyperbolae. How would a user observing these strange phenomena interpret what he saw? I have yet to receive a satisfactory answer to the question.

I believe that graphical methods for multivariate data analysis have much to offer. In the linear case, quite good progress has been made and I thank Drs. Weihs and Schmidli for their interesting contribution. Nonlinear multivariate analysis still has a long way to go. Progress will go hand-in-hand with good software, and I see that as a development of general-purpose statistical software.

Comment

Werner Stuetzle

This paper starts with a valid premise: many techniques for exploratory data analysis have been developed in an artificial context and illustrated using contrived and unconvincing examples. There is little experience as to which methods are useful in practice. Serious assessment of this issue would undoubtedly be valuable. However, the authors do not provide such an assessment. Their choice of building blocks for what they call the OMEGA pipeline appears to be largely driven by the computing environment at their disposal, and not by actual experience with a wide range of techniques. In addition to a case study, the paper presents a survey of methods and software. While such a survey could be helpful, the authors' attempt appears somewhat haphazard and incomplete. An encouraging aspect of the paper is the suggestion that techniques such as point cloud rotation, plot interpolation and Grand Tour, and brushing of scatterplots might eventually make their way from the esoteric realms of academia and research laboratories to actual consumers. I will first comment on the methodological part of the article and then on the data analysis.

Werner Stuetzle is Associate Professor, Department of Statistics, and Adjunct Professor, Department of Computer Science. His mailing address is Department of Statistics, GN 22, University of Washington, Seattle, Washington 98195.

COMMENTS ON METHODOLOGY

Simplification might be a useful idea. It comes up in other contexts, for example in Projection Pursuit (Friedman and Stuetzle, 1981), where one wants the chosen directions to involve as few of the variables as possible. The authors explain how the first principal component is simplified, although the properties of their procedure are not entirely clear. I do not see how they propose to simplify the second and higher principal components.

The motivation behind "$p\%$ resampling" is unclear. What is the distribution to be estimated? Why not simply do bootstrap resampling? Bootstrapping estimates the variability arising from repetitions of the experiment, assuming that the data can be interpreted as an iid sample from some distribution. One would then check how many principal component projections of bootstrap samples show some interesting