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

feature seen in the projection of the original data. Sequential presentation might be preferable over superposition. The problem of displaying tied observations in bootstrap samples could be solved by small amounts of jittering. If superposed display is desired, one could still match the bootstrap displays to the display of the original sample by procrustes analysis, but this should be considered a purely graphical device rather than a means for establishing some error distribution.

An entirely different approach would be to numerically summarize the variability of, say, the plane spanned by the first two principal components by considering the magnitudes of the second canonical angle between the plane for the original sample and the planes calculated from the bootstrap samples.

The authors' taxonomy of dynamic graphics techniques is unfortunate. It is hard to see how a "passive dynamic technique" (in the authors' sense) could be very useful. In the Grand Tour, for example, the user will want to control at least the speed in which the sequence of views is traversed, and the direction; the ability to back up to a previous view is obviously crucial. Whether one computes views as they are needed, or precomputes and stores an entire sequence of views, is to a large extent an implementation issue. The latter technique is what is usually referred to as "animation." Animation is preferable if the views are complex and take a long time to compute, in which case computing them on the fly would lead to a jerky movie. It is also easy to implement on some window systems, and thus can be "poor man's dynamic graphics." On the other hand, animation places restrictions on user control.

There is, however, an important distinction to be made, namely the distinction between adaptive and nonadaptive techniques. An adaptive technique would have some notion of what constitutes an "interesting" view, and it would attempt to show such views. One might, for example, run Exploratory Projection Pursuit (Friedman, 1987) on the data, order the projections in decreasing order of projection index and then show a movie interpolating between those views. In contrast, a nonadaptive technique would pick a sequence of views independent of the data.

The authors' appendix on "Hardware and Software Environment" is revealing. It confirms that a completely satisfactory computing environment for data analysis and scientific/statistical computing still does not exist. Such an environment would at least have the following properties (see McDonald and Pedersen, 1985a, b; 1988):

- Its command language should support a wide range of programming paradigms and data representations. There should be both an interpreter and a compiler. The compiled code should run at an efficiency comparable to Fortran.
- It should support multiwindow, dynamic, interactive graphics.
- It should offer access to Fortran and C libraries.
- It should run on a variety of machines.

To my knowledge, the closest to this is the XLISP-STAT system (Tierney, 1988). Its main drawbacks are that it is based on XLISP and not the standard COMMON LISP, and that there are no compilers for XLISP.

All currently existing data analysis environments have fundamental shortcomings in the area of graphics and user interfaces. The layout of plots, the placement of menus and controls and the effects of pointing at icons are hard wired and not intended for change by the user. It is impossible to configure new kinds of displays or to set up graphical user interfaces with menus, dials, sliders, buttons, etc. These capabilities could be extremely beneficial. They would, for example, allow a user who repeatedly sees similar data sets posing similar questions to customize an "analysis pipeline" with specially designed displays and controls. Identification of appropriate abstractions and integration of new programming paradigms, like constraint programming (Borning, 1981) into data analysis environments are important research problems.

COMMENTS ON THE ANALYSIS

According to the authors, the main questions to be answered by the analysis are: Can the coloristic properties of the dye be predicted from the analytic measurements, and, if so, how? Can the visual judgements be predicted from the coloristic measurements? These are questions that would commonly be tackled by regression analysis, possibly followed by an analysis of the configuration of the observations in predictor space. It would be interesting to hear the authors' motivation for using principal components and canonical correlations as their primary analysis tools.

ACKNOWLEDGMENTS

I thank Andreas Buja and John McDonald for helpful discussions.

ADDITIONAL REFERENCES

BORNING, A. H. (1981). The programming language aspects of Thinglab. *ACM Trans. Programming Languages Systems* **3** 353–387.

FRIEDMAN, J. H. (1987). Exploratory projection pursuit. *J. Amer. Statist. Assoc.* **82** 249–266.

FRIEDMAN, J. H. and STUETZLE, W. (1981). Projection pursuit regression. *J. Amer. Statist. Assoc.* **76** 817–823.

McDONALD, J. A. and PEDERSEN, J. (1985a). Computing environments for data analysis. I: Introduction. *SIAM J. Sci. Statist. Comput.* **6** 1004–1012.

McDONALD, J. A. and PEDERSEN, J. (1985b). Computing environments for data analysis. II: Hardware. *SIAM J. Sci. Statist. Comput.* **6** 1013–1021.

McDONALD, J. A. and PEDERSEN, J. (1988). Computing environments for data analysis. III: Programming environments. *SIAM J. Sci. Statist. Comput.* **9** 380–400.

TIERNEY, L. (1988). XLISP-STAT: A statistical environment based on the XLISP language. Technical Report 528, School of Statistics, Univ. Minnesota.

Comment: Industrial Strength VEDA

Forrest W. Young

Multivariate visual exploratory data analysis (**VEDA**) has withstood its "test of fire": Weihs and Schmidli are the first to try multivariate **VEDA** methods in an industrial applied statistics setting, and the methods proved useful. They are to be commended for their bravery in implementing and carrying out such a project, and are to be congratulated both on their successful application and on providing us with a model paper which shows how to turn the process of data visualization into a readable and informative report.

As one of the developers of multivariate **VEDA** methods, I am, naturally, very pleased with this paper. It is exciting to see that our methods can be used, in the words of Weihs and Schmidli, by "the investigator faced with an ongoing stream of many data sets, limited time and the need for a fairly general single routine strategy," and not just by developers who are "presenting just one more method · · · (with) examples particularly fitted to demonstrate their usefulness."

My excitement stems from three aspects of the Weihs and Schmidli paper: 1) the example, 2) the confirmatory use of exploratory methods, and 3) the emphasis on the independence of the visual methods from the multivariate methods. I will discuss these points in the next three sections of this comment. My excitement is tempered somewhat, however, by one major shortcoming: When "variables can be naturally attached to more than one group, and the predictability of one group by another is of interest" (to quote the authors), then redundancy analysis (Lambert, Wildt and Durand, 1988) should be used, not canonical analysis as suggested by the authors. I will discuss this shortcoming in the fourth section of this comment. Since no plotting tools have been proposed for redundancy analysis, in the fifth section I present the **triplot** a new **VEDA** tool for redundancy analysis with certain similarities to the biplot (Gabriel, 1971), comparing it to biplots and to the authors' approach to simplification.

Forrest W. Young is Professor of Psychometrics at University of North Carolina at Chapel Hill. His mailing address is Psychometric Laboratory, CB 3270 Davie Hall, University of North Carolina, Chapel Hill, North Carolina 27599-3270.

1. ILLUSTRATION

The application used by Weihs and Schmidli to illustrate **OMEGA** involves searching for structure in multivariate data arising in the context of a major pharmaceutical, dyestuffs and agrochemical company. The data, which concern the quality of dyestuffs, are used by Weihs and Schmidli to illustrate the kind of problem for which a routine online multivariate **VEDA** strategy is required in the industrial data analysis context.

The illustration of multivariate **VEDA** methods provided by Weihs and Schmidli is exciting because it reports the process of a real visual exploratory data analysis, not just the conclusions of the process nor a "cleaned-up" mythical version of the process. The illustration shows the dead-ends, the surprises, and the excitement of **VEDA** being applied to a typically messy set of data.

One of the major strengths of the analysis is that the authors begin with Principal Components Analysis (PCA), even though the fact that the variables fall into two groups suggests immediately that Canonical Correlation Analysis (CCA) be used. They ask the rhetorical question "But is it really justified to impose variables grouping at the beginning of the analysis?" to which they answer no, saying that they might "miss something." Thus, "following this feeling" they postponed CCA until later. My own experience is that this strategy is the best to follow. PCA is, I believe, the single most powerful multivariate exploratory tool that we have, and is nearly always my first choice with a new and unfamiliar set of data. I also find it very refreshing to see a phrase such as "following this feeling" being presented without embarrassment, since informed, scientifically based feelings—hunches, if you will—are a very important aspect of **VEDA**. They note that "we were lucky" that PCA helped them