Statistical Science
1987, Vol. 2, No. 4, 396–433

A Review of Multivariate Analysis

Mark J. Schervish

A survey of topics in multivariate analysis inspired by the publication of T. W. ANDERSON, *An Introduction to Multivariate Statistical Analysis*, 2nd ed., John Wiley & Sons, New York, 1984, xvii + 675 pages, $47.50, and WILLIAM R. DILLON and MATTHEW GOLDSTEIN, *Multivariate Analysis: Methods and Applications*, John Wiley & Sons, New York, 1984, xii + 587 pages, $39.95.

This review and discussion are dedicated to the memory of P. R. Krishnaiah, a leader in the area of Multivariate Analysis, who died of cancer on August 1, 1987.

1. INTRODUCTION

It has been a long time coming, but it is finally here. The second edition of T. W. Anderson's classic, *An Introduction to Multivariate Statistical Analysis*, will please all of those who have enjoyed the first edition for so many years. It essentially updates the material in the first edition without going far beyond the topics already included there. A reader who had spent the intervening 26 years on another planet might get the impression that work in multivariate analysis has been concentrated on just those topics with the addition of factor analysis. Of course this impression is mistaken, and Anderson himself notes in the Preface (page vii) that "It is impossible to cover all relevant material in this book." So, in the course of reviewing this book, and comparing it to the first edition, I thought it might be interesting to take a thoroughly biased and narrow look at the development of multivariate analysis over the 26 years between the two editions. A reader interested in a more complete and less personalistic review might refer to Subramaniam and Subramaniam (1973) and/or Anderson, Das Gupta and Styan (1972). Recent reviews of some contemporary multivariate texts (less cluttered by reviewer bias) were performed by Wijsman (1984) and Sen (1986).

Suppose we begin at the end. Nearly simultaneous with the publication of the second edition of Anderson's book is the release of *Multivariate Analysis* by Dillon and Goldstein (the Prefaces are dated June and May 1984, respectively). This text, which is subtitled *Methods and Applications*, is different from Anderson's in every respect except the publisher. It even seems to begin where Anderson leaves off with factor analysis and principal components. I believe that the differences between the texts reflect two very different directions in which multivariate analysis has progressed. The topics covered by Dillon and Goldstein have, by and large, been developed more recently than those covered by Anderson. As an illustration, fewer than 18% of the references cited by Dillon and Goldstein are pre-1958, whereas almost 42% of Anderson's references are pre-1958. (Of course Anderson had a headstart, but the other authors had access to his 1958 book. In three places, they cite Anderson's 1958 book in lieu of earlier work.) The major difference in emphasis is between theory and methods. To illustrate this distinction, Anderson had twelve examples worked out with data in his first edition and the same examples appear in the second edition, with no new ones (but one correction). This is due, in large part, to the fact that the topics covered in the two editions are nearly identical. (Although factor analysis has been added as a topic, no numerical examples are given, and no numerical exercises are included.) Dillon and Goldstein work out numerous examples, often reanalyzing the same data several times to illustrate the differences between various techniques.

Since 1958, the development of multivariate theory has been concentrated, to a large extent, in the general areas that Anderson covered in his first edition. Multivariate methods, on the other hand, have taken on a life of their own, with or without the theory that mathematical statisticians would like to see developed. This has led to an entire industry of exploratory and ad hoc methods for dealing with multivariate data. Researchers are not about to wait for theoreticians to develop the necessary theory when they perceive the need for methods that they think they understand. The theoretical statisticians' approach to multivariate analysis seems to have been to follow the first principle of classical inference. "If the problem is too hard, test a hypothesis." The development of procedures like cluster analysis, factor analysis, graphical

Mark J. Schervish is Associate Professor, Department of Statistics, Carnegie Mellon University, Pittsburgh, Pennsylvania 15213.

methods and the like argue that more tests are not going to be enough to satisfy the growing desire for useful multivariate methods.

2. BACK TO THE BEGINNING

2.1 What's Old

The basic theoretical results with which Anderson began his first edition are repeated in the second edition with only minor clarifications. These include the properties of the multivariate normal distribution and the sampling distributions of the sufficient statistics. They comprise the bulk of Chapters 2 and 3. Dillon and Goldstein deal with all of these concepts in fewer than 12 pages of an appendix. The new material that Anderson adds to Chapter 3 includes the noncentral χ^2 distribution for calculation of the power functions of tests with known covariance matrices. In Chapter 5, he adds a section on the power of tests based on Hotelling's T^2. The pace at which power functions have been calculated for multivariate procedures is very much slower than the pace at which tests have been proposed, even though it does not make much sense to test a hypothesis without being able to examine the power function. For a level α chosen without regard to the power, one could reject with too high a probability for alternatives fairly close to the hypothesis or with too low a probability for alternatives far away without knowing it. (See Lehmann, 1958, and Schervish, 1983, for discussions of this issue in the univariate case.) Multivariate power functions are, of course, much more difficult to produce than are tests. They are also more difficult to understand than univariate power functions. Even in the simple case of testing that the mean vector μ equals a specific value ν based on Hotelling's T^2, the power function depends on the quantity $\tau^2 = (\mu - \nu)^T\Sigma^{-1}(\mu - \nu)$. Just as in univariate analysis, it is rarely (if ever) the case that one is interested in testing that μ exactly equals ν, rather one is interested in how far μ is from ν. If one uses the T^2 test, one is implicitly assuming that τ^2 adequately measures that distance. If it does not, one needs a different test. If τ^2 is an adequate measure, what one needs is some post-data measure of how far τ^2 is likely to be from 0. The posterior distribution of τ^2 would serve this purpose. This posterior distribution is easy to derive in the conjugate prior case. In Chapter 7 (page 270), Anderson derives the posterior joint distribution of μ and Σ. This posterior is given by

$$(1) \qquad \begin{aligned} \mu \mid \Sigma &\sim N_p(\mu_1, 1/\lambda_1\Sigma), \\ \Sigma &\sim W_p^{-1}(A_1, a_1), \end{aligned}$$

where $W_p^{-1}(A_1, a_1)$ denotes the *inverse Wishart distribution* with scale matrix A_1, dimension p, and a_1 degrees of freedom. In words, the conditional distribution of μ given Σ is p-variate normal with mean vector μ_1 and covariance matrix $1/\lambda_1\Sigma$; the marginal distribution of Σ is inverse Wishart. The constants μ_1, λ_1, A_1, and a_1 are functions of both the data and the prior, but their particular values are not important to the present discussion. (For large sample sizes, a_1 and λ_1 are both approximately the size of the sample, whereas μ_1 is approximately the sample mean vector and A_1 is approximately the sample sum of squares and cross-products matrix.) It follows that, conditional on Σ, $\lambda_1\tau^2$ has noncentral χ^2 distribution with p degrees of freedom and noncentrality parameter

$$\eta = \lambda_1(\mu_1 - \nu)^T\Sigma^{-1}(\mu_1 - \nu).$$

The distribution of η is a one-dimensional Wishart or gamma distribution $\Gamma(½a_1, ½\psi^{-2})$, where

$$\psi^2 = \lambda_1(\mu_1 - \nu)^TA_1^{-1}(\mu_1 - \nu).$$

We get the marginal distribution of τ^2 by integrating η out of the joint distribution of τ^2 and η. The result is that the cumulative distribution function of τ^2 is

$$F(t) = \sum_{k=0}^{\infty}\left[\left(\frac{1}{1+\psi^2}\right)^{a_1/2}\left(\frac{\psi^2}{1+\psi^2}\right)^k \frac{\Gamma(k + ½a_1)}{k!\Gamma(½a_1)} \int_0^t \frac{(½\lambda_1)^{k+p/2}}{\Gamma(k + ½p)}\, u^{k+p/2-1}\exp\left(-\frac{\lambda_1}{2}u\right)du\right].$$

This function can be accurately calculated numerically by using an incomplete gamma function program and only a few terms in the summation, because the integral decreases as k increases. Due to the similarity that this distribution bears to the noncentral χ^2 distribution (the only difference being that the coefficients are generalized negative binomial probabilities rather than Poisson probabilities), I will call it the *alternate noncentral* $\chi^2(p, a_1, \psi^2/(1-\psi^2))$, abbreviated *ANC* χ^2. The *ANC* χ^2 distribution was derived in a discriminant analysis setting by Geisser (1967). It also turns out to be the distribution of many of the noncentrality parameters in univariate analysis of variance tests.

In other cases, when τ^2 does not adequately measure the distance between μ and ν, the experimenter will have to say exactly how he/she would like to measure that distance. Perhaps several different measures are important. One thing theoretical statisticians can do is to derive posterior distributions for a wide class of possible distance measures in the hope that at least one of them will be appropriate in a given application. What they are more likely to do is to propose more tests whose power functions depend on parameters other than τ^2. Any movement in this direction, however, would be welcome in that it would force users to think about what is important to detect before just using the easiest procedure.

2.2 What's New

An interesting addition to the chapter on Hotelling's T^2 is Section 5.5 on the multivariate Behrens-Fisher problem. Consider q samples of size N_i, $i = 1, \cdots, q$ from normal distributions with different covariance matrices. The goal is to test H_0: $\sum_{i=1}^{q} \beta_i \mu_i = \mu$. The procedures described amount to transforming the q samples into one sample of size $\min\{N_1, \cdots, N_q\}$ in such a way that the mean of the observations in the one sample is $\sum_{i=1}^{q} \beta_i \mu_i$. The usual T^2 statistic is now calculated for this transformed sample. These methods are classic illustrations of the *level α mindset*, that is, the overriding concern for having a test procedure with prechosen level α regardless of the data structure, sample size or application. Data is discarded with a vengeance by the methods described in this section, although Anderson claims (pages 178), "The sacrifice of observations in estimating a covariance matrix is not so important." Also, the results depend on the order in which observations are numbered. Of course, the posterior distribution of $\sum_{i=1}^{q} \beta_i \mu_i$ is no simple item to calculate, but some effort might usefully be devoted to its derivation or approximation.

One other unfortunate feature of Section 5.5 is the inclusion of what Anderson calls (pages 180) "Another problem that is amenable to this kind of treatment." This is a test of the hypothesis $\mu^{(1)} = \mu^{(2)}$ where

$$\mu = \begin{pmatrix} \mu^{(1)} \\ \mu^{(2)} \end{pmatrix}$$

is the mean vector of a $2q$-variate normal distribution. The test given is a special case of tł e general test of H_0: $A\mu = 0$ with A of full rank. The general test is based on $T^2 = N(A\bar{x})^T(ASA^T)^{-1}(A\bar{x})$ and it neither discards degrees of freedom nor depends on the ordering of the observations. This test is simply not another example of the type of test proposed for the Behrens-Fisher problem.

A topic that has been added to the treatment of correlation is the unbiased estimation of correlation coefficients. This topic illustrates the second principle of classical inference: "Always use an unbiased estimator except when you shouldn't." The case of the squared multiple correlation $\bar{R}^2$ is one in which you shouldn't use an unbiased estimator. When the sample multiple correlation R^2 is near 0, the unique unbiased estimator based on R^2 may be negative. This is not uncommon for unbiased estimators. Just because the average of an estimator over the sample space is equal to the parameter doesn't mean that the observed value of the estimator will be a sensible estimate of the parameter, even if the variance is as small as possible. I would suggest an alternative to the second principle of classical inference: "Only use an unbiased estimator when you can justify its use on other grounds."

3. DECISION THEORY AND BAYESIAN INFERENCE

A welcome addition to the second edition is the treatment of decision theoretic concepts in various places in the text. In Section 3.4.2, the reader first sees loss and risk as well as Bayesian estimation. Admissibility of tests based on T^2 is discussed in Section 5.6. One topic in the area of admissibility of estimators that has been studied almost furiously since 1958 is James-Stein type estimation. Stein (1956) showed that the maximum likelihood estimate (MLE) of a multivariate mean (with known covariance) is inadmissible with respect to sum of squared errors loss, when the dimension is at least 3. Then, James and Stein (1961) produced the famous "shrunken" estimator, which has everywhere smaller risk function. Since that time, the literature on shrunken estimators has expanded dramatically to include a host of results concerning their admissibility, minimaxity and proper Bayesianity. Anderson has added a brief survey of those results in a new Section 3.5. He seems, however, reluctant to recommend a procedure that acknowledges its dependence on subjective information. This is evidenced by his comment (page 91) concerning the improvement in risk for the James-Stein estimator of μ, shrunken toward ν:

> However, as seen from Table 3.2, the improvement is small if $\mu - \nu$ is very large. Thus, to be effective some knowledge of the position of μ is necessary. A disadvantage of the procedure is that it is not objective; the choice of ν is up to the investigator.

Anderson comes so close to recognizing the importance of subjective information in making good inferences, but I will not accuse him of having Bayesian tendencies based on the above remark. It should also be noted, of course, that the choice of the multivariate normal distribution as a model for the data Y is also not objective, and is probably of greater consequence than the choice of ν. For example, if the chosen distribution of Y had infinite second moments and μ were still a location vector, admissibility with respect to sum of squared errors loss would not even be studied seriously.

In addition to the simple shrinkage estimator and its varieties, Anderson reviews such estimators for the mean in the case in which the covariance matrix is unknown (Section 5.3.7) and for the covariance matrix itself (Section 7.8). He also gives the joint posterior distribution of μ and Σ based on a conjugate prior, as well as the marginal posteriors of μ and Σ. He does not give any predictive distributions, for example, the distribution of a single future random vector, or of the average of an arbitrary number of future observations. Unfortunately, he got the covariance matrix of the marginal distribution of μ incorrect. For those of you who are reading along (page 273), the correct formula is $[(N+k)(N+m-1-p)]^{-1}B$. Press (1982) gives a more detailed presentation of a Bayesian approach to inference in multivariate analysis.

Bayesian inference in multivariate analysis has not progressed by anywhere near the amount that classical inference has. An oversimplified reason may be the fact that everyone knows what to do when you use conjugate prior distributions and nobody knows what to do when you don't. There are, however many (perhaps too many) problems that can still be addressed within the conjugate prior setting. There is the issue of exactly what summaries should be calculated from the posterior distribution. The standard calculations are moments and regions of high posterior density. The first principle of Bayesian inference appears to be "Calculate something that is analogous to a classical calculation." The Bayesian paradigm is much more powerful than that, however. Having the posterior distribution theoretically allows the calculation of posterior probabilities that parameters are in arbitrary sets. It also allows the calculation of the predictive distribution of future data, which in turn includes the probabilities that future observations lie in arbitrary sets. These are the sorts of numerical summaries that people would like to see, but the technology needed to supply them is very slow in developing.

One reason for the slow progress in Bayesian methods is the computational burden of performing even the simplest of theoretical calculations. Multivariate probabilities require enormous amounts of computer time to calculate. Also, calculation of summary measures when prior distributions are not conjugate is very time consuming. Programs like those developed by Smith, Skene, Shaw, Naylor and Dransfield (1984) are making such calculations easier, but more effort is needed. Computational difficulties have also hindered the development of power function calculations for multivariate tests. Perhaps breakthroughs in one area will help researchers in the other also.

4. DISCRIMINANT ANALYSIS

Chapter 6 of Anderson, concerning classification has expanded somewhat compared to the first edition, although the introductory sections have remained basically intact. Notation has been altered to reflect standardization. In addition, the formula for the "plug-in" discriminant function W and the formula for the maximum likelihood criterion Z are introduced for future comparison in a new section on error rates. A great deal of work had been done between the two editions in the area of error rate estimation. Some of this work is discussed in Section 6.6, "Probabilities of Misclassification." The presentation consists of several theorems and corollaries giving asymptotic expansions for error rates of classification rules based both on W and on Z for the two population case. In light of the dryness of this section, perhaps the author can be forgiven for failing to discuss any results on error rate estimation in the case of several populations such as the asymptotic expansions given by Schervish (1981a, b). Surprisingly, Dillon and Goldstein say even less about error rate estimation, giving only a verbal description of a few existing methods. This is an area in which recent progress has consisted mainly of the introduction of several methods involving bootstraps, jackknives and asymptotics. The theory behind the methods is a bit sparse, which helps to explain their neglect by Anderson, but not of their shallow treatment by Dillon and Goldstein.

Anderson's treatment of the multiple group classification problem is identical in the two editions, although Dillon and Goldstein adopt the alternative approach based on the eigenanalysis of the matrix $W^{-1}B$, in their notation. In this approach, one tries to find a reduced set of discriminant functions that provides nearly the same discriminatory power as the optimal discriminant functions. For example, if one wishes to use only one discriminant function, one would choose the eigenvector of $W^{-1}B$ corresponding to the largest eigenvalue. Geisser (1977) gives an example illustrating how this first linear discriminant function can lead to poorer classification than other linear functions that are not eigenvectors of $W^{-1}B$. The problem is that discriminatory power (measured by misclassification probability) is not reflected in the squared deviations that the eigenvalues of $W^{-1}B$ measure. Guseman, Peters and Walker (1975) attack the problem of finding optimal reduced sets of discriminant functions for the purposes of classification. A simplified solution in the case of three populations was given by Schervish (1984). The theoretical analysis through the eigenstructure of $W^{-1}B$ is based on (what else?) tests of the hypotheses that successive eigenvalues are 0. I hesitate to mention that the successive tests are rarely performed conditionally on the previous hypotheses being rejected, for fear that someone may then think that this would be an interesting problem to pursue. I was surprised to see Anderson suggesting a similar sort of sequential test procedure in the related problem of determining the number of nonzero cannonical correlations. Anderson does note (page 498) that "these procedures are not statistically independent, even asymptotically." Dillon and Goldstein also give an example (11.2-2, page 405) of this successive unconditional testing. This example is noteworthy for another lapse of rigor which may be even more dangerous. They use V to denote the test statistic and say:

> Because $V = 269.59$ is approximately distributed as χ^2 with $P(K-1) = 5(3) = 15$ df, it is statistically significant at better than the 0.01 level.

Obviously, 269.59 is not approximately χ^2, but neither is V since the hypothesis is most likely false. It seems a bit strange to use the approbatory description "better than" when "less than" is meant. It is as if one were rooting for the alternative. What kind of hypothesis testing habits will a reader with little theoretical statistical training develop if this is the type of example he/she is learning from?

5. EXPLORATORY METHODS

As mentioned earlier, several well known ad hoc procedures have emerged from the need to do exploratory analysis with multivariate data. These procedures can be quite useful for gaining insight from data sets or helping to develop theories about how the data is generated. Theoreticians often think of these procedures as incomplete unless they can lead to the calculation of a significance level or a posterior probability. (This reviewer admits to being guilty of that charge on occasion.) Although some procedures are essentially exploratory, such as Chernoff's (1973) faces, others may suggest probability models, which in turn lead to inferences. I discuss a few of the better known exploratory methods below. Of course, it is impossible to cover all exploratory methods in this review. None of these methods is described in Anderson's book, presumably due to the lack of theoretical results. Dillon and Goldstein give at least some coverage to each topic. Their coverage of cluster analysis and multidimensional scaling is adequate for an introductory text on multivariate methods, but I believe they short change the reader with regard to graphical methods (as does virtually every other text on multivariate analysis). Now that the computer age is in full swing, exploratory methods will become more and more important in data analysis as researchers realize that they do not have to settle for an inferential analysis based on normal distributions when all they want is a good look at the data.

5.1 Cluster Analysis

Cluster analysis is an old topic that has flourished to a large extent in the last 30 years partly due to the advent of high speed computers that made it a feasible technique. It consists of a variety of procedures that usually require significant amounts of computation. It is essentially an exploratory tool, which helps a researcher search for groups of data values even without any clear idea of where they might be or how many there might be. Statistical concepts such as between groups and within groups dispersion have proven useful in developing such methods, but little statistical theory exists concerning the problems that give rise to the need for clustering.

Not surprisingly, some authors have begun to develop tests of the hypothesis that there is only one cluster. Here, one must distinguish two forms of cluster analysis. Cluster analysis of observations concerns ways of grouping observation vectors into homogeneous clusters. It is this form that has proven amenable to probabilistic analysis. The other form is cluster analysis of variables (or abstract objects) in which the only input is a matrix of pairwise similarities (or differences) between the objects. The actual values of the similarity measures often have no clear meaning, and when they do have clear meaning, there may be no suggestion of any population from which the objects were sampled or to which future inference will be applied. In these cases, cluster analysis may be nothing more than a technique for summarizing the similarity or difference measures in less numerical form. As an exploratory technique, cluster analysis will succeed or fail according to whether it either does or does not help a user better understand his/her data.

From a theoretical viewpoint, interesting questions arise from problems in which data clusters. Suppose we define a cluster probabilistically as a subset of the observations that arose independently (conditional on some parameters if necessary) from the same probability distribution. For convenience consider the case in which each of those specific distributions is a multivariate normal and the data all arose in one large sample. We may be interested in questions such as (i) What is probability that there are 2 clusters? (ii) What is the probability that items k and j are in separate clusters if there are 2 clusters. (iii) If there are two clusters, where are they located? Answers to the three questions raised require probabilities that there are K clusters for $K = 1, 2$. They also require conditional distributions for the cluster means and covariances given the number of clusters, and they require probabilities for the 2^n partitions of the n data values among the two clusters given that there are two clusters. There are some sensible ways to construct the above distributions, but the computations get out of hand rapidly as n increases. Furthermore, as the number of potential clusters gets larger than 2 or as the dimension of the data gets large, the theoretical problems become overwhelming. Following the first principle of classical inference Engleman and Hartigan (1969) have proposed a test, in the univariate case, of the one cluster hypothesis with the alternative being that there are two clusters. Although easier to construct than the distributions mentioned, such a test doesn't begin to answer any of the three questions raised above.

5.2 Multidimensional Scaling

Dillon and Goldstein introduce *multidimensional scaling* (MDS) as a data reduction technique. Another way to describe it would be as a data reconstruction technique. One begins with a set of pairwise similarities or differences among a set of objects and constructs a set of points in some Euclidean space (one point for each object) so that the distances between the points correspond (in some sense) to the differences or similarities between the objects (closer points being more similar). If the Euclidean space is two-dimensional, such methods can provide graphical displays of otherwise difficult to read difference matrices. For example, the dimensions of the constructed space may be interpretable as measuring gross features of the objects. Any objects that are very different in those features should be far apart along the corresponding dimension.

There are two types of MDS. When the similarities or differences are measured on interval or ratio scales, then *metric* MDS can be used to try to make the distances between points in the Euclidean representation match the differences between the objects in magnitude. This type of scaling dates back to Torgerson (1952). When the similarities or differences are only ordinal, then *nonmetric* MDS can be used to find a Euclidean representation that matches the rank order of the distances to the rank order of the original difference measures. Shepard (1962a, b) and Kruskal (1964a, b) introduced the methods and computational algorithms of nonmetric MDS. The methodology of both types of MDS is not cluttered with tests of significance or probability models. In its current state it appears to be a purely exploratory technique designed for gaining insight rather than making inference.

5.3 Graphical Methods

Graphical display of multivariate data has been performed for many years. Tufte (1983) gives some excellent historical examples of multivariate displays. Computers have made the display of multivariate data much easier and allowed the introduction of techniques not considered feasible before. Chernoff's (1973) faces are one ingenious example, as are Andrews' (1972) function plots. Such methods are often used as part of a cluster analysis in order to suggest the number of clusters or to visually assess the results of a clustering algorithm. Gnanadesikan (1977) describes several other graphical techniques that can be used to detect outliers in multivariate samples. Tukey and Tukey (1981a, b, c) describe a large number of approaches to viewing multivariate samples, including Anderson's (1957) glyphs and the trees of Kleiner and Hartigan (1981). Most of these techniques require sophisticated graphics hardware and software in order to be used routinely. Their popularity (or lack thereof) is due in large part to both the expense involved in acquiring good graphics equipment and the lack of a widely accepted graphics standard. That is, what runs on a Tektronix device will not necessarily run on an IBM PC or a CALCOMP, etc., unless the software is completely rewritten. Most statisticians (this author included) can think of more interesting things to do than rewriting graphics software to run on their own particular device. Perhaps the graphics kernel standard (GKS) will (slowly) eliminate this problem.

6. REGRESSION

Regression analysis, in one form or another, is probably the most widely used statistical method in the computer age. What would have taken many minutes or hours (if attempted at all) in the early days of multivariate analysis is now done in seconds or less even on microcomputers. Hence, we expect to see some discussion of multivariate regression in any modern multivariate analysis text. Chapter 8 of Anderson's text deals with the multivariate general linear model. The title of the chapter, unfortunately, exposes what the emphasis will be: "Testing the general linear hypothesis; MANOVA." Nevertheless, the treatment is thorough, providing more distributions, confidence regions and tests than in the first edition.

Oddly enough, however, Dillon and Goldstein devote two chapters of their text to multiple regression with a single criterion variable. This is a topic usually covered as part of a univariate analysis course, because only the criterion variable is considered random. But this reasoning only goes to further illustrate the distinction between the theoretical and methodological approaches to statistics. If the observation consists of $(X_1, \ldots, X_p, Y)$, then why not treat it as multivariate? The authors reinforce this point by denoting the regression line $E(Y \mid X)$. In addition to the mandatory tests of hypotheses, they also discuss model selection procedures, outliers, influence, leverage, multicollinearity (in some depth), weighted least squares and autocorrelation. Neither text, however, considers those additional topics in the case of multivariate regression. Gnanadesikan (1977) has some suggestions for how to deal with a few of them. As an alternative to the usual MANOVA treatment of the multivariate linear model, Dillon and Goldstein include a chapter on linear structural relations (LISREL), which I discuss in Section 10.

7. CANONICAL CORRELATIONS

A topic very closely related to multivariate regression, but usually developed separately, is canonical correlation analysis. Anderson develops it as an exploratory technique, being sure to add new material on tests of hypotheses. Dillon and Goldstein introduce the topic by saying (page 337), "The study of the relationship between a set of predictor variables and a set of response measures is known as *canonical correlation analysis.*" It seems clear that they intend this to at least replace any discussion of multivariate regression. What coverage of MANOVA they provide is a special topic under multiple discriminant analysis. Canonical correlation goes one step beyond multivariate regression, however. In regression analysis, the focus is on predicting the criterion variables **Y** from the independent variables **X**. Canonical correlation goes on to ask which linear functions of **Y** can be most effectively predicted by **X**? The canonical variables become those linear functions of **Y** together with their best linear predictors. Because the multivariate regression $\beta\mathbf{X}$ already gives the best linear predictor of **Y**, the **X** canonical variable corresponding to canonical variable $\alpha^T\mathbf{Y}$ turns out to be $\alpha^T\beta\mathbf{X}$ times a normalizing constant.

The theory and methodology of canonical correlation, as described above, has been available for many years. Anderson takes the methodology further by showing how it applies to structural equation models and linear functional relationships. For those unfamiliar with these topics, the introduction of linear functional relationships in Section 12.6.5 will be a bit confusing. It begins, essentially, as follows (page 507):

For example, the balanced one-way analysis of variance can be set up as

$$\mathbf{Y}_{\alpha j} = \nu_\alpha + \mu + \mathbf{U}_{\alpha j}, \alpha = 1, \cdots, m, \qquad j = 1, \cdots, l,$$

where

$$\mathscr{E}\mathbf{U}_\alpha = \mathbf{0}, \mathscr{E}\mathbf{U}_\alpha\mathbf{U}_\alpha^T = \psi, \textstyle\sum_{\alpha=1}^m \nu_\alpha = \mathbf{0}$$

and

$$\Theta\nu_\alpha = 0. \qquad \alpha = 1, \cdots, m,$$

where Θ is $q \times p_1$ of rank q ($\leq p_1$). No mention is given in this discussion of where the matrix Θ comes from or what it means. The inference is that it specifies linear functional relationships, but these have not been part of any discussion of the one-way analysis of variance prior to this point in the text. The discussion of structural equation models and two-stage least squares in Section 12.7 is more coherent and illustrates the author's ingenuity. Although the limited information maximum likelihood estimator introduced there appears ad hoc, it does show that canonical correlation analysis is a bit more versatile than most textbooks give it credit for being. Dillon and Goldstein present a much more grandiose treatment of linear structural relations (LISREL), which I discuss in Section 10.

8. PRINCIPAL COMPONENTS

As mentioned earlier, Dillon and Goldstein begin where Anderson leaves off by discussing principal components. Although both authors give only a brief treatment of this topic, their treatments differ dramatically. Anderson gives asymptotic distributions for the vectors and eigenvalues. He even adds some new discussion of efficient methods of computing the eigenstructure. Other new material includes confidence bounds for the characteristic roots and tests of various hypotheses about the roots. Dillon and Goldstein, in contrast, say next to nothing about how to calculate principal components, aside from the mathematical formulas. They give brief mention of one hypothesis test (lip service to the first principle of classical inference, no doubt). They describe the geometry of principal components in extensive detail, and they present a brief treatment of some ad hoc methods for choosing how many components to keep. The major difference between the two treatments, however, is that Dillon and Goldstein present principal components analysis as one part of a larger factor analysis rather than as a separate procedure.

An interesting alternative derivation and interpretation of principal components is suggested by results of O'Hagan (1984). Let R be the correlation matrix of a random vector **X** that has been standardized so that R is also the covariance matrix. In most treatments, the first principal component is that linear function of **X** that has the highest variance subject to the coefficient vector having norm 1. It also happens to be that linear function whose average squared correlation with each of the X_i's is largest. That is, if $r_i(\mathbf{c}) = \text{corr}(\mathbf{c}^T\mathbf{X}, X_i)$ then the **c**, which maximizes $\sum_{i=1}^n r_i^2(\mathbf{c})$, is the first principal component. So the first principal component is that linear function of **X** that would best serve as a regressor variable if one wished to predict all coordinates of **X** from the same regressor. Suppose now that we regress **X** on the first principal component and calculate the residual covariance matrix. In the residual problem, the second principal component is that linear function of **X** that maximizes the weighted average of the squared correlations with the coordinates of X_i. The weights are the residual variances after regression on the first principal component. That is, the second principal component is the best regressor variable for predicting all of the residuals of the X_i's after regression on the first principal component. The remaining principal components are generated in a similar fashion. The advantages to this approach over the more standard approaches are 2-fold. First, if one wishes to reduce dimensionality, the goal should be to be able to predict the whole data vector as well as possible from the reduced data vector. That this is achieved by principal components is not at all obvious from their derivation as linear functions with maximum variance. Second, there is no need to introduce the artificial constraints that the principal components have norm 1 and that they be uncorrelated or orthogonal. One can scale them any way one wishes for uniqueness, and they are automatically uncorrelated because each one lies in the space of residuals from regression on the previous ones. Hence, the maximization problem one solves for each principal component is identical with all of the others except that the covariance matrix keeps changing. This approach is described in more detail by Schervish (1986).

9. FACTOR ANALYSIS

Factor analysis has been described both as a data reduction technique and as a data expansion technique. The basic goal is to find a small number of underlying *factors* such that the observed variables are all linear combinations of these few factors plus small amounts of independent noise. Because the factors are not observable variables, it turns out that there is a great deal of indeterminacy in any particular factor solution. That is, given a particular solution, there are many alternative solutions that produce the very same estimated covariance structure for the observed variables, but with different factors. Some arbitrary restrictions must be placed on the solution in order to obtain a unique answer. Chapter 14 of Anderson's second edition is all new and contains a good exposition of the maximum likelihood approach to factor analysis. This is the only approach in which statistical theory has played an important role. It includes a particular arbitrary restriction that allows calculation of a unique solution.

9.1 Exploratory Factor Analysis

There are traditionally two modes in which one can perform factor analysis. First, there is *exploratory factor analysis*. In this mode, one is trying to determined both how many (if any) factors there are and what they mean, if there are any. Once one has fit a model with a specific number of factors, one can rotate the factors through all of the equivalent solutions by using any of several exotically named techniques. With the maximum likelihood approach, one can also test the hypothesis that there are only m common factors where m is smaller than the dimension of the observation vectors. If the test rejects the hypothesis, one is free to add more factors until the result is insignificant. This practice is deplorable in the usual hypothesis testing framework, although I am sure that some unfortunate person somewhere is currently trying to solve the problem of determining the level of this procedure, or sequences of critical values to guarantee a specified level. Because it is never conclusively decideable how many factors there are in a given application, it would be worth while to have a model that would allow calculation of the probability distribution of the number of factors. This would require subjective information about the factor structure.

Consider the example analyzed in Section 3.4 of Dillon and Goldstein by both the principal factor method and maximum likelihood. The example concerns ten political and economic variables measured on 14 countries. Dillon and Goldstein present a principal factor solution with four factors and a maximum likelihood solution with three factors. The fourth principal factor contributes almost as much to the solution as does the third. But Dillon and Goldstein claim that the likelihood ratio test of the three-factor model (using the maximum likelihood method) produces a χ^2 value of 20.36 with 18 degrees of freedom, and accepts the model at any commonly used α level. They do not report the result of a test of the two-factor model, and they claim that the fitting of a four-factor model failed to converge. I used BMDP4M (cf. Dixon, 1985) to fit the two- three- and four-factor models so that I could compare them. Unfortunately, I was unable to reproduce Dillon and Goldstein's results. The two- three- and four-factor models converged in 7, 17 and 8 iterations, respectively. The χ^2 values were 50.475, 38.400 and 19.857 for two, three, and four factors, respectively, with 26, 18 and 11 degrees of freedom. (Note that BMDP4M does not calculate the χ^2 value so I had to work with the output, which was rounded to three digits. Hence, some rounding error has been introduced into my calculation. I used both the raw data and the correlation matrix and got similar results.) The results of the three-factor fit with a varimax rotation are given in Table 1. The results of the four-factor fit with a varimax rotation are given in Table 2.

The point of this example is to illustrate the difficulty one has in determining the number of factors. The hypothesis test is not conclusive (regardless of whether Dillon and Goldstein's or my calculations are correct). The fourth factor in Table 2 is certainly not easy to interpret, but does that mean that we should believe there are only three factors? The fourth factor contributes 84% as much variance as does the third factor. One has to look carefully at the meanings of

TABLE 1
Maximum likelihood solution with 3 factors and varimax rotation

	Factor		
Variable	1	2	3
1	0.846	0.298	0.338
2	0.870	0.471	0.145
3	0.769	0.010	−0.095
4	0.442	0.141	0.658
5	−0.102	0.929	0.356
6	0.510	−0.375	0.224
7	0.237	0.754	0.192
8	0.814	−0.076	0.241
9	0.341	−0.254	−0.034
10	−0.038	0.288	0.823

TABLE 2
Maximum likelihood solution with 4 factors and varimax rotation

	Factor			
Variable	1	2	3	4
1	0.588	0.570	0.352	0.444
2	0.714	0.606	0.177	0.219
3	0.993	−0.100	−0.007	0.055
4	0.275	0.302	0.572	0.204
5	0.005	0.750	0.360	−0.555
6	0.212	−0.105	0.169	0.547
7	0.028	0.929	0.112	−0.014
8	0.653	0.181	0.212	0.510
9	0.042	0.028	−0.159	0.444
10	0.013	0.140	0.970	−0.200

the variables and try to imagine what, if anything, could contribute to the variables in the proportions given by each of the columns. If this is not possible, rotate the factors and try again. When done, one may have a deeper understanding of the data set or even have developed a new theory for explaining the data. One does not (in this case at least) have a conclusion as to how many factors there are. I am beginning to understand why Anderson did not include any numerical examples of factor analysis in his second edition.

9.2 Confirmatory Factor Analysis

In the second mode of operation, namely *confirmatory factor analysis*, one hypothesizes a factor structure of a particular sort and then uses the data to find the best fitting model satisfying the hypothesized structure. The specified structure may be extremely specific (going so far as to specify all of the factor loadings) or less specific, such as only saying that some loadings are required to be zero. In general, confirmatory analysis does not permit arbitrary rotations of the factors because the specified structure might be destroyed by the rotation. After fitting the model, one is compelled to test the hypothesis that the model fits, presumably by using the likelihood ratio test. Dillon and Goldstein present an example of this procedure in Section 3.8.5. The example concerns eleven variables on $n = 840$ subjects and three hypothesized factors with certain specified loadings equal to zero. They calculate the likelihood ratio χ^2 statistic as 50.99 with 35 df ($p = 0.0395$) and claim (page 104), "The fit of this model is not satisfactory." First of all, a χ^2 value so close to the degrees of freedom with $n = 840$ is not bad if the hypothesized model has any a priori credibility. Aside from this often neglected point, one must ask, "Then what?" Dillon and Goldstein fit a second model with comparable results and conclude (page 106) "that the data do not confirm the *a priori* assumptions about their structure." I suggest that this is more a failure of the hypothesis testing mentality than of the hypothesized model. I took the same data and used BMDP4M to find the unrestricted maximum likelihood solution with three factors and a varimax rotation. The χ^2 statistic was 26.99 with 25 df (I refuse to look up the p-value). This is presumably a pretty good fit. The solution bares a good deal of resemblance to the hypothesized solution and only has high loadings in two of the thirteen places hypothesized to be zero. This is not to say that the hypothesis should be accepted, but rather that one should not (just) calculate the p-value and ignore how close the data really are to the hypothesis.

9.3 Interpretation

As an exploratory technique, factor analysis is as good as the insights its users gain from using it. As an inferential technique, however, it suffers from a lack of predictive validity. One cannot observe factor scores and then predict observables. However, there is no arguing the fact that the statement of the factor analysis problem is very appealing intuitively. Large sets of moderately correlated variables probably have some common structure, the discovery of which might shed considerable light on the process generating the variables. What seems so mystifying about factor analysis is how that discovery occurs. After forming a factor solution, one is still left with the question of whether the original variables are linear combinations of the factors or if the factors are just linear combinations of the original variables. Certainly the estimated factor scores are just linear combinations of the original variables. If these later prove useful in some as yet unspecified problem, it may still be the original variables and not the hypothesized factors that are doing the work. Put more simply, the way the common factor model is implemented, it is as if the user is regressing the original variables on each other to find a few best linear predictors. This is essentially what principal components analysis does, and that is why the two methods are often used for similar purposes. This discussion is not intended to discourage or denigrate work in the area of factor analysis, but rather to encourage those, who feel that the common factor model has something to offer, to develop experiments in which the use of that model can be distinguished from regression.

10. PATH ANALYSIS AND LISREL

The path analysis and LISREL models are generally not well known to mathematical statisticians, because they are most commonly discussed in writings by and for psychometricians. In this section, I present a very cursory overview of the ideas underlying these models and some examples of how they can be used and misused.

10.1 Path Analysis

When dealing with a large collection of variables, it is very useful to sort out which of them one would like to be able to predict from which other ones. The same variables may play the role of predictor in one situation and criterion in another. The power of multivariate analysis is its ability to treat joint distributions, not just conditional ones like traditional regression analysis. Hence, the initial stages of a path analysis can be quite useful. A diagram illustrating which variables one thinks influence which others, and which ought to be correlated with each other can help one to organize the analysis more sensibly. (See Darroch, Lauritzen and Speed, 1980, for an introduction to general graphical models. Also, see Howard and Matheson (1981) and Shachter (1986) for descriptions of how influence diagrams can be used to model subjective probabilistic dependence between variables. Spiegelhalter (1986) and Lauritzen and Spiegelhalter (1987) show how such diagrams can be useful in expert systems.)

What I would object to in the practice of path analysis are the attempts to interpret the coefficients placed along the path arrows. Take the following trivial example in which two correlated exogenous variables X_1, X_2 are thought to influence the endogenous variable Y. The residual of Y is e_Y. The notation is borrowed from Dillon and Goldstein (Chapter 12). Figure 1 is a typical path diagram. The single-headed arrows denote effect or causation, whereas the double-headed arrows denote correlation. Suppose all three variables have variance one and intercorrelations of 0.9. Without going into details, the path coefficients would be as follows:

$$r_{X_1X_2} = .9, \qquad P_{YX_1} = .4737,$$
$$P_{YX_2} = .4737, \qquad P_{e_Y} = .2768.$$

One would be led, by the path analysis methodology, to interpret $P_{YX_1} = .4737$ as the direct effect of X_1 on Y.

The remainder of the correlation between Y and X_1 is $.9 - .4737 = P_{YX_2} r_{X_1X_2} = .4263$ and is attributed to "unanalyzed effects." (If X_2 had not been in the picture, P_{YX_1} would equal .9 and there would be no unanalyzed effects.) Suppose, that we know that $X_2 = X_1 + Z$, and we set $X_3 = \sqrt{5}Z$ (the standardized version of Z). Then $r_{X_1X_3} = -.2236$. Replacing X_2 by X_3 in the path analysis leads to the following path coefficients:

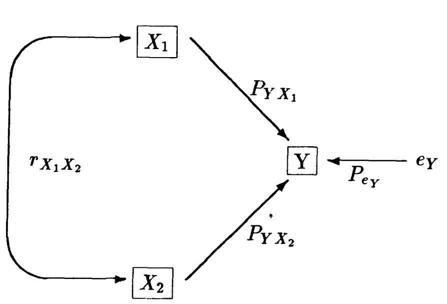

FIG. 1. *A path diagram.*

$$r_{X_1X_3} = -.2236, \qquad P_{YX_1} = .9474,$$
$$P_{YX_3} = .2120, \qquad P_{e_Y} = .2768.$$

Now the direct effect of X_1 is .9474 and the unanalyzed effect is $-.0474$. For a simple path diagram like Figure 1, such an ambiguous definition of "direct effect of X_1 on Y" is easy to understand. But in more complicated analyses, such ambiguity will affect and indirect effects of X_1 on variables in other parts of the diagram making any interpretation tenuous at best.

Of course, the ambiguity of regression coefficients is not news to most readers. For this reason, it is surprising that Dillon and Goldstein do not mention multicolinearity as one of the potential drawbacks to such models. Statisticians constantly tell their students to be careful not to interpret a regression coefficient as measuring the effect of one variable on another when the data arises from an observational study. It is not even the effect of one variable on the other ceteris paribus. In the example above, it would be impossible to vary X_1 while keeping X_2 and X_3 fixed. The only safe interpretation of a regression coefficient is simply as the number you multiply by the independent variable X_i in a *specific* regression model to try to predict the dependent variable Y, assuming that Y and the X_i's all arise in a fashion similar to the way they arose in the original data set. When the variables all arise in a designed experiment, in which each X_i is fixed at each of several values and the other X_j are chosen equal to one of their several values, then the interpretation is clearer due to the way the data arose. If one now fixes all of the X_i but one, the coefficient of the other variable does measure how much we expect the response to change for one unit of change in that variable (assuming the change occurs in a manner consistent with how the variable changed in the experiment). If, on the other hand, one merely observes the X_i for a new observation and then wishes to predict Y, based on the results of a designed experiment, one has the problem of assuming that the conditions of the experiment were sufficiently similar to those under which the new observation is generated. This is closely related to Rubin's (1978) notion of *ignorable treatment assignments*. The basic question to be answered is, "What effect, if any, does a deliberate intervention to affect the exogenous variables have on the relationship between the endogenous and exogenous variables?" This question can only be addressed by people with significant subject matter knowledge.

10.2 Linear Structural Relations

A more general method for analyzing path diagrams is the LISREL model for linear structural relations. This model is quite general and allows the fitting of hybrids of factor analysis and general linear models. Its generality also makes it very easy to misuse, however. In Section 12.5.3, Dillon and Goldstein consider an example borrowed from Bagozzi (1980). The goal of the example was (Bagozzi, 1980, page 65) "... to discover the true relationship between performance and satisfaction in an industrial sales force." More specifically (same page) "... four possibilities exist: (1) satisfaction causes performance, (2) performance causes satisfaction, (3) the two variables are related reciprocally, or (4) the variables are not causally related at all and any empirical association must be a spurious one due to common antecedents." The linear structural relations are stated in terms of latent variables ξ_1 = achievement motivation, ξ_2 = task specific self esteem, ξ_3 = verbal intelligence, η_1 = performance, η_2 = job satisfaction. The exogenous latent variables ξ_i are introduced as possible "common antecedents." Based on the above statement of goals, one would now expect to see models in which η_1 and η_2 were causally related to each other along with models in which they were causally unrelated, but in which causal effects existed from the ξ_i to the η_i. The initial model of Bagozzi (1980) is described by the equation

$$(2)\quad \begin{pmatrix} 1 & -\beta_1 \\ -\beta_2 & 1 \end{pmatrix}\begin{pmatrix} \eta_1 \\ \eta_2 \end{pmatrix} = \begin{pmatrix} 0 & \gamma_2 & \gamma_3 \\ \gamma_1 & 0 & 0 \end{pmatrix}\begin{pmatrix} \xi_1 \\ \xi_2 \\ \xi_3 \end{pmatrix} + \begin{pmatrix} \zeta_1 \\ \zeta_2 \end{pmatrix},$$

where the ζ_i are disturbance terms and the matrix multiplying the η's is assumed nonsingular. This equation is the algebraic representation of the path diagram in Figure 2. Figure 2 is the portion of the path diagram that concerns the latent variables only. The observed variables can be appended with more arrows to make a much more impressive diagram. The paths in Figure 2 with coefficients β_1 and β_2 represent reciprocal causation between η_1 and η_2. The covariance matrices of the ξ and η vectors are, respectively,

$$\begin{pmatrix} \phi_{11} & & \\ \phi_{21} & \phi_{22} & \\ \phi_{31} & \phi_{32} & \phi_{33} \end{pmatrix} \quad\text{and}\quad \begin{pmatrix} \psi_{11} & \\ \psi_{21} & \psi_{22} \end{pmatrix}.$$

Bagozzi deletes those paths with coefficients β_1 and ψ_{21} because the estimates are not significant at level .05 and arrives at his final model. It has a likelihood ratio χ^2 of 15.4 with 15 degrees of freedom and is depicted in Figure 3. Because the β_1 coefficient is estimated to be zero (more precisely, because the hypothesis that $\beta_1 = 0$ is not rejected), Bagozzi claims (page 71), "Perhaps the most striking finding is that job satisfaction does not necessarily lead to better performance." He then goes on to offer advice to management based on this finding, such as (page 71) "... resources should be devoted to enhancement of job satisfaction only if this is valued as an end in and of itself ... " Bagozzi appears to have fallen into a common trap described by Pratt and Schlaifer (1984, page 14) (but presumably known in 1980):

> Exclusion of a regressor because it contributes little to R^2 or because its estimated coefficient is not statistically significant may make sense when one wants to predict y given a naturally occurring $\mathbf{x}$, but not when one wants to know how two or more x's affect y. Here it implies that if the data provide very little information about the separate effects of two factors, it is better to attribute almost all of their joint effect to one and none to the other than to acknowledge the unavoidable uncertainty about their separate effects.

As an example of how to fit a specified LISREL model, the Bagozzi example is excellent in that it illustrates several features of the model and allows comparison of the initial and final models. As an example of how causal analysis should be done, however, I find this example disappointing. First of all, it was an expressed goal of the project to see if common antecedents can explain the association between performance and satisfaction. No causal models involving only paths from common antecedents were described

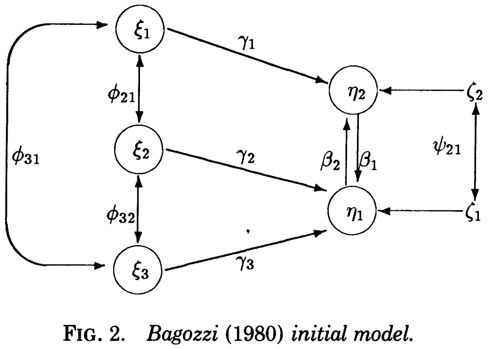

FIG. 2. *Bagozzi* (1980) *initial model.*

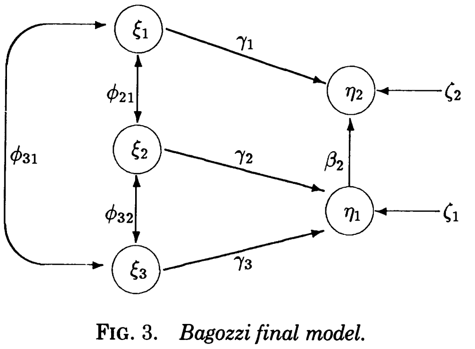

FIG. 3. *Bagozzi final model.*

in the example. Some hypothesis tests on the partial correlation between performance and satisfaction given some other variables were performed, but the other variables did not include all three of the ξ variables. In fact, there are models involving no causal arrows between performance and satisfaction, which are *equivalent* (not just similar) to the models in (2). It is well known that, in many cases, several causal models are equivalent in the sense that the parameters are one-to-one functions of each other. As an example, the following model is equivalent to (2):

$$(3)\qquad \begin{pmatrix}\eta_1\\ \eta_2\end{pmatrix} = \begin{pmatrix}\alpha_1 & \alpha_2 & \alpha_3\\ \alpha_4 & c\alpha_1 & c\alpha_3\end{pmatrix}\begin{pmatrix}\xi_1\\ \xi_2\\ \xi_3\end{pmatrix} + \begin{pmatrix}\zeta_1^*\\ \zeta_2^*\end{pmatrix}.$$

The model of (3) is is not linear in the parameters, hence, it cannot be fit with the computer program LISREL IV of Jöreskog and Sörbom (1978), nor can it be fit with the EQS program of Bentler (1985). However, it can be fit via straightforward maximum likelihood. The equations relating the two models are

$$\beta_1 = \alpha_1/\alpha_4,$$

$$\beta_2 = c,$$

$$\gamma_1 = \alpha_4 - c\alpha_1,$$

$$\gamma_2 = \alpha_2(1 - c\alpha_1/\alpha_4),$$

$$\gamma_3 = \alpha_3(1 - c\alpha_1/\alpha_4),$$

$$\begin{pmatrix}\psi_{11} & \\ \psi_{21} & \psi_{22}\end{pmatrix} = \begin{pmatrix}1 & -\beta_1\\ -\beta_2 & 1\end{pmatrix}\begin{pmatrix}\psi_{11}^* & \\ \psi_{21}^* & \psi_{22}^*\end{pmatrix}\begin{pmatrix}1 & -\beta_2\\ -\beta_1 & 1\end{pmatrix},$$

with some restrictions on the parameters. The model (3) corresponds to the path diagram in Figure 4. Notice that there are no paths between η_1 and η_2, although there are extra paths from the ξ_i to the η_j. In this model, the η variables are not causally related, but are both affected by the three common antecedents. One could just as easily start with a model of this sort and delete paths until one had a model that made sense and fit acceptably. The final model would lead to different conclusions from the model that Bagozzi arrived at, and one would be hard pressed to distinguish them based on the data.

As an example, I replaced the coefficients $c\alpha_2$ and $c\alpha_3$ in Figure 4 with α_5 and α_6, respectively, so that I could use the program EQS of Bentler (1985). The model had a likelihood ratio χ^2 of 9.3 with 12 degrees of freedom. To fit a model more like the final model of Bagozzi, I set $\psi_{21}^* = 0$ and $\alpha_4 = 0$ and got a likelihood ratio $\chi^2 = 14.2$ with 14 degrees of freedom. If I set $\alpha_6 = 0$, I get $\chi^2 = 16$ with 15 degrees of freedom. This last model is depicted in Figure 5. All of these models (the ones depicted in Figures 2 to 5) fit the data comparably with an average absolute difference between the observed and fitted correlations of about 11% of the average absolute correlation. The model of Figure 5 is not equivalent to that of Figure 3, due to the deleted paths, and the causal conclusions that would be drawn from the two models would be different. Because I do not claim to be an expert in management science, I will not begin to offer advice to managers. Nor will I recommend the model of Figure 5 over that of Figure 3. (In fact, I would recommend that neither model be used for causal inference, but rather only for prediction, as suggested by Pratt and Schlaifer.) But I would offer advice to users of structural equation models: Don't start drawing conclusions from your models until you have spent more time looking at alternative but nearly equivalent models that have different causal links. (See Glymour, Scheines, Spirtes and Kelly, 1987, for a description of one way to examine alternative causal models.)

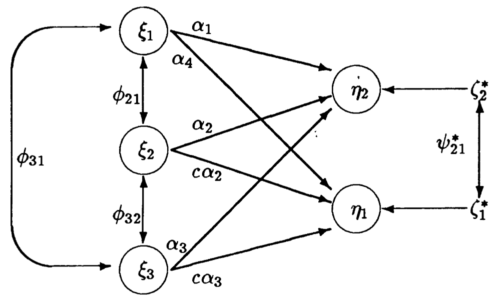

FIG. 4. *Model equivalent to Bagozzi initial model.*

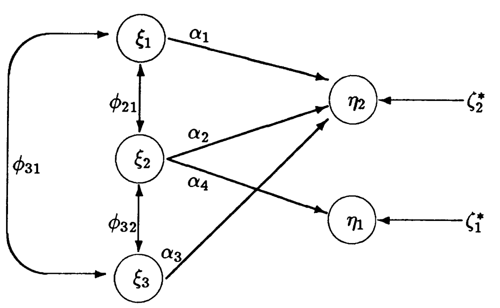

FIG. 5. *Final model with no causation between η's.*

10.3 Interpretations

The issue of how to detect causation is a difficult one. Philosophers have been arguing about it for centuries, and I do not propose to settle it here. Holland (1986) describes a precise but narrow view of how to define and detect causation. Pratt and Schlaifer (1984) offer a different account of causation in statistical models. The discussions of these papers suggest that we are no closer to understanding causation than were Aristotle and Hume. Fortunately, the sensible practice of statistical techniques does not require that one even pretend to have an understanding of causation. It is in the various subject matter disciplines in which statistics is used that researchers can attempt to model and understand causation. Take Bagozzi's model for example. It may or may not be reasonable within the various theories of management science to model a causal relationship between the various latent constructs described in the example. The statistical methods merely give you ways to quantify your uncertainty about those relationships, given that you believe a particular model for the generation of the data. It is the beliefs about those relationships, whether stated explicitly or implied by the form of the model, that express the causal relations. Two different researchers who believed strongly in two different, but predictively equivalent, causal models for the data could collect data for an eternity and never be able to distinguish the two models based on the data. Only by arguing from subject matter considerations (or designing different experiments) would they be able to conclude that one model is better supported than the other. Perhaps the Bagozzi example is an isolated instance, but I would remind the reader that Bentler (1985) also presents it as an example of the use of EQS. If this example is being singled out as exemplary or prototypical, then those who teach the use of LISREL models to their students ought to look for some better examples.

The most important thing which Dillon and Goldstein have to say about the use of the LISREL model is contained in a paragraph at the end of Chapter 12 entitled "Indeterminacy":

> If the analysis is data driven and not grounded in strong a priori theoretical notions, it is always possible to find an acceptable χ^2-fit, and it is always possible to find several models that fit the data equally well. Thus, in the absence of theoretical knowledge, covariance structure analysis becomes a limitless exercise in data snooping contributing little, if anything, to scientific progress. It is a simple fact that exploratory analysis is better performed by other methods that impose fewer restrictive assumptions [e.g., principal components analysis (Chapter 2)].

It is possible, of course, to make use of structural equation models without getting hog-tied by the ambiguity of causal interpretations. By making only predictive inferences, one gives up the compulsion to draw causal inferences from observational data and concentrates on simply modeling the joint distributions of the unknown quantities. For example, if I were to learn the value of η_2 "job satisfaction" for a salesperson selected from a population like that in this study, then what would be the (conditional) distribution of η_1 "performance"? Which one "causes" the other is not an issue. In fact, Lauritzen and Spiegelhalter (1987) drop the directional arrows from the paths in their graphical models to further emphasize that inference is a two-way street. One can condition on whatever variables become known and make inference about the others. On the other hand, if I need to make some policy decisions as to whether to try to increase job satisfaction or something else in the hopes of affecting performance, I must raise the question of whether the associations of the variables measured in the observational study remain the same when I intervene with new policies. This is a subject matter question that mere statistics alone cannot address (at least not without a different data set). Such issues do not invalidate the use of structural equations models, but rather, they make it clear that it is irresponsible to teach causal modeling without preparing the students to make the appropriate subject matter judgments.

11. TESTING HYPOTHESES

As mentioned earlier, a great deal of the theoretical research performed in multivariate analysis since 1958 has been in the area of hypothesis testing. Hence, it is not surprising that Chapters 8, 9 and 10 of Anderson's book have been substantially rewritten. These chapters consider testing everything under the sun. Discussion of more invariant tests has been added, where just the likelihood ratio tests were discussed before. Distributions of the test statistics have been developed in the intervening years and these are given for all of the tests considered. New results on admissibility of tests and properties of power functions have been included. There is also an expanded treatment of confidence regions. A remark from the Preface of the first edition seems to have been adopted as a battle cry by an entire generation of multivariate researchers: "In many situations, however, the theory of desirable or optimum procedures is lacking." Unfortunately, the emphasis has been on the procedures and not on the desirability and/or optimality of them. The result is that the likelihood ratio criterion has been augmented by a battery of uniformly most wonderful invariant tests and confidence regions.

One possible explanation for the plethora of invariant multivariate tests, despite their dubious inferential relevance is the fact that the distributions of the test statistics depend only on the small dimensional maximal invariant, and are therefore easier to derive mathematically. Power function calculations are largely ignored, even when they are available, because the maximal invariant is generally not the parameter of interest to the researcher who collected the data. Ease of derivation is also a reason why so much of the Bayesian methodology in multivariate analysis relies on conjugate priors. This situation is reminiscent of the following story of a man who lost his room key:

> A man lost his room key one night and began searching for it under a street lamp. A police officer happened by and began to help him look.
> *Officer*: What are you looking for?
> *Man*: My room key. I heard it drop from my key-chain.
> *Officer*: Where were you standing when you heard it drop?
> *Man*: About half-way up the next block.
> *Officer*: Then why are you looking for it here?
> *Man*: Because the light is better under the street lamp.

In multivariate analysis (if not in the entire field of statistics), we have taken to solving problems because we can solve them and not because somebody needs the solution. If a problem is hard to solve, it makes more sense to try to approximate a solution to the problem than to make up and solve a problem whose solution nobody wants. The theory of invariant tests is elegant mathematically, but it does not begin to address the questions of interest to researchers, such as "How much better or worse will my predictions be if I use model B instead of model A?" or "To what extent has the treatment improved the response and how certain can I be of my conclusion?"

This point about the relevance of the maximal invariant parameter was raised in Section 2.1 with regard to Hotelling's T^2. As Lehmann (1959, page 275) puts it:

> When applying the principle of invariance, it is important to make sure that the underlying symmetry assumptions really are satisfied. In the problem of testing the equality of a number of normal means $\mu_1, \ldots, \mu_s$, for example, all parameter points, which have the same value of $\psi^2 = \sum n_i(\mu_i - \mu.)^2/\sigma^2$, are identified under the principle of invariance. This is appropriate only when these alternatives can be considered as being equidistant from the hypothesis. In particular, it should then be immaterial whether the given value of ψ^2 is built up by a number of small contributions or a single large one. Situations where instead the main emphasis is on the detection of large individual deviations do not possess the required symmetry, . . .

The justification for the use of invariant procedures has always been mystifying. Anderson (page 322) gives the only legitimate reason of which I am aware for using invariant procedures: "We shall use the principle of invariance to reduce the set of tests to be considered."

Perhaps in the next 26 years, those who feel compelled to develop tests for null hypotheses will at least enlarge their horizons and consider tests whose power functions depend on more general parameters that might be of interest in specific applications. Implicit also is the hope that the deviation of the power function will be treated as equal in importance to the derivation of the test.

But, it will take more than a new battery of variant (opposite of invariant?) tests to get the focus of multivariate analysis straight. The entire hypothesis testing mentality needs to be reassessed. The level α mindset has caused people to lose sight of what they are actually testing. The following example is taken from one of the few numerical problems worked out in Anderson's text (page 341) and is attributed to Barnard (1935) and Bartlett (1947). It concerns $p = 4$ measurements taken on a total of $N = 398$ skulls from $q = 4$ different periods. The hypothesis is that the mean vectors $\mu^{(i)}$ for the four different periods are the same. Anderson uses the likelihood ratio criterion $-k \log U_{p,q-1,n}$, where $n = N - q$ and $k = n - ½(p - q + 2)$, and writes (page 342):

> Since n is very large, we may assume $-k \log U_{4,3,394}$ is distributed as χ^2 with 12 degrees of freedom (when the null hypothesis is true). Here $-k \log U = 77.30$. Since the 1% point of the χ^2_{12} distribution is 26.2, the hypothesis of $\mu^{(1)} = \mu^{(2)} = \mu^{(3)} = \mu^{(4)}$ is rejected.

The corresponding coordinates of the sample mean vectors do not differ very much compared to the sample standard deviations. If we were to consider the problem of sampling a new observation and classifying it into one of the four populations, we could calculate the correct classification rates for the four populations (assuming a uniform prior over the four populations). By using the asymptotic expansions of Schervish (1981a), we get the results in Table 3. The reason these numbers are so small (we could get 0.25 by just guessing), despite the low p-value for the hypothesis, is that the mean vectors are actually quite close. The square roots of the estimated Mahalanobis distances between the pairs of populations $(\bar{y}^{(i)} - \bar{y}^{(j)})^T\hat{\Sigma}^{-1}(\bar{y}^{(i)} - \bar{y}^{(j)})$ are given in Table 4. Population 4 does seem to be uniformly separated from the others, accounting for it having the largest correct classification rate. Even so, it is no more than one estimated standard deviation (in the observation scale) from any of the other three populations. A one standard deviation difference between

TABLE 3
Estimated correct classification rates

Population	1	2	3	4
Estimated rate	0.41	0.32	0.30	0.54

TABLE 4
Pairwise estimated distances

	Population		
Population	1	2	3
2	.653		
3	.632	.407	
4	.986	.946	.923

two populations allows a correct classification rate of 0.69 compared to the 0.5 you would get by mere guessing. On the average, the correct classification rates are not much larger than what one could obtain by guessing, except for population 4. Simply rejecting the hypothesis does not tell the story of how little the mean vectors differ. The low p-value is due as much to the large sample size as it is to the differences between the mean vectors.

12. DISCRETE MULTIVARIATE ANALYSIS

Some people do not consider categorical data analysis as "multivariate." Bishop, Fienberg and Holland (1975) are notable exceptions. Anderson does not say a word about it. Nor does he even acknowledge it as a multivariate topic which he will not cover. On the other hand, Dillon and Goldstein devote two chapters to discrete multivariate analysis. These two chapters, however, are as distant in approach as they are in location in the book (Chapters 8 and 13). The earlier chapter discusses classical methods like χ^2 tests and log-linear models. The later chapter describes an approach more familiar to psychometricians, namely latent structure analysis.

Latent structure analysis attempts to construct an additional discrete unobserved variable X, whose values are called *latent classes*, to go with the observed categorical variables Y_i. The Y_i, in turn, are modeled as conditionally independent given X. This sounds a lot like the construction of factors in factor analysis. In fact, latent class modeling is actually quite a bit like discrete factor analysis. In particular, it shares some of, but not all of, the identifiability problems of factor analysis. Take the first example of Chapter 13 given in Table 13.1-1 on page 492 of Dillon and Goldstein. It is a hypothetical two-way table exhibiting significant dependence between rows and columns. Below the table are two subtables corresponding to levels of a third (unobserved) variable (in this case education). In each of the subtables, the two observed variables are independent. This is an example of conditional independence of two categorical variables given a third latent variable. Because the actual tables given by Dillon and Goldstein have errors in them (for example, the subtables do not add up to the aggregate table and one of the subtables does not have the rows and columns independent, I have revised the data as little as possible to make them correspond to the description above. The data are in Table 5. I have converted the subtables to probabilities, so as to avoid the embarassment of fractional persons. The subtables give the conditional probabilities given the corresponding level of the latent variable. The probability in the lower right corner of each subtable is the marginal probability of that latent class.

The strange feature of this example is that it would be impossible to use the latent class modeling methodology to arrive at the solution given in Table 5 without placing arbitrary restrictions on the parameters of the solution. The reason is that a latent class model with two latent classes is nonidentifiable in a $2 \times K$ table. Such a model would require $4K - 3$ parameters to be estimated, whereas there are only $2K - 1$ degrees of freedom in the table. The nonidentifiability in this example is disguised by the fact that the latent classes have been named "Low Education" and "High Education" and corresponded to actually observable variables. Had they been unspecified, as in most problems in which latent class modeling is applied, the user would have had a two-dimensional space of possible latent classes from which to choose. With expressed prior beliefs, about the classes, one can at least find an "average" solution by finding the posterior mean, say, of the cell probabilities under the model. For example, suppose I have a uniform prior over the five probabilities p_1, the probability of being in the first class (assumed less than 0.5 for identifiability), $p_{T|1}$, the conditional probability of reading the Times given class 1, $p_{D|1}$, the conditional probability of reading the Daily News given class 1, and $p_{T|2}$ and $p_{D|2}$ similarly defined for Class 2. The posterior means of the conditional and marginal cell counts are given in Table 6. The estimation was done by using the

TABLE 5
Hypothetical two-way tables

Regularly Read *Times*	Regularly Read *Daily News*		
	Yes	No	
Aggregate table			
Yes	116	244	360
No	524	116	640
Total	640	360	1000
Latent class 1 (high education)			
Yes	.2311	.5689	.8000
No	.0578	.1422	.2000
Total	.2889	.7111	.3714
Latent class 2 (low education)			
Yes	.0812	.0188	.1000
No	.7308	.1692	.9000
Total	.8120	.1880	.6286

TABLE 6
Posterior from uniform prior

Regularly Read *Times*	Regularly Read *Daily News*		
	Yes	No	
Smaller latent class (1)			
Yes	.0232	.0005	.0236
No	.9532	.0232	.9764
Total	.9764	.0236	.4830
Larger latent class (2)			
Yes	.2137	.4766	.6904
No	.0959	.2137	.3096
Total	.3096	.6904	.5170
Marginal probabilities			
Yes	.1218	.2465	.3683
No	.7308	.1692	.6317
Total	.6317	.3683	1.0

program of Smith, Skene, Shaw, Naylor and Dransfield (1984). The numbers in the lower right corners of the subtables are the means of p_1 and $1 - p_1$. The marginal table is not identical with the original table, but we do not expect it to be due to substantial uncertainty and asymmetry in the posterior distribution. I also used a different prior distribution that had high prior means in the cells with low probabilities in Table 6 to see how sensitive the fit was to the prior. The posterior means of the cell probabilities were very close to those in Table 6. The important thing to keep in mind when estimating latent class parameters is that, unless one has an a priori reason to believe there are such classes and what they are, one will be hard pressed to offer any explanation for what the estimates are estimates of.

If one has prior beliefs about what the latent classes are, we saw how a Bayesian analysis can help to deal with the nonidentifiability in small tables. Identifiability is not a problem in larger tables in which the number of cells is much larger than the number of parameters fit by a latent class model. Also, "rotation" of latent classes is not an option as was rotation of factors in factor analysis. However, there is still more to the analogy between latent class models and factor analysis. The analogy extends to the two modes in which they can operate. *Exploratory latent class modeling* is searching for latent classes and hoping you can interpret them. There is also a mode, which I would call *confirmatory latent class modeling*. Just as in confirmatory factor analysis, one can incorporate prior assumptions about the latent classes and then fit models to see how well or badly they fit. But, when operating in the confirmatory mode, one should also not fall into the trap of rejecting a model without looking at how far it is from fitting well (as was done in the example in Section 9.2). In their Section 13.3, Dillon and Goldstein give a good example of performing a confirmatory latent class analysis and avoiding the trap.

13. CONCLUSION

The theory and practice of multivariate analysis has come a long way since 1958, and a great many talented people have contributed to the progress. The books by Anderson and Dillon and Goldstein give an excellent overview of that progress. Each one does a good job of what it sets out to do. Were one to teach a purely theoretical course in multivariate statistics to graduate students, one could do much worse than follow Anderson's text. One could do slightly better by augmenting it with a supplementary text offering a different perspective, such as that of Press (1982). Theory alone, however, can be quite sterile, and Anderson's text does not give the reader much exposure to the broad range of multivariate methods that are available for data analysis. On the other hand, teaching statistical methods without theory gives statistics a "black box" image. I do not believe that a student would gain much understanding of multivariate analysis if his/her sole exposure to the subject were a course from Dillon and Goldstein's text. Some of what each of these texts lacks is supplied in abundance by the other. But, as I have probably made clear elsewhere in this review, I believe that the two of them together still paint a somewhat incomplete picture of what multivariate theory and methods should be.

Despite the progress in multivariate statistics since 1958, there is still a long way to go. There is an old saying that one must crawl before one can walk. In multivariate analysis, I am afraid that we may have made the crawl an artform of its own. That is not to say that no good work is being done, but only that many of the results obtained are second and third order improvements on existing results. The field has great unexplored potential. In its current state, multivariate inference consists largely of invariant tests and the corresponding confidence regions, although very little effort has gone into predictive inference. Techniques such as latent class modeling and factor analysis have taken the journey one turn further away from prediction. They have taken the focus off of parameters, which, although unobservable, can usually be identified within the stochastic model, and placed the focus on unobservable *and* unidentifiable latent constructs.

The job for which statistical inference is best suited is to make statements about things which will or could (at least in theory) be observed, that is, to make predictive inference. This job has been neglected by multivariate researchers for too long. Fortunately, the methods and models that people develop to solve other interesting problems (such as LISREL, discriminant analysis, MANOVA, etc.) are still of use in predictive inference when appropriately interpreted and modified. For example, in a one-sample problem with p variate normal data $X_i \sim N_p(\mu, \Sigma)$, the marginal posterior distribution of the the mean vector μ is multivariate t with mean vector μ_1 and covariance matrix $[\lambda_1(a_1 - 1 - p)]^{-1}A_1$, where μ_1, λ_1, A_1 and a_1 have the same meanings as they did in (1). It is easy to show that the predictive distribution of the average of m future X_i observations is also multivariate t with mean vector μ_1, but with covariance matrix $(m^{-1} + \lambda_1^{-1})(a_1 - 1 - p)^{-1}A_1$. But predictive inference is not solely a rewording of parametric inference. It also needs to have its own unique set of tools developed, and I hope that a generation of researchers can begin to develop them with the same fervor with which they have developed what we now call multivariate analysis.

ACKNOWLEDGMENTS

I gratefully acknowledge the support of the Bayesian Statistics Study Year 1986, University of Warwick, Coventry, United Kingdom, during which much of this review was conducted. Although not apparent from the views expressed herein, I am also grateful to Clark Glymour for discussions on causal models. Finally, I would like to thank the Executive Editor and Robert Wijsman for several insightful comments on an earlier draft.

REFERENCES

ANDERSON, E. (1957). A semigraphical method for the analysis of complex problems. *Proc. Nat. Acad. Sci. U.S.A.* **43** 923–927.

ANDERSON, T. W., DAS GUPTA, S. and STYAN, G. P. H. (1972). *A Bibliography of Multivariate Statistical Analysis.* Oliver and Boyd, Edinburgh.

ANDREWS, D. F. (1972). Plots of high-dimensional data. *Biometrics* **28** 125–136.

BAGOZZI, R. P. (1980). Performance and satisfaction in an industrial sales force: an examination of their antecedents and simultaneity. *J. Marketing* **44** 65–77.

BARNARD, M. M. (1935). The secular variations of skull characters in four series of Egyptian skulls. *Ann. Eugenics* **6** 352–371.

BARTLETT, M. S. (1947). Multivariate analysis. *J. Roy. Statist. Soc. Suppl.* **9** 176–197.

BENTLER, P. M. (1985). *Theory and Implementation of EQS, a Structural Equations Program.* BMDP Statistical Software, Los Angeles.

BISHOP, Y. M. M., FIENBERG, S. E. and HOLLAND, P. W. (1975). *Discrete Multivariate Analysis.* MIT Press, Cambridge, Mass.

CHERNOFF, H. (1973). Using faces to represent points in k-dimensional space graphically. *J. Amer. Statist. Assoc.* **68** 361–368.

DARROCH, J. N., LAURITZEN, S. L. and SPEED, T. P. (1980). Markov fields and log-linear interaction models for contingency tables. *Ann. Statist.* **8** 522–539.

DIXON, W. J., ed. (1985). *BMDP Statistical Software.* Univ. California Press, Los Angeles.

ENGLEMAN, L. and HARTIGAN, J. A. (1969). Percentage points of a test for clusters. *J. Amer. Statist. Assoc.* **64** 1647–1648.

GEISSER, S. (1967). Estimation associated with linear discriminants. *Ann. Math. Statist.* **38** 807–817.

GEISSER, S. (1977). Discrimination, allocatory and separatory, linear aspects. In *Classification and Clustering* (J. Van Ryzin, ed.) 301–330. Academic, New York.

GLYMOUR, C., SCHEINES, R., SPIRTES, P. and KELLY, K. (1987). *Discovering Causal Structure: Artificial Intelligence, Philosophy of Science, and Statistical Modelling.* Academic, New York.

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

GUSEMAN, L. F., PETERS, B. C. and WALKER, H. F. (1975). On minimizing the probability of misclassification for linear feature selection. *Ann. Statist.* **3** 661–668.

HOLLAND, P. W. (1986). Statistics and causal inference (with discussion). *J. Amer. Statist. Assoc.* **81** 945–970.

HOWARD, R. A. and MATHESON, J. E. (1981). Influence diagrams. Technical Report, Dept. Engineering-Economic Systems, Stanford Univ.

JAMES, A. T. and STEIN, C. (1961). Estimation with quadratic loss. *Proc. Fourth Berkeley Symp. Math. Statist. Probab.* **1** 361–379. Univ. California Press.

JÖRESKOG, K. G. and SÖRBOM, D. (1978). *LISREL IV: Analysis of Linear Structural Relationships by the Method of Maximum Likelihood.* International Educational Services, Chicago.

KLEINER, B. and HARTIGAN, J. A. (1981). Representing points in many dimensions by trees and castles. *J. Amer. Statist. Assoc.* **76** 260–269.

KRUSKAL, J. B. (1964a). Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika* **29** 1–27.

KRUSKAL, J. B. (1964b). Nonmetric multidimensional scaling: A numerical method. *Psychometrika* **29** 115–129.

LAURITZEN, S. L. and SPIEGELHALTER, D. J. (1987). Fast manipulation of probabilities with local representations, with applications to expert systems. Technical Report, Institute of Electronic Systems, Aalborg Univ.

LEHMANN, E. L. (1958). Significance level and power. *Ann. Math. Statist.* **29** 1167–1176.

LEHMANN, E. L. (1959). *Testing Statistical Hypotheses.* Wiley, New York.

O'HAGAN, A. (1984). Motivating principal components, and a stronger optimality result. *Statistician* **33** 313–315.

PRATT, J. W. and SCHLAIFER, R. (1984). On the nature and discovery of structure (with discussion). *J. Amer. Statist. Assoc.* **79** 9–33.

PRESS, S. J. (1982). *Applied Multivariate Analysis.* Robert E. Krieger, Malabar, Fla.

RUBIN, D. B. (1978). Bayesian inference for causal effects: The role of randomization. *Ann. Statist.* **6** 34–58.

SCHERVISH, M. J. (1981a). Asymptotic expansions for means and variances of error rates. *Biometrika* **68** 295–299.

SCHERVISH, M. J. (1981b). Asymptotic expansions for correct classification rates in discriminant analysis. *Ann. Statist.* **9** 1002–1009.

SCHERVISH, M. J. (1983). User-oriented inference. *J. Amer. Statist. Assoc.* **78** 611–615.

SCHERVISH, M. J. (1984). Linear discrimination for three known normal populations. *J. Statist. Plann. Inference* **10** 167–175.

SCHERVISH, M. J. (1986). A predictive derivation of principal components. Technical Report 378, Dept. Statistics, Carnegie Mellon Univ.

SEN, P. K. (1986). Contemporary textbooks on multivariate statistical analysis: A panoramic appraisal and critique. *J. Amer. Statist. Assoc.* **81** 560–564.

SHACHTER, R. D. (1986). Evaluating influence diagrams. Technical Report, Dept. Engineering-Economic Systems, Stanford Univ.

SHEPARD, R. N. (1962a). The analysis of proximities: Multidimensional scaling with an unknown distance function. I. *Psychometrika* **27** 125–140.

SHEPARD, R. N. (1962b). The analysis of proximities: Multidimensional scaling with an unknown distance function. II. *Psychometrika* **27** 219–246.

SMITH, A. F. M., SKENE, A. M., SHAW, J. E. H., NAYLOR, J. C. and DRANSFIELD, M. (1984). The implementation of the Bayesian paradigm. *Comm. Statist. A—Theory Methods* **14** 1079–1102.

SPIEGELHALTER, D. J. (1986). A statistical view of uncertainty in expert systems. In *Artificial Intelligence and Statistics* (W. Gale, ed.) 17–55. Addison-Wesley, Reading, Mass.

STEIN, C. (1956). Inadmissibility of the usual estimator for the mean of a multivariate normal distribution. *Proc. Third Berkeley Symp. Math. Statist. Probab.* **1** 197–206. Univ. California Press.

SUBRAMANIAM, K. and SUBRAMANIAM, K. (1973). *Multivariate Analysis: A Selected and Abstracted Bibliography, 1957–1972.* Dekker, New York.

TORGERSON, W. S. (1952). Multidimensional scaling: I. Theory and method. *Psychometrika* **17** 401–419.

TUFTE, E. (1983). *The Visual Display of Quantitative Information.* Graphics Press, Cheshire, Conn.

TUKEY, P. A. and TUKEY, J. W. (1981a). Preparation; prechosen sequences of views. In *Interpreting Multivariate Data* (V. Barnett, ed.) 189–213. Wiley, New York.

TUKEY, P. A. and TUKEY, J. W. (1981b). Data-driven view selection; agglomeration and sharpening. In *Interpreting Multivariate Data* (V. Barnett, ed.) 215–243. Wiley, New York.

TUKEY, P. A. and TUKEY, J. W. (1981c). Summarization; smoothing; supplemented views. In *Interpreting Multivariate Data* (V. Barnett, ed.) 245–275. Wiley, New York.

WIJSMAN, R. A. (1984). Two books on multivariate analysis. *Ann. Statist.* **12** 1145–1150.

Comment

T. W. Anderson

1. OBJECTIVES

I am pleased that *Statistical Science* has furnished the opportunity to review my second edition and stimulate a discussion of the development and future of multivariate statistical analysis. The reviewer makes clear that his paper is "a thoroughly biased and narrow look"; I look forward to an unbiased, broad and comprehensive view in the future.

This article contrasts two books on multivariate statistical analysis that are very different in content and objectives. I shall hold my discussion to Schervish's remarks concerning my book. Let me first elucidate my criteria for inclusion of material. Writing a book on multivariate statistical analysis originated as an idea some forty years ago. It was accomplished over a period of years in connection with teaching courses in the Department of Mathematical Statistics at Columbia University. I wanted to write about statistical analysis that I thought has a sound foundation, about methods that were widely accepted. When the first edition was published in 1958, I had no thought that a quarter of a century would pass before the second edition would appear. When I finally came to revise the book, I found that most of the contents had stood the test of time; there was little that I wanted to change or delete, although there was a good deal that could be added. It has been a great satisfaction to me that the book has stood up so well; the initial selection of material has been justified. The objectives and organization of the first edition have been retained. In fact, the headings of the chapters and of most of the sections have been kept.

Although the book includes a considerable amount of mathematics, the primary objective is to provide and explain the methods and their properties. I think that the purpose of statistical theory is to initiate, develop, clarify and evaluate statistical methods. One criterion for inclusion of a topic is that it contributes to understanding useful procedures. Accordingly, there is not much theory in the book for its own sake, but I will admit that the relevance of some material is a matter of personal taste and some theory is to satisfy intellectual curiosity.

A second criterion, as the reviewer has surmised, is that a topic has a mathematical backing. For a confident and thorough understanding, the mathematical theory is necessary. This implies a rigorous treatment.

Thirdly, I wanted to organize the contents coherently. This desire is partly for the sake of clarity and efficiency of exposition and partly for personal satisfaction—aesthetics, if you will.

An outcome of following these criteria was that the inference treated here is based on normal distributions as models. There was not a place for ad hoc methods, valuable though they may be. Normal distributions serve as suitable models for generating many sets of data, but, of course, not for all sets.

Because the book is aimed at statistical practice, I included a number of examples, perhaps not enough. Beside the twelve examples mentioned by

T. W. Anderson is Professor of Statistics and Economics, Stanford University, Stanford, California 94305.