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.

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

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.

discover an important feature of their data. To my thinking, they were not so much lucky as they were smart to follow their hunch that PCA would be a good tool to start with.

According to the true nature of data visualization, the PCA showed Weihs and Schmidli something surprising. They checked the finding by resampling and trying again. Still, the same result. They then compared it to the first result by using procrustean rotation. They then tried to simplify the loadings by making some zero and rounding others to 1 significant digit, only to continue seeing the same structure. Apparently, the finding is robust. Not only is the result surprising and robust, but it is also interpretable. The end result is that the authors suggested to the producer of the dyestuff that these results must be a measurement artefact. This was confirmed by the producer, who in turn revised the measurement process. These are very good steps to take, and should be a typical set of steps to follow when exploring data visually. It is particularly refreshing to see all these steps in the real analysis reported. In many cases, this whole initial visual exploration of the data would not be reported.

2. EXPLORING DATA TO CONFIRM HYPOTHESES

One of the most interesting aspects of this paper is that the authors use exploratory tools for confirmatory purposes. They state that

> a graphical test on point symmetry about zero for ··· residuals was carried out to test for bivariate normality. Indeed this hypothesis is supported, since the corresponding strategic random fluctuations appear to be qualitatively the same as in the original plot.

It is very refreshing to see such a description made without qualification or embarrassment. Indeed, it is my experience that when I use exploratory tools (visual or not) I often use them this way, even though, at least in North America, most of us have been taught that one only supports hypotheses by using a specified significance test and obtaining a significant p-value. To me, this is an important point in the philosophy of data analysis. I may very well have a specific structural hypothesis in mind. For example, I might believe that the points in the data space form a specific hierarchical structure, or that they fall in a circle (Young, 1974). I then go looking for that structure in the data. When I see the hypothesized structure, I certainly believe that I have "confirmed" the hypothesis, and it can be very easy to convince others that I have confirmed the hypothesis, even though there is no formal statistical test involved, and even though no p-values are calculated or reported.

A data analyst often explores data with some idea about what type of structure is being sought. The idea, if it exists, is an informal hypothesis, and the process of exploring data is focused on trying to find evidence to "support the hypothesis." The hypothesis is "tested" via the "inter-ocular-impact test": If the evidence "hits you between the eyes," then the hypothesis is supported; otherwise, it is not. This is the route taken by Weihs and Schmidli in Section 5.3, where they have developed an idea (hypothesis) that the three variables measuring color strength are redundant, and that one of these variables is the most fundamental. They then "test" the hypothesis by computing residuals that should be distributed bivariate normally if the hypothesis is correct. The resulting plots "hit them between the eyes," leading them to conclude "this hypothesis is supported." Of course, I have nothing at all against formal statistical tests being used in the ordinary confirmatory way to yield p-values. My point here is simply that it is very nice to see visual data exploration leading to hypothesis confirmation.

3. INDEPENDENCE OF THE MULTIVARIATE FROM THE VISUAL TOOLS

In several places throughout their paper, Weihs and Schmidli mention that the multivariate exploratory data analysis (**MEDA**) techniques which have become fairly familiar to many of us can be visualized by using them in conjunction with many of the **VEDA** techniques. For example, in Section 3.7, where they are discussing resampling and procrustean rotation, they mention that a method for constructing confidence ellipsoids

> can be applied to projections from PCA-COR, CDA, the two groups of variables in CCA individually, and SOG, since in every case projections are represented corresponding to an orthogonal basis.

This is a very important point which should not be overlooked by the reader. The basic point is that all of these methods involve orthogonal projections and that any visualization tool which uses orthogonal projections and which has been proposed in the context of one of these multivariate tools can be legitimately used with others of the multivariate tools mentioned in the quote (and with redundancy analysis, mentioned below). In particular, the rotation and interpolation visualization tools discussed in Section 4.3, and the scatterplot-matrix tool discussed in Section 4.5 can be used with either the original data, or with any linear combinations resulting from one of the above analyses.

4. REDUNDANCY ANALYSIS

Redundancy analysis (RDA) has been presented recently in a very readable article by Lambert, Wildt and Durand (1988) as an alternative to canonical correlation analysis (CCA) (and to multivariate multiple regression, which we will not discuss here). The most complete technical accounts of RDA have been presented by Muller (1981, 1982). Like CCA, RDA can be used when there are two sets of variables. The crucial distinction, however, between CCA and RDA is that CCA is appropriate when neither set of variables is seen to be dependent on the other, whereas RDA is appropriate when one set of variables is being predicted from the other set, the Weihs and Schmidli situation.

The two analyses have the following relationship: CCA estimates *two* sets of mutually orthogonal linear combinations which have the strongest possible association. One set is the "predictor" variates (linear combinations of the "predictor" variables), while the other is the set of "criterion" variates (linear combinations of the "criterion" variables). In no way does CCA explain variance in the criterion variables. Rather, CCA maximizes association between orthogonal linear combinations of one set of variables with orthogonal linear combinations of another set of variables. In fact, since both sets of variables are treated identically, neither set can be viewed as criterion nor predictor (which explains my use of quotation marks in this paragraph).

On the other hand, RDA estimates *one set* of mutually orthogonal linear combinations, the set of predictor variates (no quotation marks) called the redundancy variates. The redundancy variates are computed successively so that each one explains the *maximum proportion of the variance of the entire set of criterion variables* which is unexplained by the previous predictor variates. That is, the first redundancy variate is the linear combination of the predictor variables which explains as much of the variance in the entire set of criterion variables as can be explained by any linear combination of the predictor variables. The next redundancy variate is the linear combination of the predictor variables which is orthogonal to the first and which explains as much of the unexplained variance in the set of criterion variables as can be explained by any linear combination of the predictor variables. Successive redundancy variates have similar properties.

As you might expect, the equations for solving the two problems are quite similar. For RDA, the vectors of standardized weights are the eigenvectors of $R_{xx}^{-}R_{xy}R_{yx}$, arrayed in decreasing order of the magnitude of the eigenvalues (R_{xx} is the correlation matrix for the predictor variables, and R_{xy} and R_{yx} are the intercorrelation matrices of the predictor and criterion variables). However, for CCA the standardized weights are the eigenvectors of $R_{xx}^{-}R_{xy}R_{yy}^{-}R_{yx}$, for the "predictor" variates and the eigenvectors of $R_{yy}^{-}R_{yx}R_{xx}^{-}R_{xy}$ for the "criterion" variates.

In my experience with these two analyses, CCA very often does *not* produce interpretable results, especially when it is misapplied to a set of criterion variables and a set of predictor variables. In contrast, RDA seems to often produce useful results in this situation. Thus, it may be that the relative paucity of results in the sections of Weihs and Schmidli's paper which involve CCA is due to the inappropriate use of CCA with criterion and predictor variables. It would be very interesting to see an RDA of these data.

5. SIMPLIFICATION, BIPLOTS AND TRIPLOTS

Weihs and Schmidli, in Section 2.2, make the point that the linear combinations of the original variables computed by multivariate techniques are often difficult to interpret, especially when the original variables do not measure the same phenomena. They propose a method which they call "simplification" to deal with this problem. This method involves two ways of simplifying the coefficients of the linear combinations. One simplification is to eliminate effects of unimportant variables by changing their coefficients to zero. The other simplification is to round off the coefficients of important variables to a single decimal place.

They propose this method as an alternative to Gabriel's biplot method (1971), a method which is frequently used for multivariate visualization. The biplot is a two-dimensional plot whose dimensions are the first two linear combinations computed by some multivariate technique (i.e., the first two principal components, discriminant variates, canonical variates, redundancy variates, etc.). The biplot displays the scores of the observations on the two linear combinations as points and the coefficients of the variables in the two linear combinations as vectors. The biplot is used frequently because it correctly portrays the geometry of the multivariate analysis: it shows the two-dimensional projection of the high-dimensional data space whose dimensions are the original variables and which contains points for each observation. Sometimes the biplot is three-dimensional, involving the first three linear combinations. Young and Rheingans (1990) have created a video of a dynamic six-dimensional biplot using VISUALS.

While I have no argument with the Weihs and Schmidli simplification technique; I am concerned that biplots are not also used. Weihs and Schmidli imply that there is an inherent conflict between the two techniques. However, they are complimentary and can be profitably used together, as shown by the work 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.