CLEVELAND, W. S., MCGILL, M. E. and MCGILL, R. (1986). The shape parameter of a two-variable graph. *Proc. ASA Sec. Statist. Graphics* 1–10.

DE LEEUW, J. (1984). The GIFI system of nonlinear multivariate analysis. In *Data Analysis and Informatics III* (E. Diday, M. Jambu, L. Lebart, J. Pagès and R. Tomassone, eds.) 415–424. North-Holland, Amsterdam.

DONOHO, A. W., DONOHO, D. L. and GASKO, M. (1988). MACSPIN: Dynamic graphics on a desktop computer. In *Dynamic Graphics for Statistics* (W. S. Cleveland and M. E. McGill, eds.) 331–351. Wadsworth, Belmont, Calif.

EASTMENT, H. T. and KRZANOWSKI, W. J. (1982). Cross-validatory choice of the number of components from a principal component analysis. *Technometrics* **24** 73–77.

FISHERKELLER, M. A., FRIEDMAN, J. H. and TUKEY, J. W. (1988). PRIM-9: An interactive multidimensional data display and analysis system. In *Dynamic Graphics for Statistics* (W. S. Cleveland and M. E. McGill, eds.) 91–109. Wadsworth, Belmont, Calif.

FRIEDMAN, J. H. and TUKEY, J. W. (1974). A projection pursuit algorithm for exploratory data analysis. *IEEE Trans. Comput.* **C-23** 881–890.

GABRIEL, K. R. (1971). The biplot graphic display of matrices with application to principal component analysis. *Biometrika* **58** 453–467.

GNANADESIKAN, R. (1977). *Methods for Statistical Data Analysis of Multivariate Observations.* Wiley, New York.

GOODALL, C. and THOMA, H. M. (1987). Interpolation of multivariate data. *Proc. ASA Sec. Statist. Graphics* 64–67.

HUBER, P. J. (1985). Projection pursuit. *Ann. Statist.* **13** 435–475.

ISP (1988). *ISP User's Guide.* Artemis Systems, Carlisle, Pa.

JOLLIFFE, I. T. (1972). Discarding variables in a principal components analysis. I: Artificial data. *Appl. Statist.* **21** 160–173.

KRZANOWSKI, W. J. (1987). Selection of variables to preserve multivariate data structure, using principal components. *Appl. Statist.* **36** 22–33.

SIBSON, R. (1978). Studies in the robustness of multidimensional scaling: Procrustes statistics. *J. Roy. Statist. Soc. Ser. B* **40** 234–238.

SIBSON, R. (1979). Studies in the robustness of multidimensional scaling: Perturbational analysis of classical scaling. *J. Roy. Statist. Soc. Ser. B* **41** 217–229.

S-PLUS (1988). *S-PLUS User's Manual.* Statistical Sciences, Seattle, Wash.

STUETZLE, W. (1988). Plot windows. In *Dynamic Graphics for Statistics* (W. S. Cleveland and M. E. McGill, eds.) 225–245. Wadsworth, Belmont, Calif.

WANG, C. M. and GUGEL, H. W. (1988). A high performance color graphics facility for exploring multivariate data. In *Dynamic Graphics for Statistics* (W. S. Cleveland and M. E. McGill, eds.) 379–389. Wadsworth, Belmont, Calif.

WEIHS, C. (1988). Dynamic graphical methods in multivariate exploratory data-analysis: An overview. Technical Report 8802, Mathematical Applications, CIBA-GEIGY, Basel.

WOLD, S. (1978). Cross-validatory estimation of the number of components in factor and principal component models. *Technometrics* **20** 397–405.

YOUNG, F. W., KENT, D. P. and KUHFELD, W. F. (1988). Dynamic graphics for exploring multivariate data. In *Dynamic Graphics for Statistics* (W. S. Cleveland and M. E. McGill, eds.) 391–424. Wadsworth, Belmont, Calif.

Comment

A. Buja and C. Hurley

Reading the authors' paper was very gratifying for us: as it happens, we have been working on the integration of multivariate analysis and graphical data analysis as well. We are delighted to observe that our separate efforts converged to some degree. While we may differ in details of implementation, human interface and computing philosophy, our independent efforts indicate a certain necessity in the idea of marrying classical multivariate analysis and the more recent high-interaction graphics tools. A paper by us on this subject is in press in the *SIAM Journal on Scientific and Statistical Computing* (Hurley and Buja, 1990). It is based on the Ph.D. thesis of Hurley (1987). The multivariate methods which we considered were the same as those of the authors with the exception of their successive orthogonalization procedure. The authors carried certain ideas of visual inference and assessment considerably further than we did (for now, we have not gone beyond what is documented in Buja, Asimov, Hurley and McDonald, 1988). On the other hand, we may claim a tighter coupling of multivariate analysis and graphics, as we will show below.

A. Buja is a member of the technical staff, Bellcore, 445 South Street, MRE 2Q-362, Box 1910, Morristown, New Jersey 07960-1910. C. Hurley is Assistant Professor, Department of Statistics, The George Washington University, Washington, DC 20052.

MULTIVARIATE ANALYSIS (MA) AND GRAPHICAL METHODS

A basic motivation behind the authors' and our endeavor is the simple insight that MA allows us to generate a wealth of potentially illuminating data projections. Curiously, the first attempts at combining interactive graphics with automatic methods for finding informative projections were based on projection pursuit rather than classical MA. Surely, the latter can be interpreted as a subset of the former, but this view does not do justice to MA. It is more useful to interpret MA as a set of methods for changing coordinate systems in a data-driven way. One reason for the initial lack of interest in the graphical and exploratory side of MA among graphics people in statistics may derive from the inferential tilt of the statistics literature on this subject. Projection pursuit seemed more natural as it was always presented as an exploratory tool, unencumbered by theories of inference which tend to dominate their subjects. Another reason for the higher profile of projection pursuit could be the fact that it was initially presented as an improvement over principal components: Friedman and Tukey (1974) constructed some artificial data (spherical clusters sitting on the vertices of a simplex) whose structure was partially revealed by projection pursuit, while the first two principal components were uninformative. We doubt that this example is a sufficient reason for discarding the more accessible methods of MA, which moreover are adapted to simple correlational structure found commonly in real data.

In our implementation of a system for graphical MA, we adopt the view that MA produces potentially informative coordinate systems. We consider coordinate systems as means for specifying subspaces to which data analysts may wish to confine projection planes. Within such subspaces, they can freely explore data projections via 3D rotations, plot interpolation and grand tour motion. A major difference of our framework from the authors' is that we provide visual clues for the position of the current projection plane in two coordinate systems: the canonical basis corresponding to raw variables on the one hand, and the basis derived from principal components or any other MA method on the other. This is achieved by two sets of variable icons (called variable boxes, Buja, Asimov, Hurley and McDonald, 1988; Hurley and Buja 1990) which largely replace the information usually supplied by tables of coefficients or loadings, such as the authors' Table 3a). The dual clues in terms of raw and derived variables allow one to read off at any time how the current projection "loads" on raw variables and variates obtained from MA. In addition, the variable icons play an active role as input devices in activating and deactivating variables of either kind via mouse clicks. An interactive approximation to the authors' simplification method in our framework would be as follows: activate, say, the projection onto the first two principal components; then give control to the raw variables and deactivate those which display only marginal loadings for the current projection; our system will then automatically perform a general 4D motion of the projection plane in order to zero out the deactivated variables. Such an operation would be part of what we call a "guided tour," i.e., guiding projections by playing with subspace restrictions.

Motion is based on the principal of geodesic interpolation of pairs of planes. If applied to sequences of unrestricted random planes, one obtains an implementation of the grand tour (Asimov, 1985; Buja and Asimov, 1985). The numerical methods used for interpolation of projection planes are described in detail in Buja, Asimov and Hurley (1989).

We have considered additional tools for performing parallel analyses such as the authors describe in Section 4.5. Quite often, in a parallel analysis one compares 2D scatterplots of different data subsets in the same or analogous coordinates on the screen. Similarly 3D (or higher dimensional) scatterplots may be compared by performing simultaneous rotations of the plots, while ensuring that at any moment the plots employ the same projection coefficients. In this way one could, for example, compare the 3D scatterplot yielded by the first three principal components, with the 3D structure obtained by performing a principal components analysis of a subset of the variables (or observations). In general, we note that in a graphical parallel analysis one compares multiple views which differ by a few of the transformations composed in the viewing pipeline (e.g., different nonlinear transformations of the variables, various random permutations). Our implementation allows dynamic linking of such plots so that when a plot changes, all plots linked to it change automatically in a manner determined by the common pipeline element (e.g., the projection operation; see Buja, Hurley and McDonald, 1986).

PROGRAMMING ENVIRONMENTS IN GENERAL

However useful the authors' (or our) proposal for a viewing pipeline may be, it is not the last word, and no final version should ever be expected. The problem has to do with the fortunate situation that data analysis requires creativity and allows for personal styles to some extent. The OMEGA pipeline may suit 1) specific types of data and problems, 2) the tastes of the authors, and 3) the computing environment at their disposal. In other places and for other data analysts with other computing resources, a useful viewing pipeline may look very different. What, under these circumstances, can we offer in ways of research that is of wider interest? We do not think that the answer is a monster pipeline which does everything for everyone, although it is necessary that some well-developed prototypes be implemented and published to give existence proofs of the concepts. We believe that an answer can be found in the direction of programmable pipeline modules, which give mildly sophisticated users the opportunity to concoct their own viewing machinery. This implies that a reasonable set of building blocks be found, and that they be accessible at a reasonably high level of abstraction, i.e., in a language which expresses the desired manipulations not too differently from the way we think about them. And, of course, this language should be part of a larger system which provides statistical and general purpose scientific computing at an equally high level of abstraction. It appears that computing environments close to this ideal are just now emerging. We know of at least one that is inexpensive and easily accessible on common hardware: Tierney's LISP-STAT (1990) system. It brings within everyone's reach the kinds of tools which some of the more "exotic" authors (e.g., McDonald and Pedersen, 1988) have been writing about. Besides offering basic statistical computing and a host of programmable high-interaction graphics tools, LISP-STAT also allows one to implement other abstractions, "statistical models" for example. At the base of this high-level language is an extension of LISP for so-called object-oriented programming, possibly the most important contribution of applied artificial intelligence to computing.

What is the point of this excursion, apart from being a sales pitch for a particular piece of software? We mean to indicate that "exotic" research, which tries to bring to statistical computing such alien notions as object-oriented programming, has a bearing on complex methodologies like the one presented by the authors. One is forced to rethink the software tools at hand if viewing pipelines for data analysis should 1) feature as much graphical and statistical functionality as the OMEGA pipeline, 2) be capable of providing high-interaction control, and 3) yet be user programmable for creative experimentation and tailoring to specific problems. While it is certainly true that anything can be done on a computer, say, in assembler language, the challenge is to raise the level of communication between humans and machines. And here is where the jargon of "high-level abstraction" takes on a more technical meaning: it refers to the expressive wealth of high-level programming languages (Fortran is not one of them)—a wealth which reduces the number of steps that humans take when translating their mental models (of, say, a viewing pipeline) into machine-readable form.

THE DATA ANALYSIS EXAMPLES

The authors' presentation of their analysis is refreshing in that it does not hide some rough edges and some of the history of the analysis. One could ask several questions about what was done and propose a number of additional things that could have been done. On the other hand, in exploration it can occur that a priori unmotivated actions find justifications simply by their success. The authors' initial PCA is a case in point. We feel that squabbling over details and matters of taste is beside the point of the analysis. The only question worth mentioning concerns the cross-validation/Procrustes procedure: we do not understand what kind of projection variability is assessed here. See W. Stuetzle's comments for some further thoughts.

Some of the lessons we learned (or had confirmed) from the exercise of the authors' analysis are the following.

1. Multivariate analysis can be a powerful tool in revealing structure which has nothing to do with conventional distribution theory.
2. In the presence of large numbers of variables, MA can help to locate some of the critical ones. However, canonical correlation analysis has the same collinearity problem as regression, and therefore, assessing how strongly a certain variable contributes to a canonical variate depends heavily on the other included variables.
3. Informal inference is useful. As data analysis becomes more qualitative due to the pervasiveness of graphics, assessment of complex plots is needed in the form of simulation of null situations, resampling or leave-out methods. Results can be displayed as real-time movies (sequential presentation) or superposition plots (simultaneous presentation), or simply arranged in parallel.

CONCLUDING REMARKS

One of the more important aspects of the authors' paper is how it integrates tools in a computational framework which allows one to actually carry out complete analyses. It is one of the biases of our publishing culture that microscopic investigations of very specialized methods are easier to place in journals than attempts to integrate tools in global strategies. As is indicated by the authors' work, in an applied context (be it industrial or academic consulting) there is no patience with partial answers and incomplete tools. To get a job done, one needs a set of strategies for data analysis and a computational framework (such as the OMEGA pipeline) to facilitate the application of these strategies. In this sense, the computational framework can be regarded as an expression of the underlying strategic ideas. If the computational framework reflects a set of strategies properly, it will allow one to perform with greatest ease those actions which are the most important ones according to the strategic ideas. It would be an error to regard strategy as a rigid game plan. A better notion is that of a hierarchy of options which an analyst may or may not choose to apply in a sensible sequence in the course of an analysis. On the other hand, the notion of a computational framework is related (although not identical) in that it describes the implementation of such a hierarchy of options on a computer. If this diagnosis of the situation is appropriate, we should expect that a discussion of data analytic strategies is helped by the precision obtained by casting strategies in terms of computational frameworks.

We would like to thank the authors for a stimulating paper and hope that this is not the end but the beginning of a discussion.

ADDITIONAL REFERENCES

BUJA, A. and ASIMOV, D. (1985). Grand tour methods: An outline. In *Computer Science and Statistics: Proc. of the 17th Symposium on the Interface* 63–67. North-Holland, Amsterdam.

BUJA, A., ASIMOV, D. and HURLEY, C. (1989). Methods for subspace interpolation in dynamic graphics. Technical Memorandum, Bellcore, Morristown, N.J.

BUJA, A., HURLEY, C. and MCDONALD, J. A. (1986). A data viewer for multivariate data. In *Computer Science and Statistics: Proc. of the 18th Symposium on the Interface* 171–174. North-Holland, Amsterdam.

HURLEY, C. (1987). The data viewer: A program for graphical data analysis. Ph.D. dissertation and Technical Report, Dept. Statistics, Univ. Washington.

HURLEY, C. and BUJA, A. (1990). Analyzing high dimensional data with motion graphics. *SIAM J. Sci. Statist. Comput.* To appear.

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. (1990). *LISP-STAT.* Wiley, New York. To appear.

Comment

Frank Critchley

It is a pleasure to welcome this paper by Weihs and Schmidli with its emphasis on the practical benefits which derive from combining classical dimensionality reduction methods with recent advances in interactive, dynamic graphics in a single integrated computing environment. At the same time, however pressing the practical need, asking for "a fairly general *single* routine strategy" (Section 1.1) for multivariate exploratory analysis seems, to me at least, to be asking for the moon. A more realistic objective might be to establish a framework of methods through which the user is guided by an expert system. We elaborate a little on this possibility below.

With one exception, my comments are of two types: possible extensions and remarks on the example. The exception is a detail which we dispose of first. In the context of resampling and Procrustes transformation (Section 3.7), the authors suggest that "it may be worth looking for analytic expressions derived from data disturbances analogously to Sibson (1979)." At least for PCA-COV and PCA-COR, some relevant formulae are given in Sections 3.6.2 and 6.3 of Critchley (1985). Note that the covariance matrix used there has divisor n. Trivial modifications apply when the divisor is $(n - 1)$. The formulae given are essentially expansions in inverse powers of $(n - 1)$. In practice, these expansions are usually truncated to obtain approximations. In this case, greater accuracy can be achieved by renormalization of the eigenvalues to sum to the easily computed perturbed trace and of the eigenvectors to have unit length. Exact orthogonalization is also possible.

Frank Critchley is Chairman, Department of Statistics, University of Warwick, Coventry CV4 7AL, United Kingdom.

POSSIBLE EXTENSIONS

The following remarks are partly taken from the unpublished conference paper by Critchley (1987) on graphical data analysis. They relate principally to the dimensionality reduction methods employed.

1. In that paper I suggested that healthy progress requires constructive interaction between five ingredients: (a) important practical problems, (b) sufficient computing power, (c) a sound mathematical/statistical basis, (d) a good framework of methods, and (e) international cooperation. The present paper is an excellent example of the first three ingredients, while hopefully its publication in this format in this journal will encourage the last of these!

2. It is within the fourth ingredient that there is perhaps the greatest scope for fruitful extensions. The authors offer in Table 1 a classification of multivariate techniques in terms of two "dimensions": the preinformation required and the aspects of the data that are optimally represented. This framework of methods can be fruitfully extended by adding new methods (as the authors remark in Section 6) and also, we note here, by adding new "dimensions" to the classification of methods.

3. The methods currently considered can be characterized as corresponding to one of several possibilities on each of a (nonexhaustive) number of additional