of Young and Harris (1990). Once the unimportant variables have been eliminated by simplification, then biplots can be used to display not only the distribution of observations in the two-dimensional multivariate space (as in the Weihs and Schmidli figures), but also the distribution of retained variables. This provides a more informative plot which displays relationships within the variables and between variables and observations, as well as within observations. The paper, the OMEGA pipeline, and the richness of the results would be strengthened with the inclusion of biplots.

The biplot can be extended in a very interesting way for redundancy analysis, as was originally proposed by Young and Sarle (1981). The extension uses the first two redundancy variates as the dimensions for a two-dimensional plot of the "redundancy plane." This is the plane in the predictor space which shares the most variance with the criterion space. A biplot can be constructed in this plane in the ordinary way, using the scores of the observations on the two redundancy variates as coordinates of observation-points, and the coefficients of the predictor variables on the redundancy variates as coordinates of the end points of predictor-variable-vectors which extend from the origin of the space. This biplot can be extended to become a *triplot* by adding to the biplot the projection of the criterion variables into the redundancy plane. They should be displayed as vectors. The plot of the redundancy plane now contains three kinds of information: the observations are represented as points, while the two sets of variables are represented as vectors.

The algebra underlying the redundancy triplot is as follows. The redundancy model is expressed by the equation $Y = XL$, subject to suitable restrictions on L. Since L is nonnegative definite, it is the case that $L = AB$, and we can re-express the model by the fundamental RDA equation $Y = XAB$. The rank two approximation to the criterion variables Y is given by the approximation $Y \simeq XA_2B$, where the subscript 2 indicates we are using only the two sets of linear combinations that correspond to the largest two eigenvectors. The redundancy model can now be re-written as $Y \simeq Z_2B$, where $Z_2 = XA_2$. The values in Z_2, which are the scores on the first two redundancy variables, are displayed as points in the triplot, whereas the values in A_2 (the coefficients of the predictor variables) and B_2 (the coefficients of the criterion variables) specify the endpoints of vectors emanating from the origin of the biplot.

ADDITIONAL REFERENCES

LAMBERT, Z. V., WILDT, A. R. and DURAND, R. M. (1988). Redundancy analysis: An alternative to canonical correlation and multivariate multiple regression in exploring interset associations. *Psychological Bull.* **104** 282–289.

MULLER, K. E. (1981). Relationships between redundancy analysis, canonical correlation, and multivariate regression. *Psychometrika* **46** 139–142.

MULLER, K. E. (1982). Understanding canonical correlation through the general linear model and principal components. *Amer. Statist.* **36** 342–354.

YOUNG, F. W. (1974). Scaling replicated conditional rank-order data. In *Sociological Methodology* (D. Heisse, ed.). 129–170. Amer. Sociological Assoc.

YOUNG, F. W. and HARRIS, D. F. (1990). Estimating missing values with dynamics for principal components. (Unpublished manuscript.)

YOUNG, F. W. and RHEINGANS, P. (1990). Visualizing structure in high-dimensional multivariate data. *IBM J. Res. Develop.* **34**. To appear.

YOUNG, F. W. and SARLE, W. S. (1981). *Exploratory Multivariate Data Analysis*. SAS Institute, Cary, N.C.

Rejoinder

Claus Weihs and Heinz Schmidli

We would like to thank the discussants for initiating the debate on our conceptual framework of interactive data analysis. Our responses cover five areas: the actual implementation of the OMEGA pipeline concerning software and methods, the data analysis example, possible extensions of the tool box, and a desirable ideal strategy.

SOFTWARE IMPLEMENTATION

The implementation of the OMEGA pipeline has always been, and remains, restricted by the graphical power of the underlying software (ISP). We have never attempted to program our own graphics system. Therefore, the concept of the OMEGA pipeline goes far beyond our implementation (as described in Appendix 2). We were not intending to describe one more software tool, as Gower seems to assume, but rather a working implementation of a concept. Nevertheless, even the capabilities of the implemented version cannot be demonstrated on paper (see also Section 4.2). In fact, no real attempt was made to illustrate dynamics or to describe details of the software, like variable selection or interactive elimination of observations. Instead, we tried to demonstrate the power of the concept by showing what actions lead to which results.

We believed that this kind of demonstration would also make clear the flexibility of the implementation of the OMEGA pipeline. Unfortunately, Gower seems to take its structure (as illustrated in Figure 1 of the paper) literally, in that he interprets the ordering of the blocks as fixed. Let us, therefore, make clear that the ordering of the blocks is controllable by the user, apart from some natural restrictions (e.g., dimension reduction before simplification).

IMPLEMENTED METHODS

Buja and Hurley proposed a graphical substitute for our importance criterion based simplification method. This might work for PCA, since then the absolute value of the loadings is a natural criterion for effect elimination (see Section 3.6). For CDA and CCA, appropriately scaled loadings would have to be displayed (see Appendix A1.6). Moreover, our simplification procedure includes an additional feature, namely the rounding of the loadings to a given number of decimal places (not just to a single decimal place as Young seems to believe). Stuetzle indicates that the simplification procedure for the second and higher principal components might be somewhat unclear. As noted in Section 3.6, once the first simplified coordinate is found, "the whole orthogonal system ··· [is] rotated so that the first axis coincides with the 'simplified' first new coordinate." The simplification procedure as described for the first coordinate is then applied to the second coordinate of the rotated system, and so on.

Stuetzle, Buja and Hurley had some difficulties in understanding the kind of prediction variability we try to assess with our resampling/Procrustes procedure. Rephrasing the motivation, for PCA say, might clarify ideas. Our first idea was to look mainly at the distribution of the projections corresponding to the complements of the subsamples used to generate the principal components (prediction) and not only at the distribution of the principal components themselves corresponding to those subsamples. This is closely related to the proposal of Stuetzle, apart from the different resampling strategy. But, how would an observer judge the similarity of the simulated projections? Would he or she not consider as equivalent all those projections which can be obtained by rotations/reflections, translations or global scale change (Procrustes)? If so, the problem remains how to choose the actual representative from these equivalent projections. For obvious reasons, we decided to choose the one that is the nearest to the projection obtained by full sample PCA. Thus, we are representing the pointwise distribution of predictions of the PCA projections after elimination of effects arising from randomly arriving at one of many equivalent projections.

Fisher also proposed relooking at the resampling procedure in order "to see whether we can do rather better than simply displaying a cloud of grapeshot in the vicinity of a point estimate, or a skein of spaghetti strangling a density." As a response, let us remark that we are using this kind of graphics as a diagnostic tool to identify abnormalities in the pointwise distribution and to identify influential observations (see Figure 19 of the paper).

Critchley mentioned that he had already developed analytic expressions for the influence of data disturbances for PCA. Unfortunately, his analysis (Critchley, 1985) stops with the principal components and does not expand on Procrustes transformed projections. Krzanowski (1984) advanced to a similar stage.

Gower's discussion reflects some small misunderstandings (see also above) indicating that our paper is, perhaps, not as clear as it ought to be. So we repeat: different nonlinear transformations can be applied to different variables; there is no restriction causing en bloc handling (see also Section 4.6.1 of the paper).

Stuetzle seems to have a general uneasiness with our choice of methods for the OMEGA pipeline. But, on the one hand, the pipeline has room for improvement and also for personal styles, and, on the other hand, the methods have proven their usefulness in routine application.

Since the paper was written, we have ourselves changed at least one important block of the OMEGA pipeline. For plot interpolation we now use the static technique of disconnected arrows in the superposition of original and simplified projections. In Section 4.3, we already indicated that this static technique totally reflects the dynamic interpolation. Moreover, the disconnected arrows technique is surely more informative than the static simulation of dynamic interpolation (compare Figure 4 of the paper and Figure A1).

Furthermore, we now also recommend Procrustes transformations after simplification, since the arguments for qualitative projection similarity are also applicable to simplification (see also Sections 3.6 and 3.7). The way in which the impression of a simplification changes by the use of Procrustes transformations is illustrated in Figures A1 and A2.

EXAMPLE

Stuetzle proposed regression analysis to tackle the two main questions to be answered by the analysis, i.e., can coloristic properties be predicted by analytical measurements, and can visual judgements be predicted by coloristic measurements? First note that the visual judgements are ordinal variables, and hence our use of discriminant analysis appears to be preferable to regression. Concerning the technical measurements of coloristic impression, in fact numerical variables, regression analysis might be an appropriate method. However, very good predictable linear combinations

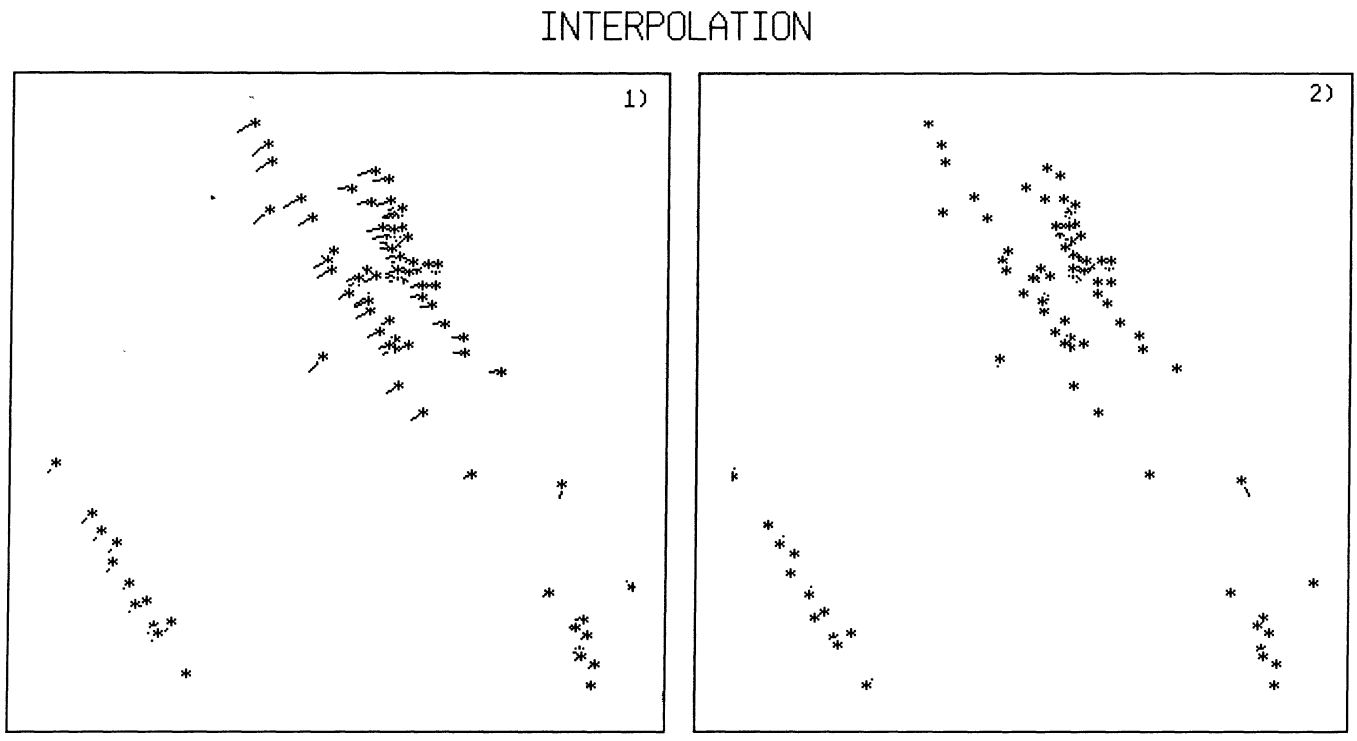

FIG. A. *Interpolation illustrated by disconnected arrows in the superposition of original and simplified projection, "∗" indicating simplifications, before* (1) *and after* (2) *Procrustes transformations.*

of the coloristic variables, e.g., differences in hue under daylight and artificial light, are of equal interest to the producer as long as they can be interpreted.

There are two other kinds of comments concerning the example. One concerns the use of PCA-COV, the other concerns the outliers.

Critchley wonders whether PCA-COR delivers similar results to PCA-COV. This is, in fact, not the case. On the one hand, this is not surprising due to the very different scales of the variables. Gower already noted that the variable TOTORG should be related to the sum of variables 1–14 (see Table 2 of the paper). On the other hand, it is not very important, since PCA-COV delivered an important result, the change of the measurement procedure. Other methods may, or may not, deliver other insights.

Some of the discussants are interested to obtain more information about "outliers," in particular about batches 84 and 93, "misplaced" when plotting TOTORG versus SUMDYE (see Figure 8 of the paper), and about the obvious outlier in Figures 6 and 7 of the paper, which refers to batch 85. Unfortunately, it has proved impossible to identify the causes for such outliers so long after the actual measurements were taken.

TOOL BOX EXTENSIONS

Critchley, Fisher, Gower and Young took the opportunity to recommend extensions to the OMEGA tool box. Critchley proposed to include a whole, well structured, bunch of multivariate techniques in order to guarantee a rich framework of methods. We agree that the realization of such an ambitious project is only possible by means of international cooperation.

Critchley also proposes a "constructive interplay between ··· the exploratory, graphical approach (to data analysis) and the confirmatory, modeling approach." In fact, some of his proposals were all along included in the OMEGA pipeline. In particular, brushing of influential points was demonstrated together with CDA (see Section 5.4 and Figure 19), and the examination of the robustness of the dimension reduction methods by resampling is a standard tool in the OMEGA pipeline.

Fisher recommended the use of his chi-plot matrix in conjunction with the scatterplot matrix in order to test, graphically, for independence of pairs of variables. We compared the outcomes of the permutation test (superposition mode, see Figure 11 of the paper) and the corresponding chi-plot (see Figure B). At least to us, "near-independence" is much clearer in the superposition.

Gower proposed the extension of the tool box by the Gifi-methods or, at least, by some variant of Multiple Correspondence Analysis/Homogeneity Analysis/Fisher's Method of Optimal Scores. We already mentioned in Section 6 that the Gifi-methods are on the list of planned extensions.

In practice, however, our first essential extension of the OMEGA pipeline was the Partial-Least-Squares

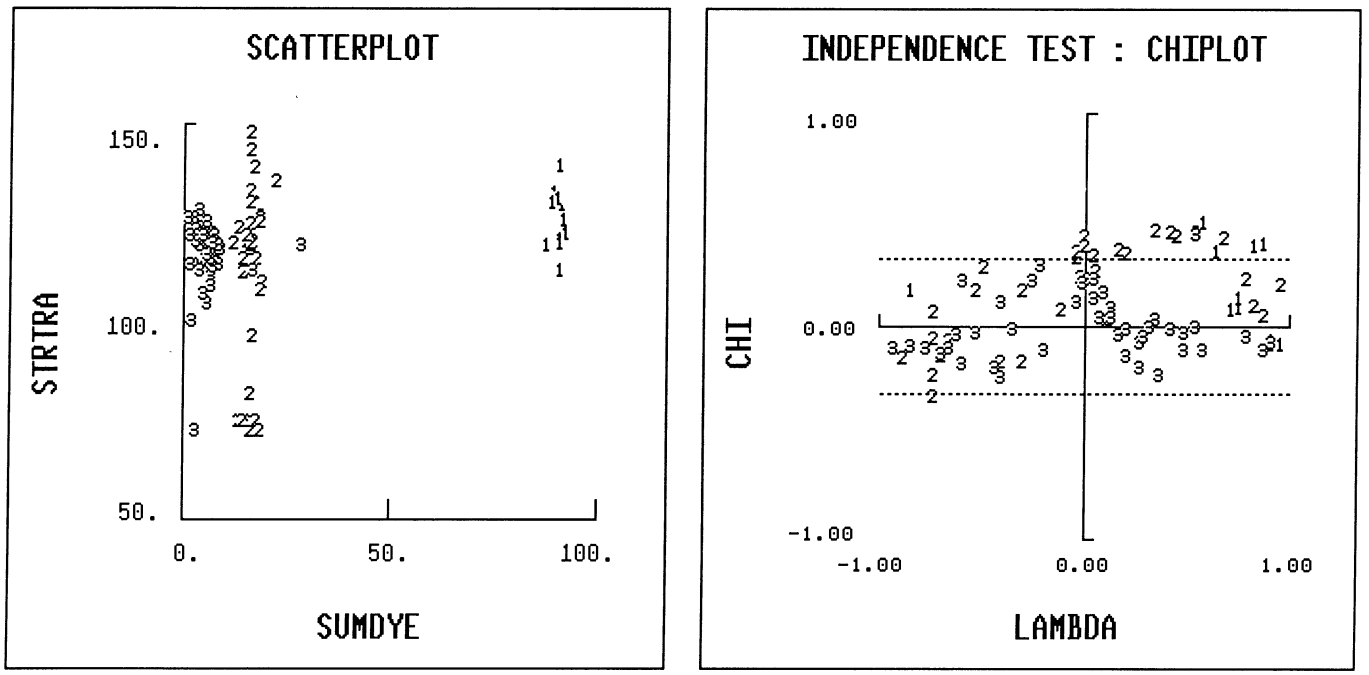

FIG. B. *Test of independence by means of a chi-plot.*

(PLS–) method (see Hoeskuldsson, 1988) not listed in Section 6. This was because PLS is very popular in chemometrics, which is our main application field.

Young argued for the use of Redundancy Analysis (RDA) instead of CCA if a set of criterion variates is to be predicted by a set of predictor variates (unsymmetric situation). We applied RDA in the two cases to which we had applied CCA in the paper. This led to very similar results to those obtained with CCA. Indeed, RDA, like CCA, selected SUMGRD, TERTMC, and SUMRED as predictors for HUEREM, HUEREMAL (compare Table 4 of the paper and Table A), and the prediction of the measured brightnesses was poor for both RDA and CCA.

Last, but not least, Young proposed the use of the biplot (or triplot) complementary to simplification in order to visualize the distribution of the retained variables. A biplot version of Figure 2 of the paper (see Figure C) convinced us that this is indeed helpful. Note that the diagonals of this projection are interpreted in Section 5.2.

TABLE A
Redundancy analysis

	Linear combinations: Analytical variables				
	(a)		(b)		
	RD1	RD2	Importance criterion	Initial simplification	RD1S
Redundancy	0.75	0.02			0.60
TERCUP	0.379		0.200	0.0	
DNANDI	−0.014		−0.010	0.0	
DNBZDI	0.319		0.298	0.0	
SECMC	−2.447		−7.862	0.0	
TERTMC	3.580		12.099	3.6	0.293
PRIMSC	−2.394		−1.324	0.0	
SECSC	−2.839		−1.062	0.0	
DNSEC	−1.983		−0.591	0.0	
TERTSC	4.015		3.864	0.0	
SUNKUV	−0.312		−0.437	0.0	
SUMUV	−0.731		−1.094	0.0	
SUMRED	2.166		8.221	2.2	−0.135
SUMGRD	−4.247		−16.244	−4.2	−0.113
SUNKDY	−0.536		−0.749	0.0	
SUMDYE	0.000		0.009	0.0	
TOTORG	−0.022		−0.163	0.0	
SEC/TE	−0.664		−0.211	0.0	
LMBDAC	−0.312		−0.127	0.0	

The loadings RD1 of the first component of the RDA on the analytical variables predicting the measured hues (HUEREM and HUEREMAL) are given together with their importance criterion for each variable and the resulting (initial) simplification. A new RDA was performed on the selected variables giving the loadings RD1S.

IDEAL STRATEGY

Buja and Hurley brought up the idea of a programmable pipeline, "which gives mildly sophisticated users the opportunity to concoct their own viewing machinery." We agree that this would be a natural extension of the concept of the OMEGA pipeline, since we also believe in personal styles and creativity in data analysis. We also consider Lisp-Stat (see Tierney, 1989) as a first step in the direction of such a programmable pipeline. On the other hand, we do not consider Lisp to be an attractive computer language.

A nonprogrammable but very attractive computer environment for exploratory data analysis is the JMP product of the SAS-Institute (see Held, Lehmann and Sall, 1989).

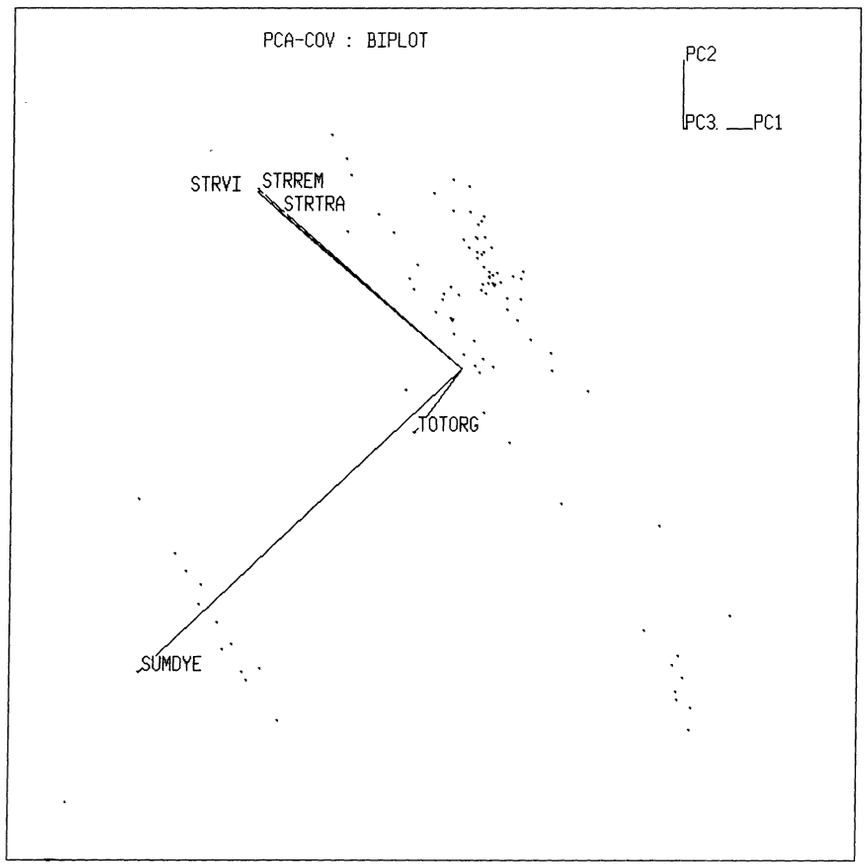

FIG. C. *First two principal components* PC1, PC2: *biplot with variables retained after simplification.*

Finally, Critchley proposed an expert system to guide the user through his voluminous tool box. This really appears to be the ultimate aim, but, looking at other so-called expert systems in statistics might lead one to have some doubts about the practicability of such a project.

CONCLUDING REMARKS

We would like to express our great satisfaction that, independently, Buja and Hurley have extended their original viewing pipeline in a similar direction to ours, and that our thoughts about sensible data analysis strategies appear to be very similar. We would also like to express our great delight at Young's re-presentation of parts of our paper, which illuminates, in a different, very lively way, many of the things we tried to show. In conclusion, we would like to thank the editors of *Statistical Science* for the opportunity to publish a paper exposing a whole concept of data analysis in an applied context.

ADDITIONAL REFERENCES

CRITCHLEY, F. (1985). Influence in principal components analysis. *Biometrika* **72** 627–636.

HELD, G., LEHMANN, A. and SALL, J. (1989). Advances in graphical data analysis from SAS Institute. *Statistical Software Newsletter* **15** 85–90.

HOESKULDSSON, A. (1988). PLS regression methods. *J. Chemometrics* **2** 211–228.

KRZANOWSKI, W. J. (1984). Sensitivity of principal components. *J. Roy. Statist. Soc. Ser. B* **46** 558–563.

TIERNEY, L. (1989). Lisp-Stat: A statistical environment based on the Lisp language. *Bull. ISI Proc. 47th Session*, book 3, 91–104.