On Some Invariant Criteria for Grouping Data

Author(s): H. P. Friedman and J. Rubin

Source: *Journal of the American Statistical Association*, Dec., 1967, Vol. 62, No. 320 (Dec., 1967), pp. 1159-1178

Published by: Taylor & Francis, Ltd. on behalf of the American Statistical Association

Stable URL: https://www.jstor.org/stable/2283767

JSTOR is a not-for-profit service that helps scholars, researchers, and students discover, use, and build upon a wide range of content in a trusted digital archive. We use information technology and tools to increase productivity and facilitate new forms of scholarship. For more information about JSTOR, please contact support@jstor.org.

Your use of the JSTOR archive indicates your acceptance of the Terms & Conditions of Use, available at https://about.jstor.org/terms

Taylor & Francis, Ltd. and *American Statistical Association* are collaborating with JSTOR to digitize, preserve and extend access to *Journal of the American Statistical Association*

This content downloaded from
86.59.13.237 on Wed, 27 Jan 2021 17:36:17 UTC
All use subject to https://about.jstor.org/terms

ON SOME INVARIANT CRITERIA FOR GROUPING DATA

H. P. FRIEDMAN AND J. RUBIN
International Business Machines Corporation

This paper deals with methods of "cluster analysis". In particular we attack the problem of exploring the structure of multivariate data in search of "clusters".

The approach taken is to use a computer procedure to obtain the "best" partition of n objects into g groups. A number of mathematical criteria for "best" are discussed and related to statistical theory. A procedure for optimizing the criteria is outlined. Some of the criteria are compared with respect to their behavior on actual data. Results of data analysis are presented and discussed.

1. INTRODUCTION

"THE problems of discrimination and classification are insistent in sciences." This is the opening sentence of a paper by R. C. Bose and S. N. Roy (1938). In this paper they showed the invariance under non-singular linear transformation of the studentized Mahalanobis D^2 statistic and found its distribution. Their theoretical work was motivated by the following classification problem.

Suppose that we have three samples S_1, S_2, S_3, from populations π_1, π_2, π_3, respectively and that by suitable tests of significance[1] we have decided that the populations are different. Can we say in some significant sense whether population π_1 is closer to π_2 or to π_3? C. R. Rao (1952) applied the studentized D^2 to such problems arising in anthropometric studies. More recently T. Cacoullos (1965a) (1965b) considered this problem and some of its generalizations in the framework of statistical decision theory.

The Mahalanobis distance is very closely related to Hotelling's T^2 statistic and to R. A. Fisher's Linear discriminant Functions. These relationships have been described by Hotelling (1954). The above references as well as those in the sequel are not exhaustive. However, they should be indicative of the major work in the area of classification and discrimination.

It is clear that before applying any of the above mentioned methods the investigator must have a model that includes a suitable characterization of the populations that he is interested in. There are many problem areas in the Behaviorial and Life Sciences (See G. Ball (1965), Sneath and Sokal (1963), H. Solomon (1955)) where the investigator has gathered N objects to study. He doesn't have an explicit model. Since he must start somewhere, (the philosophical problem of an ultimate beginning to the process of classification is well described by I. J. Good (1965)), the investigator decides to characterize each object by a set of measurements. He has reason to believe that there should be meaningful subcategories of the N objects. However, he does not have an explicit *external criterion* with which to define these categories. Rather he is willing to tentatively accept an *internal criterion*; that is, he wants the data itself to suggest "natural" categories.

[1] W. Fisher (1953) discusses the problem of "significant difference" from a decision theoretic point of view.

This content downloaded from
86.59.13.237 on Wed, 27 Jan 2021 17:36:17 UTC
All use subject to https://about.jstor.org/terms

The methods we describe are most useful in what may be called the area of pre-classification. These methods are not strictly formal and at present involve no probability distribution theory.

The objective is to analyze multivariate heterogeneous data and to present the results in such a way as to lend insight into the structure of the data so as to suggest more formal models for further analysis as well as to provide guide lines for the collection of other data.

The methods to be described apply to data consisting of p measurements on each of n objects where there is some reason to believe that these n objects are a heterogeneous collection (i.e., the data may consist of clusters of points in p dimensional space). Further, the data should be such that the spatial distribution of the objects represented as points, can be meaningfully summarized by the location of the center of gravity of each cluster and by the sample scatter matrix of each cluster.

The output of the analysis will be partitions of the N objects into classes along with the empirical distribution of the measurements within each class. A method for appropriate graphical summary, portraying the relations of the groups to each other, will be given and should aid in the detection of outliers and hybrid objects.

Hopefully this type of analysis will be a step forward in helping to define clinically relevant subcategories of poorly defined illnesses such as schizophrenia, in isolating different disease syndromes or in defining useful categories in such fields as biological taxonomy.

In the sequel along with a description of methodology we present the results of application of these methods to actual data.

2. DESCRIPTION OF METHODS

The approach we have taken is to answer the question—What is the "best partition" of n objects (represented as points in a p—dimensional space) into g groups. We leave informal the process for deciding the "best number" of groups, or whether an object is an "outlier" or a "hybrid".

As part of our method we define "best partition" by introducing a numerical valued function defined for all partitions of the objects into g groups, and selecting a partition for which the numerical measure is maximal. In the sequel we describe three such functions and compare them on data.

2.1 *Criteria for grouping*

To begin we assume our data are given in the form of a matrix X (nxp) with the ith row given by the $(1 \times p)$ vector $P_i = (x_{i1}, \cdots, x_{ip})$ representing the observation vector of the ith object. Thus a decomposition of the n objects into groups will be given by a partition of the row vectors of X. Without loss in generality we may assume that the center of gravity of the total n points is the zero vector. Thus the total scatter matrix of the n points (in the sense of S. Wilks (1960) is given by

$$T = X^T X = \sum_{i=1}^{n} P_i^T P_i$$

This content downloaded from
86.59.13.237 on Wed, 27 Jan 2021 17:36:17 UTC
All use subject to https://about.jstor.org/terms

Now suppose that we have a partition of the n objects into g groups with n_1, $n_2, \cdots, n_g$ objects in each group and $n = \sum_{i=1}^{g} n_i$. Then for the kth group the row vectors P_{lk} for $l = 1, \cdots, n_k$ represent the objects in group G_k. We now define the scatter matrix for each group G_k with center of gravity vector C_k by

$$W_k = \sum_{l=1}^{n_k} (P_{lk} - C_k)^T (P_{lk} - C_k).$$

The pooled-within groups scatter matrix is defined by

$$W = \sum_{k=1}^{g} W_k.$$

The between groups scatter matrix is defined by

$$B = \sum_{k=1}^{g} n_k C_k^T C_k.$$

Hence for each partition of the n objects into g groups we have the following well known matrix identity (see S. Wilks (1962)).

$$T = W + B. \tag{1}$$

For $p = 1$ (one variable) equation (1) is a statement about scalars and since the total scatter T is fixed a natural criterion for grouping is to minimize W. This is equivalent to maximizing B. Also for $p = 1$ we have $T/W = 1 + B/W$ where B/W multiplied by the ratio of the degrees of freedom is familiar to statisticians as an F ratio. This ratio is invariant under non-singular linear transformations of the data. The criterion may thus be restated as partitioning the n objects into g groups so as to maximize the ratio B/W or equivalently T/W. W. Fisher (1958) for the case $p = 1$ describes a procedure to find a partition into g groups for given n objects for which W is minimized.

For $p > 1$ equation (1) is a matrix equation and the question of criteria for grouping is more complex. One criterion that has been suggested by A. Edwards and L. Cavalli-Sforza (1965) and by R. Singleton (1965) is to minimize trace W over all partitions into g groups. Since T is constant over all the partitions, minimizing trace W is equivalent to maximizing trace B, because Trace T = Trace W + Trace B. Edwards (1965) uses this criterion by first partitioning into two groups say G_1, G_2, and then he continues by partitioning G_1 and G_2 respectively into two groups and so on. Singleton (1965) partitions directly into g groups. Although Trace W is invariant under an orthogonal transformation it is not invariant under any non-singular linear transformation.

If we measure distance squared between objects i and j by (ordinary Euclidean distance squared)

$$d_{ij} = (P_i - P_j)(P_i - P_j)^T$$

then Trace W can be directly computed from the pairwise distances d_{ij}. Note also that these distances are invariant under orthogonal transformations.

Assuming the p variables are not linearly dependent, then as long as $p \leq n - g$, $W(p \times p)$ is positive definite symmetric and hence so is W^{-1}. Thus we may de-

This content downloaded from
86.59.13.237 on Wed, 27 Jan 2021 17:36:17 UTC
All use subject to https://about.jstor.org/terms

fine as a squared distance between points (Mahalanobis Euclidean Distance),

$$md_{ij} = (P_i - P_j)W^{-1}(P_i - P_j)^T$$

This distance md_{ij} is invariant under any non-singular linear transformation of the points P_i. Hence we see that for any partition we can define an invariant metric.

Here we begin to see the "bootstrap" nature of the problem. If we knew the groups then we could define an appropriate distance. If we knew the appropriate distance then we would be much closer to knowing the groups. The Trace criterion hides this circularity by assuming ordinary Euclidean distance.

We deal with the circularity by utilizing criteria invariant under non-singular linear transformations of the original data matrix. These criteria are derived from the basic relation $T = W + B$. One criterion is the ratio of determinants

$$\frac{|T|}{|W|} = |I + W^{-1}B| \tag{2}$$

The left hand side of (2) is a scalar function. Its reciprocal $|W|\,|T|$ was introduced by S. Wilks as a statistic in the situation where there are g given groups normally distributed with equal covariance matrices and we want to test whether at least two of the g groups differ in location (i.e., mean values). We use $|T|/|W|$ as a criterion function to be maximized. That is in principle we consider all partitions of the n objects into g groups and choose that partition into g groups for which this ratio is a maximum. Note that $|T|$ the total scatter is fixed and thus it is sufficient to minimize $|W|$. Further note that this measure is not comparable for different values of g since its value for $k+1$ groups will be greater than or equal to its value for k groups. However, we will show how the the values of log (max $|T|/|W|$) may be used as informal indicators for the number of groups.

Once having decided on the number of groups, the matrix W for the partition which maximizes $|T|/|W|$ determines the pairwise Mahalanobis distance between objects.

Another criterion function related to the basic identity (1) is the maximum of the Trace$[W^{-1}B]$ over all partitions into g groups. The function Trace $[W^{-1}B]$ has been used as a test statistic in the same way as the ratio of the two determinants mentioned previously. It has been called Hotelling's Trace Criterion and is also equivalent to what C. R. Rao (1952) called the Generalization to $K(>2)$ groups of the Mahalanobis distance between two groups.

Indeed, both Trace $W^{-1}B$ and $|T|/|W|$ may be expressed in terms of the eigenvalues of $W^{-1}B$.

In particular

$$\frac{|T|}{|W|} = \prod_{i=1}^{t} (1 + \lambda_i) \quad \text{and,}$$

$$\text{Trace } W^{-1}B = \sum_{i=1}^{t} \lambda_i.$$

This content downloaded from
86.59.13.237 on Wed, 27 Jan 2021 17:36:17 UTC
All use subject to https://about.jstor.org/terms

These eigenvalues are solutions of the determinant equation, $|B-\lambda W| = 0$. All the eigenvalues of this equation are known to be invariant under non-singular linear transformations of the original data matrix. In fact they are the only invariants of W and B under such transformations. There is a proof of this statement in T. Anderson (1958).

The distributions of $|W|/|T|$ and Trace $W^{-1}B$ are known under the Null-Hypothesis that the g groups are samples from the same multivariate normal population. The power functions of these test statistics (i.e., the non-central distributions) have not been studied over a wide class of alternative hypotheses. Thus statistical theory does not provide guide-lines for choice between these two criteria even under standard usage. Some recent work in this direction are the papers by R. Bargmann and H. Posten (1964) and M. Schatzoff (1966).

The foregoing refers to the distribution of $|T|/|W|$ over all partitions of n objects into g groups. This value in general would be larger than the one based on the original samples. Indeed the distribution of $\max|T|/|W|$ is certainly different from the distribution of $|T|/|W|$ under the Null hypothesis. It is the distribution of $\max|T|/|W|$ that one must study to begin to answer questions of statistical significance of the groups found by our methods. Similar remarks also hold for the other statistical criterion functions that we consider.

2.1.1 *Pairwise distance criteria*

Many clustering procedures take as a point of departure a matrix of pairwise distances or similarities between objects. Thus, for n objects one would have an $n \times n$ symmetric matrix of distances or similarities. For the problems we deal with, the investigator does not want to assign *a priori* a fixed distance between objects. However, if he uses the trace W criterion then he has implicitly chosen ordinary euclidean distance, since

$$\text{Trace } W_k = \frac{\sum_{\substack{l,m=1 \\ l<m}}^{n_k} (P_{lk} - P_{mk})(P_{lk} - P_{mk})^T}{n_k}, \quad \text{and}$$

$$\text{Trace } W = \sum_{k=1}^{g} \text{Trace } W_k$$

Ordinary euclidean distance has been advocated by Sneath and Sokal (1963) for use in biological taxonomy. One could extend our methods to any pairwise distance d_{ij} by defining a criterion suggested by the above remarks as

$$\text{minimum} \sum_{k=1}^{g} \frac{\sum_{\substack{i,j=1 \\ i<j}}^{n_k} d_{ij}}{n_k},$$

where the minimum is taken over all partitions into g groups. For example M. Kendall (1965) has given a distance function that depends only on rank orders. Hence, the above criterion would be invariant under a rank preserving transformation.

This content downloaded from
86.59.13.237 on Wed, 27 Jan 2021 17:36:17 UTC
All use subject to https://about.jstor.org/terms

Finally we feel compelled to mention the important work of R. Shepard (1962) and J. Kruskal (1964) in multi-dimensional scaling. By their methods it is possible, given only rank order of pairwise resemblances between objects to imbed them in a euclidean space such that the rank order of the pairwise distances in the euclidean space preserves (or minimally distorts) the given rank order of resemblances. One could then use metric methods or any of the criteria we have mentioned for grouping the objects in this space. A big problem here is one of interpreting the dimensions. In this context we recommend an article on scaling by W. Torgerson (1965).

A very recent paper by J. Gower (1966) deals with distance functions and, their use in classification and factor analysis.

3. COMPUTATIONAL CONSIDERATIONS

The computational problem of evaluating all partitions of n objects into g groups in order to select one that maximizes a given criterion function is solvable in principle but not in practice since the number of partitions is enormously large (see G. Rota (1964). An attempt at non-exhaustive procedures by a sampling technique has been given by Fortier and Solomon (1964). The procedure we use is also non-exhaustive and is due to Rubin (1967). The procedure is as follows.

We start with any given partition into g groups. Consider moving a single object into every group other than the one it is in. If no move will create a partition for which the criterion is increased, leave the object where it is. Otherwise move it so that the maximum increase in the criterion occurs. Using the partition thus created, we process the second object in the same way, then the third, etc. A "hill-climbing pass" is defined as the application of this procedure once to each object, in some given order. After several hill-climbing passes, one must reach a point at which no move of a single object from the group it is in to a different group will cause an increase in the criterion function. At this point we say we have found a "single move local maximum" of our criterion function. This rarely takes more than half a dozen passes. To try to rise above this we apply some heuristic procedures, in particular:

1. "Forcing passes"—We start with the best partition yet known. Processing one group at a time, in sequence, we place each object of the group being processed into the *outside* group with nearest center of gravity (using the metric defined by the current partition or, for the Trace W criterion, the ordinary Euclidean metric), recalculating the criterion function after each move. This is done in sorted order, the object nearest an outside group being moved first. Although the criterion initially decreases, it may at some point during the process achieve a value higher than previously found. This will especially be the case if the group being processed consists of two clusters widely separated in space. After processing all the objects of one group, we restore the best partition yet found, and proceed to process the next group. A "forcing pass" is defined as the application of this procedure once to each group, in sequence. Forcing passes are repeated until they produce no improvement. These passes are relatively fast, compared to hill-climbing, since we need not evaluate every possible move for an object.

This content downloaded from
86.59.13.237 on Wed, 27 Jan 2021 17:36:17 UTC
All use subject to https://about.jstor.org/terms

2. "Reassignment passes"—A procedure due to Edward Forgy (1965) involves starting with a partition Q (we use the best partition currently known) and reassigning each object to the group with nearest center of gravity. The value of the newly formed partition is then calculated. When using Trace W, as Forgy did, the distance from Point P to center of gravity C_k is the ordinary Euclidean distance. With either of the other two criteria, we use the metric defined by the matrix W^{-1} computed from the partition P—i.e., $d(P, C_k) = (P-C_k)W^{-1}(P-C_k)^T$. The centers of gravity C_k and the scatter matrix W are maintained as those of the original partition Q until all n objects have been reassigned, at which time new values for C_k and W are computed. This contrasts with the hill-climbing, for which the partition and the derived W change with each move of an object.

The reassignment of each object in the above manner is termed a "reassignment pass." Reassignment passes are repeated until a partition with higher value is no longer achieved. Sets of forcing passes and reassignment passes are alternated until neither produces improvement, and the hill-climbing is reentered to get to a new local maximum. Certain other heuristics are also applied (see Rubin (1967)) but when it proves impossible to reach a higher local maximum, the procedure is terminated. However, if one is willing to spend the computer time (which is substantial) one can repeat the entire procedure using random starting partitions. (Initially, the procedure can start by applying a forcing pass to the conjoint partition—i.e., the partition with all objects in one group—or else with a random partition.)

When applying this procedure to the Bee data described later in this paper, we were able to obtain the highest known value 10 times out of 14 runs from different random starting partitions. With some less well-structured data not described in this paper, the highest value was reached 3 out of 11 runs from different random starting partitions—and we are not at all sure that there are no partitions of higher value.

The forcing and reassignment passes are fast, but only occasionally helpful. Restarting from random partitions is of course slow, but provides more confidence in the result.

One very nice advantage of this algorithm from the point of view of computation is that usually only one object is moved at a time. Hence, it has been possible to design the computations so that W does not have to be completely recalculated each time. We only compute the change in W due to the moving of a single object.

3.1 *Too many dimensions and singular matrices*

As we have commented on in previous sections the use of the ratio of determinants criterion and the Hotelling trace criterion requires the non-singularity of W, the pooled within groups scatter matrix. This requires that the number of variables P be less than or equal to $n-g$. In many problems the data does not satisfy this condition. What we do under these conditions is to perform a principal component analysis on T, the total scatter matrix or sometimes the total correlation matrix.

If we partition into g groups then we can keep at most $n-g$ components. We

This content downloaded from
86.59.13.237 on Wed, 27 Jan 2021 17:36:17 UTC
All use subject to https://about.jstor.org/terms

usually keep less than this number making an informal judgement based on the way the eigenvalues decrease in size. If we group the data based on a choice of k components it is good practice to check for consistency by repeating the analysis with more than k components. It is also desirable once having decided on a set of groups, to go back to the space of original variables and do some form of stepwise discrimination analysis (as mentioned later) so that in the end result the groups are described in terms of a subset of the original variables. This procedure may be used for very large p even in the non-singular case to cut down the cost of computation.

The foregoing remarks once again reveal the "bootstrap" nature of the clustering problem. Until we know the groups it is difficult to select a subset of relevant variables.

The criterion of minimum Trace W is not affected by problems of singularity. The invariance of the trace of W under an orthogonal transformation of the original p dimensional space is also unaffected. From a computational standpoint it is the easiest to compute of the three criteria considered in this report. This criterion leads to what Macqueen (1966) calls the minimum variance partition.

4. RELATION TO LINEAR DISCRIMINANT ANALYSIS (p DIMENSIONS, g GROUPS)

The procedure of linear discriminant analysis as described by S. Wilks (1962) utilizes the matrices W and B to determine a new set of coordinate axes in which to describe the observation vectors.

The observations originally described as vectors in a p dimensional coordinate system are described as vectors in a t dimensional coordinate system with $t=\min\ (p,\ g-1)$.

The total scatter, pooled within scatter and between groups scatter matrices may be defined for the points described in this space. We denote them T_t, W_t, and B_t. For $p>g-1$ the procedure of Linear Discriminant Analysis results in choosing that subspace of dimension $g-1$ for which $|T_t|/|W_t|$ is maximal and equal to $|T|/|W|$. That is, the maximal ratio of scatters is equal to the ratio of scatters of the points as described in the original coordinate system. (Hence for example if we have 25 variables and 4 groups then we may describe our points in a 3 dimensional coordinate system with no loss in within-to-between group scatter.)

It is further known that the coordinate directions in the $t=\min[p,\ g-1]$ dimensional subspace are the eigenvectors associated with the eigenvalues λ_1, $\lambda_2, \cdots, \lambda_t$ of the matrix $W^{-1}B$. Furthermore the eigenvalue λ_i is the ratio of the one dimensional between group scatter to the one dimensional within group scatter as measured along the direction of eigenvector V_i associated with λ_1.

In discriminant function analysis a partition of n objects in p dimensions into g groups is given. Thus $|T|/|W|$ in p dimensions is known. The object is to find a space of dimension $t<p$ for which $|T_t|/|W_t| = |T|/|W|$. In our procedure the main objective is to *find* the partition for which $|T|/|W|$ or some other criterion function is maximal in the original p dimensions. We may then, as we describe in section (5), apply linear discriminant analysis to the results of our procedure to provide a descriptive summary of the results. We just add a

This content downloaded from
86.59.13.237 on Wed, 27 Jan 2021 17:36:17 UTC
All use subject to https://about.jstor.org/terms

note of caution that the usual tests of significance used in discriminant analysis do not apply because of our use of max $|T|/|W|$. Even if it is known that one had samples from g groups and computed $|T|/|W|$ from these samples the value would be less than or equal to the value of max $|T|/|W|$ that one could get by reshuffling the sample.

In the singular case $p>n$, where we have taken principal components to reduce the dimensionality, the discriminant function weights are on the principal components rather than on the original variables. One may obtain the weights on the original variables by forming the product of the matrix of principal components with each eigenvector of $W^{-1}B$. A procedure we recommend is, once having found the groups, using the principal components to go back to the space of original variables and do some form of step-wise discriminant analysis in order to eliminate variables. A relevant paper is given by Weiner and Dunn (1966).

5. GRAPHIC REPRESENTATION OF GROUPS

Due to the nature of the criteria that we used to select the groups the natural space to represent the groups in is the vector space generated by the t linearly independent eigenvectors of $W^{-1}B$. This space has nice properties. From the remarks made in section [4] it follows that we lose nothing with respect to $|T|/|W|$ if $t<p$ (p is the dimension of the original measurement space). Furthermore if $t=p$ ordinary euclidean distance in the space of eigenvectors is equivalent to Mahalanobis (Euclidean) distance in the original measurement space. A discussion of related ideas may be found in M. Wilk and Gnanadesikan (1965).

We have $W(p\times p)$ positive definite. B is $(p\times p)$ and positive semi-definite. Both W and B are symmetric. Hence, there exists non-singular R $(p\times p)$ such that

$$R^TWR = I = (p \times p) \text{ identity} \tag{1}$$

$$R^TBR = D = (p \times p) \text{ diagonal} \tag{2}$$

It is known that $D=(d_{ik})=(\delta_{ik}\lambda_k)$ (where λ_k is the kth eigenvalue of $W^{-1}B$) and that the columns of R are the eigenvectors of $W^{-1}B$.

Without loss in generality if $t=p$ we may assume

$$D = \begin{pmatrix} \lambda_1 & & & \\ & \lambda_2 & & \\ & & \ddots & \\ & & & \lambda_p \end{pmatrix} \quad \text{with} \quad \lambda_1 > \lambda_2 > \lambda_p$$

If we denote the jth column of R by v_j^T then v_j^T is the eigenvector of $W^{-1}B$ associated with λ_j. Let P be any $(1\times p)$ observation vector in the original measurement space. Then

$$Pv_j^T \quad \text{for } j = 1, \cdots, p$$

yields the coordinates of P in the space of the eigenvector. Let the space of vectors Q be defined (for all vectors P) by

This content downloaded from
86.59.13.237 on Wed, 27 Jan 2021 17:36:17 UTC
All use subject to https://about.jstor.org/terms

$$Q = PR \tag{5}$$

Hence, for $t=p$ it follows from (1) and (3) that

$$PW^{-1}P^T = PRR^TP^T = QQ^T. \tag{4}$$

Thus, since an observation vector P in the original space becomes $Q=PR$ in the space of discriminant functions (i.e., eigenvectors), (4) asserts that ordinary Euclidean distance in the space of discriminant functions is equivalent to Mahalanobis distance in the original space. If $t>p$ we cannot preserve actual distance but we can preserve $|T|/|W|$. In practice we can only plot in two dimensions. We plot the coordinates Pv_1^T and Pv_2^T for the two largest eigenvalues. We can also plot any pair (Pv_1^T, Pv_j^T). The information about scatter lost by going to two dimensions is readily determined since

$$\frac{|T|}{|W|} = \prod_{i=1}^{t} (1 + \lambda_i).$$

An alternate index of the relevance of the two dimensional representation has been suggested by C. R. Rao (1964), namely

$$\frac{\lambda_1 + \lambda_2}{\lambda_1 + \cdots + \lambda_t}$$

Finally we remark that if the groups were from p-variate normal distributions with equal covariances then if $t=p$ the transformation into discriminant function space may be viewed as first transforming the original coordinates so that equal probability ellipses appear as spheres and transforming again to best describe the dispersion of the centers of gravity of the spheres.

6. RESULTS

In this section we present and discuss the results of applying our methods to data. The main purpose of this presentation is to make more explicit the ideas presented in the previous sections as well as to compare the various criterion functions.

6.1 *Artificial data*

In order to initially test our procedures we generated in the computer a set of fifty objects consisting of five groups of ten objects each with each object characterized by a vector of five measurements. The groups were constructed so as to represent five samples from five different multivariate normal distributions with equal covariance structure but differing in location. The five groups were labeled 1–10, 11–20, 21–30, 31–40, and 41–50.

It is instructive to look at the plots of the data. For the determinant criterion, we see in figure 1 the result of partitioning into two groups. In figure 2 we see the result for a three group partition, figure 3 for four groups, figure 4 for five groups and figure 5 for six groups. We see structure at the two, three, and five group levels. All samples were recovered at the five group level except for object 19 which went with the 1–10 group. At the six group level object 13

This content downloaded from
86.59.13.237 on Wed, 27 Jan 2021 17:36:17 UTC
All use subject to https://about.jstor.org/terms

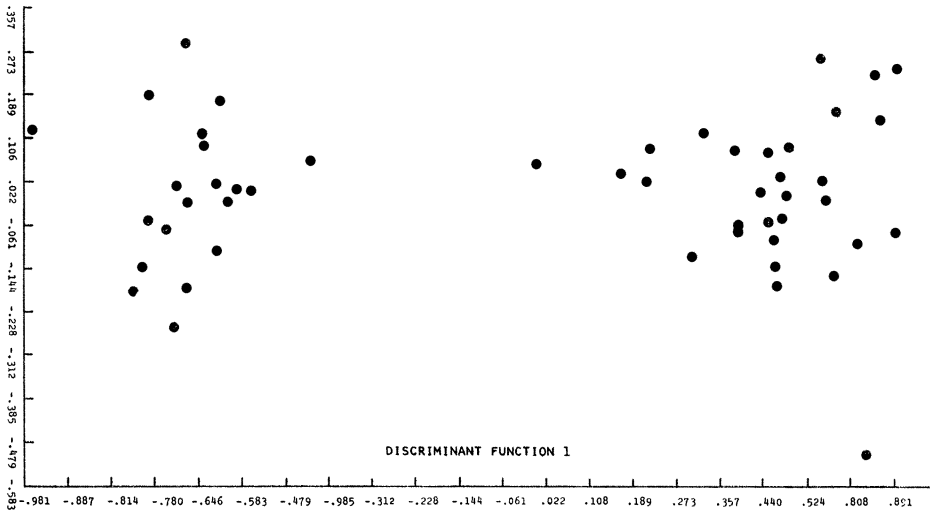

FIG. 1. 5 Variable random data, determinant criterion
50 Objects 2 Groups

formed a group by itself and object 19 went with its original sample. The trace $W^{-1}B$ criterion gave the same results as the determinant criterion for five groups.

We present some summary statistics for the artificial data below.

Artificial Data

Number of Groups	2	3	4	5	6
Log max $\lvert T\rvert/\lvert W\rvert$	1.22	1.91	2.28	2.83	3.12

Eigenvalues of $W^{-1}B$ for five groups

Criterion	λ_1	λ_2	λ_3	λ_4
max $\lvert T\rvert/\lvert W\rvert$	229.0	1.38	.34	.03
Trace $W^{-1}B$	229.0	1.38	.34	.03

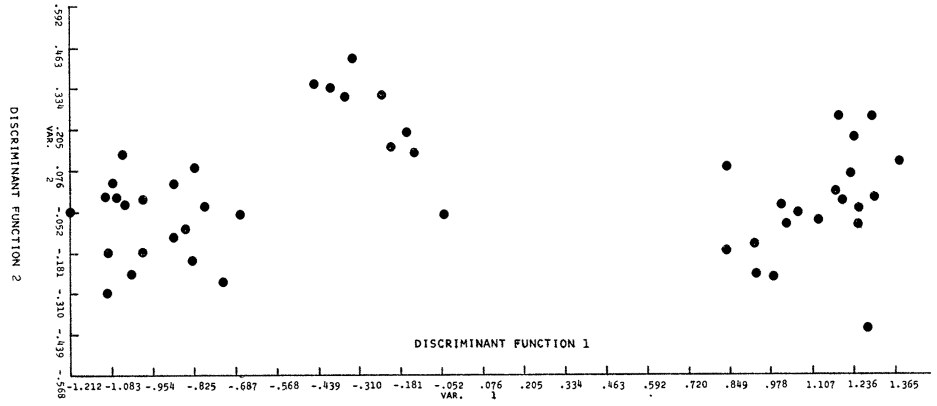

FIG. 2. 5 Variable random data, determinant criterion
50 Objects 3 Groups

This content downloaded from
86.59.13.237 on Wed, 27 Jan 2021 17:36:17 UTC
All use subject to https://about.jstor.org/terms

Eigenvalues of $W^{-1}B$ for six groups

Criterion	λ_1	λ_2	λ_3	λ_4	λ_5
max $\lvert T \rvert / \lvert W \rvert$	222.0	2.22	.71	.07	.004
Trace $W^{-1}B$	303.0	1.5	.34	.05	.000

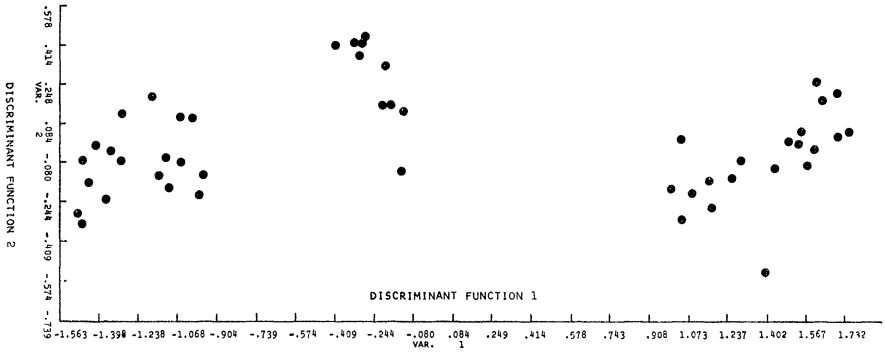

FIG. 3. 5 Variable random data, determinant criterion
50 Objects 4 Groups

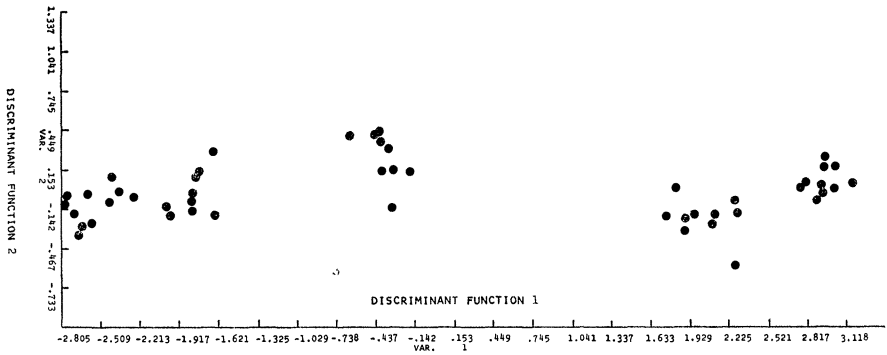

FIG. 4. 5 Variable random data, determinant criterion
50 Objects 5 Groups

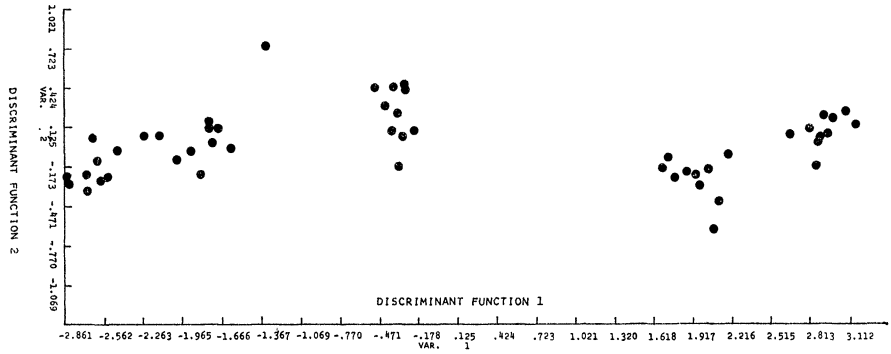

FIG. 5. 5 Variable random data, determinant criterion
50 Objects 6 Groups

This content downloaded from
86.59.13.237 on Wed, 27 Jan 2021 17:36:17 UTC
All use subject to https://about.jstor.org/terms

Two points are of interest. First, note how the values of log max $|T|/|W|$ reflect the group structure. In particular note the large rise in value from 2 to 3 compared to the small rise in value from 3 to 4. Note also the rise in value from 4 to 5 compared to the rise in value from 5 to 6. This reflects the structure at the three group and five group partitions as seen in the plots. Two sets of groups were close to each other with one separate from all the rest, thus, accounting for the structure when partitioning into three groups. Secondly, note the relative magnitudes of the eigenvalues. Essentially, all the group structure is in one dimension. The difference between the criteria is shown in the six group partition. The trace $W^{-1}B$ increased the largest eigenvalue, while the determinant criterion gave an increase to the smaller ones. This tendency will be highlighted with the Bee data we describe later.

6.2 *Iris data*

The Iris data we analyze was published by R. A. Fisher (1936). There are three species of Iris, Iris setosa, Iris versicolor, and Iris virginica. There are fifty plants of each species with four measurements on each plant. We have labeled them 1–50, 51–100, 101–150, where the labels are taken consecutively down the rows of the originally published table by R. A. Fisher. Two of the species, Iris setosa (1–50) and Iris versicolor (51–100) were found growing in the same colony by the botanist E. Anderson. Iris virginica was found in a different colony. R. A. Fisher (1936) applied his procedure of Linear Discriminant Analysis to the above three groups and found that Iris setosa could be separated from the other two species. However, Iris versicolor and Iris virginica overlapped somewhat. More recently M. G. Kendall (1965) applied a clustering technique based on a pairwise distance function utilizing only the rank order of the measurements and he was able to recover Iris setosa from the lumped sample of 150 objects. He experienced difficulty with about a dozen plants from the intersection of Iris versicolor and virginica.

We applied the min trace W criterion to this data. The breakup into two groups gave all Setosa in one group plus plants 58, 94, and 99 of Versicolor with the rest of the plants in the other group. When we partitioned into three groups we again found Iris Setosa appearing as a separate group with the two other groups being predominantly Versicolor and Virginica except for about ten plants where some Virginica went to Versicolor and vice versa. At four groups we found all Setosa in one group; a group consisting solely of Virginica; a group solely of Versicolor except for plant 107 of Virginica; and a group containing a mixture of Versicolor and Virginica.

We next applied the max $|T|/|W|$ and trace $W^{-1}B$ to the Iris data. The partition into two groups is shown in figure 6. Again Setosa appears well separated from the other species. The partition into three groups recovered the three species except for three plants for both criteria (see figures 7, 8). For trace $W^{-1}B$ plants 71, 78, and 84 went with Virginica, while for the determinant criterion 71 and 84 went with Virginica, and 134 went with Versicolor.

We present some summary statistics for the determinant and trace $W^{-1}B$ criteria below.

This content downloaded from
86.59.13.237 on Wed, 27 Jan 2021 17:36:17 UTC
All use subject to https://about.jstor.org/terms

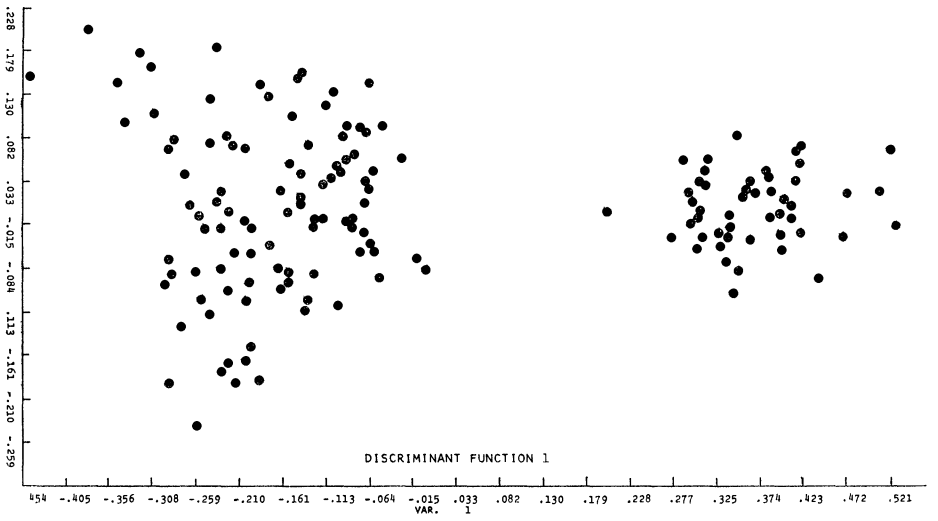

FIG. 6. Iris data determinant criterion
2 Groups

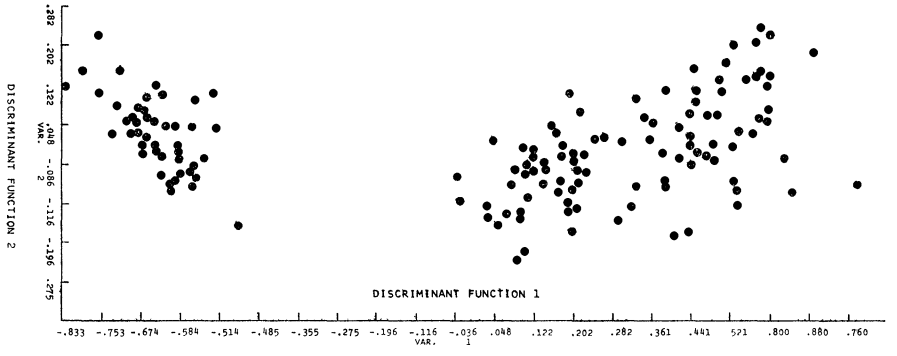

FIG. 7. Iris data determinant criterion
3 Groups

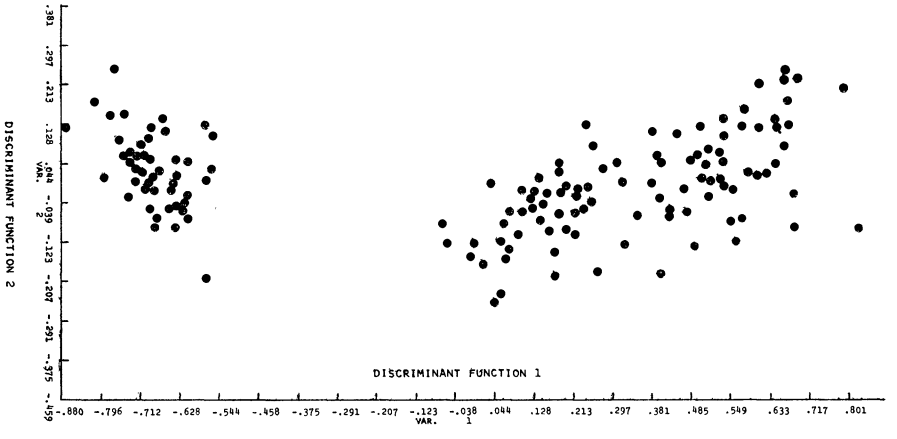

FIG. 8. Iris data determinant criterion
4 Groups

This content downloaded from
86.59.13.237 on Wed, 27 Jan 2021 17:36:17 UTC
All use subject to https://about.jstor.org/terms

Iris Data

Number of Groups	2	3	4
Log max $\|T\|/\|W\|$	1.04	1.66	1.87
Max trace $W^{-1}B$	9.88	34.20	55.90
Min trace W	152.50	79.00	57.30

Eigenvalues of $W^{-1}B$ for three groups

Criterion	λ_1	λ_2
Max $\|T\|/\|W\|$	33.63	.31
Max trace $W^{-1}B$	33.95	.27

Eigenvalues of $W^{-1}B$ for four groups

Criterion	λ_1	λ_2	λ_3
Max $\|T\|/\|W\|$	39.37	.46	.28
Max trace $W^{-1}B$	55.4	.42	.02

Note that all the structure is in one dimension. However, the direction given by the largest eigenvector of $W^{-1}B$ is different from the largest principal component direction. The eigenvector of $W^{-1}B$ indicates the proper direction of separation between groups and thus accounts for the better results with the determinant and trace $W^{-1}B$ criteria as compared with Trace W. The discriminant function we found is slightly different than the one found by R. A. Fisher (1936). This is because we put in the total set of 150 plants and let the procedure find the "best" separated three groups. These groups were slightly different (as described earlier) than the original three species. This is a nice illustration of what we described in section (4) as being able to reshuffle the samples to effect a better separation.

Note also that values of log max $|T|/|W|$ give an indication of structure at three groups. We would also have to say that the value of min trace W gives a similiar indication although this criterion did not do as well in recovering the original species.

In summarizing the behavior of the criterion functions on the Iris Data, one is led to reject the min trace W criterion since it did not separate Setosa completely at the two group level. Our interpretation of the reason for this is that as a global criterion it imposed a euclidean metric on the data such that three plants in Setosa were closer to the center of gravity of the lumped sample of Virginica and Versicolor than they were to the center of gravity of their own group.

One could not choose between the determinant and trace $W^{-1}B$ criteria since there was essentially only one eigenvector direction defining the separation of the groups.

6.3 *Bee data*

The data analyzed consists of 97 American species of the megachilid genera, Ashmeadiella, Anthocopa, Hoplitis, and Proteriades. These four genera are numbered 1–23, 24–41, 42–62, and 63–97 respectively. The data were obtained from Professors Michener and Sokal of the University of Kansas. Michener

This content downloaded from
86.59.13.237 on Wed, 27 Jan 2021 17:36:17 UTC
All use subject to https://about.jstor.org/terms

and Sokal (1957) reported a revised classification of these species based on a pairwise distance procedure. We are not reporting a complete analysis of this data here. The results included are to make certain points with respect to the various criteria under study.

The data consists of 122 measurements on 97 species coded by Professor Michener. We first performed a principal component analysis of the total correlation matrix. We kept the first 3 components. Thus the input data matrix $(n \times p)$ to the classification computer program was (97×3).

Shown in figure A is a plot of the first two principal components of the Bee Data. Enclosed are the four genera as originally defined by Michener and Sokal.

We analyzed this data with the three different criteria, min Trace W, max $|T|/|W|$ and max Trace $W^{-1}B$. Some of the graphic output is shown in figures 9–11.

In figure 10 we see the result of using three components. All four genera are recovered. Contrast figure 10 with figure A. Note the separation of the genera in figure 10. This is due to the fact that the projection of figure 10 was determined by the eigenvalues of $W^{-1}B$ rather than of T.

The min Trace W Criterion recovered the four genera except for species 41.

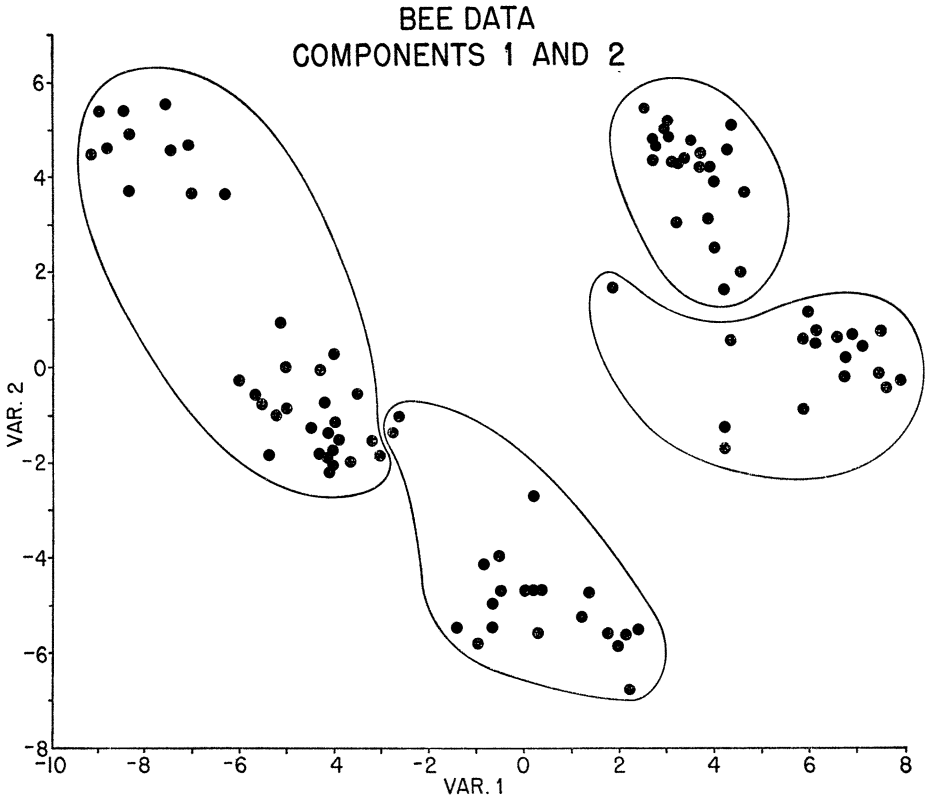

FIG. A

This content downloaded from
86.59.13.237 on Wed, 27 Jan 2021 17:36:17 UTC
All use subject to https://about.jstor.org/terms

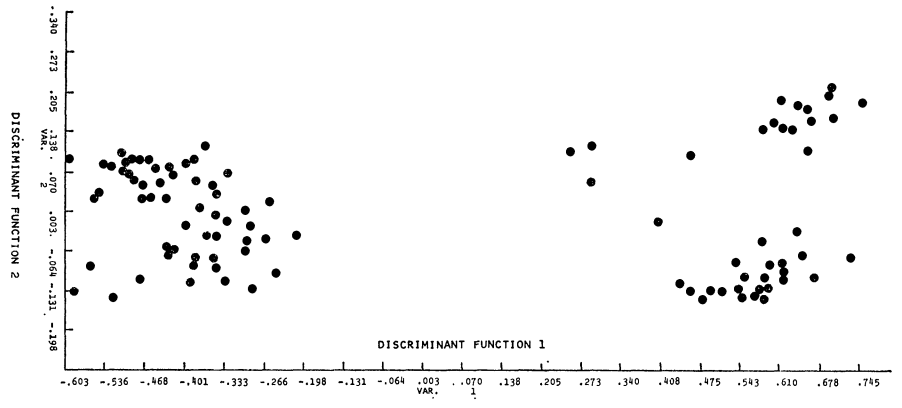

FIG. 9. Bee data determinant criterion
2 Groups 3 Components

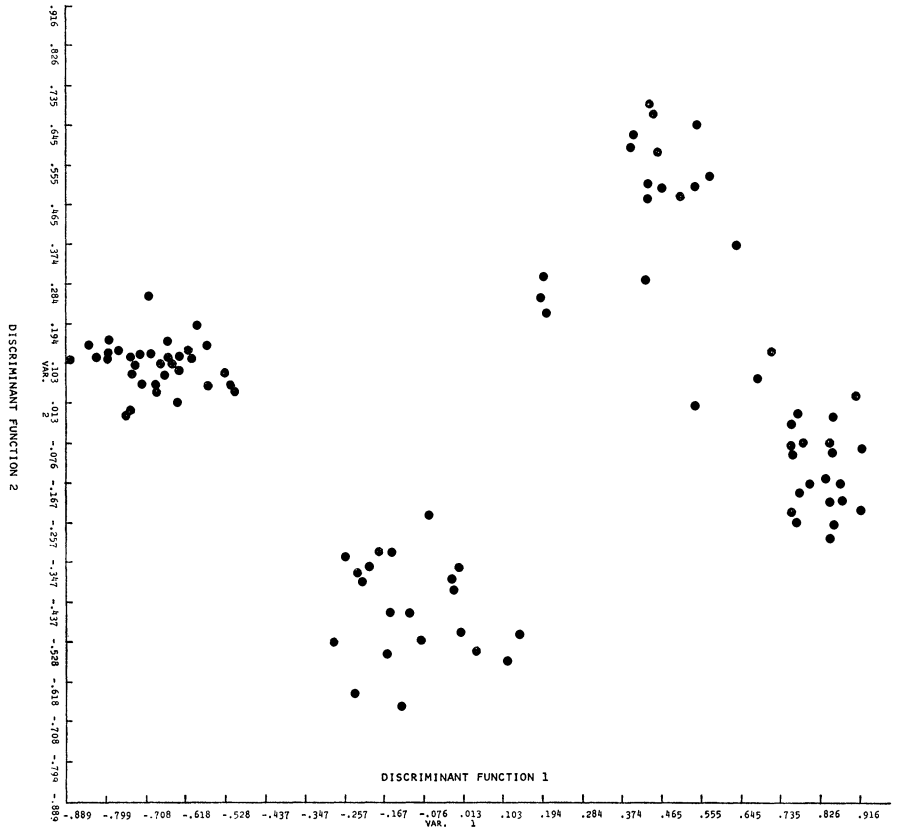

FIG. 10. Bee data determinant criterion
4 Groups 3 Components

This content downloaded from
86.59.13.237 on Wed, 27 Jan 2021 17:36:17 UTC
All use subject to https://about.jstor.org/terms

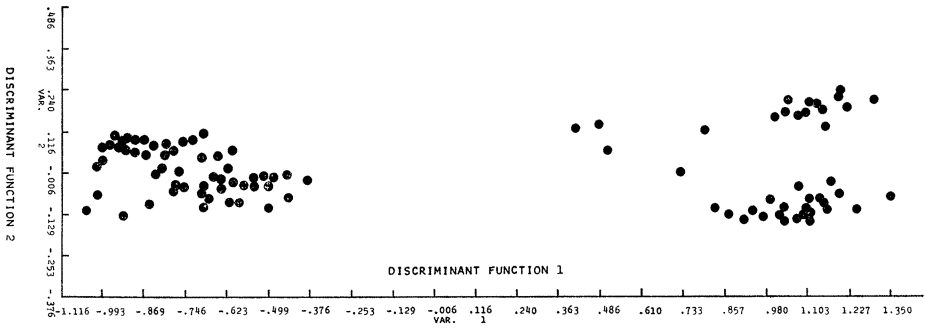

FIG. 11. Bee data, trace W^{-1} B criterion
4 Groups 3 Components

The max Trace $W^{-1}B$ criteria split the data correctly into two groups. However, in going to four groups it continued to split the data along the eigenvector direction determined by the two group partition. The result was a distortion of the data as shown in figure 11.

In summary if we keep only the three largest principal components and ask for the "best" partition of the 97 species of bees into four groups using either min Trace W or max $|T|/|W|$ as a criterion we produce the four genera as defined by Professor Michener & Sokal (except for species 41 as indicated for min Trace W.)

7. CONCLUDING REMARKS

If we must choose a single criterion function with which to explore the structure of heterogeneous multivariate data, we would choose $|T|/|W|$ since it is invariant under non-singular linear transformations and has demonstrated on the data analyzed for this report a greater sensitivity to the local structure of data than the other criteria. In particular the groups resulting from the trace $W^{-1}B$ criterion were always separable by a single discriminant function (a single direction in space). This was not true for the $|T|/|W|$ criterion.

We think this was due to the nature of the functions, namely:

$$\text{Trace } W^{-1}B = \sum_{i=1}^{t} \lambda_i \quad \text{and} \quad \frac{|T|}{|W|} = \prod_{i=1}^{t} (1 + \lambda_i)$$

Trace $W^{-1}B$ could be increased most easily by increasing the largest eigenvalue, while the determinant criterion, since it includes product terms, reflects changes in the smaller eigenvalues. See figures 10 and 11 for a dramatic instance of the foregoing remarks.

This result is nice from the point of view of computation since it is faster to compute $|W|$ than W^{-1}. In the use of the determinant criterion W^{-1} is computed only for the final output.

It is also substantially more time consuming to compute all eigenvalues of $W^{-1}B$ at each iterative step of the "hill climbing" procedure. However, it is conceivable that one could formulate other functions of the eigenvalues of $W^{-1}B$

This content downloaded from
86.59.13.237 on Wed, 27 Jan 2021 17:36:17 UTC
All use subject to https://about.jstor.org/terms

that might be as satisfactory as the determinant criterion for the purpose of "cluster analysis".

The minimum Trace W criterion is much less costly in computer time than the other criteria. Its major fault is that it does not take into account the within-group covariance of the measurements (as in the Iris data). Transforming variables to have unit variance does not always help.

In summary, we think a major reason for applying these procedures is to get a better insight into the geometric structure of the data in the p dimensional measurement space, and consequently a better understanding of the measurements themselves.

From this point of view if there was little prior knowledge about the group structure of a set of data we would analyze this data with several different criteria before coming to any strong conclusions.

8. ACKNOWLEDGEMENTS

We would like to thank Professors T. W. Anderson, D. R. Cox and I. Olkin for helpful discussion and encouragement. We would also like to thank the editors for useful suggestions.

REFERENCES

[1] Anderson, T. W. (1958). *An Introduction to Multivariate Statistical Analysis*, New York, John Wiley and Sons.

[2] Ball, G. (1965). "Data Analysis in the Social Sciences." Volume 27, Part 1, pages 533–560 in American Federation of Information Processing Societies Conference Proceedings. Fall Joint Computer Conference. Washington: Spartan Books; London: Macmillan.

[3] Bargmann, R. and H. Posten (1964). "Power of the Likelihood-ratio Test of the General Linear Hypothesis in Multivariate Analysis," *Biometrika*, 51, #304, pp. 467–80.

[4] Bose, R. C. and S. N. Roy (1938). "The Distribution of the Studentized D^2 Statistic." *Sankhya* Vol. 4, Part 1, pp. 19–38.

[5] Cacoullos, T. (1965). Comparing Mahalanobis Distances I. Sankhya Series A, Vol. 27, Part 1, pp. 1–22.

[6] Cacoullos, T. (1965). Comparing Mahalanobis Distances II, Sankhya Series A, Vol. 27, Part 1, pp. 23–32.

[7] Dempster, A. P. (1958). "A High Dimensional Two Sample Significance Test." *Annals of Math. Statistics* Vol. 29, No. 4, pp. 995–1010.

[8] Edwards, A. W. F. and Cavalli-Sforza, L. L. (1965). "A Method for Cluster Analysis." *Biometrics*, Vol. 21, No. 2, pp. 362–75.

[9] Fisher, R. A. (1936). "Multiple Measurements in Taxonomic Problems." *Annals of Eugenics*, Vol. VII part II, pp. 179–88.

[10] Fisher, W. (1953). "On a Pooling Problem From the Statistical Decision Viewpoint." *Econometrica*, 21, 567–85.

[11] Fisher, W. (1958). "On Grouping for Maximum Homogeneity," *Journal of American Stat. Assn.*, Vol. 53, pp. 789–98.

[12] Forgy, Edward [1965] Cluster analysis of multivariate data: Efficiency vs. interpretability of Classifications. WNAR meetings, University of California, Riverside, June 22–23, 1965. (see abstract, *Biometrics*, 21(3) p. 768).

[13] Fortier, J. and H. Solomon (1966). "Clustering Procedures," pp. 493–506, Multivariate Analysis, Edited by P. R. Krishnaiah, Academic Press, N. Y.

[14] Good, I. J. (1965). "Categorization of Classification," *Mathematics and Computer Science in Biology and Medicine*. Her Majesty's Stationery Office, London.

This content downloaded from
86.59.13.237 on Wed, 27 Jan 2021 17:36:17 UTC
All use subject to https://about.jstor.org/terms

[15] Gower, J. C. (1966). "Some Distance Properties of Latent Root and Vector Methods Used in Multivariate Analysis" *Biometrika*, Vol. 53, December 1966, pp. 325–28.
[16] Hotelling, H. (1954). "Multivariate Analysis," *Statistics and Mathematics in Biology*, Edited by Kempthorne. Hafner, N. Y.
[17] Kendall, M. G. (1966). "Discrimination and Classification," pp. 165–84, *Multivariate Analysis*, Edited by P. R. Krishnaiah, Academic Press, N. Y.
[18] Kruskal, J. (1964). Multidimensional Scaling by Optimizing Goodness of Fit to a Non-Metric Hypothesis. *Psychometrika*, 29, 1–28.
[19] MacQueen, J. B. (1966). Some Methods for Classification and Analysis of Multivariate Observations. Working paper No. 96. UCLA, W.M.S.I.
[20] Michener, C. D. and R. R. Sokal (1957). "A Quantitative Approach to a Problem in Classification." *Evolution* 11: 130–62.
[21] Rao, C. R. (1952). *Advanced Statistical Methods in Biometric Research*. John Wiley & Sons., Inc., N. Y.
[22] Rao, C. R. (1964) The Use and Interpretation of Principal Component Analysis in Applied Research. *Sankhya*, Series A, Vol. 26, Part 4, Dec., pp. 329–58
[23] Rota, G. (1964). "The Number of Partitions of a Set." *American Mathematical Monthly*, Vol. 71, No. 5, pp. 498–504.
[24] Rubin, J. (1967). "Optimal Classification into Groups: An Approach for Solving the Taxonomy Problem." *Journal of Theoretical Biology*, 15, pp. 103–144.
[25] Schatzoff, M. (1966). "Sensitivity Comparisons Among Tests of the General Linear Hypothesis," *Journal of American Stat. Assoc.*, June 1966, V. 61, No. 314, Pt. 1, pp. 415–35.
[26] Shepard, R. N. (1962). The Analysis of Proximities: Multidimensional Scaling with an Unknown Distance Function I. *Psychometrika*, 27, 125–40.
[27] Singleton, R. (1965). Personal Communication. Stanford Research Institute, Menlo Park, California.
[28] Sneath, P. H. A. and R. R. Sokal (1963). *Principles of Numerical Taxonomy*. W. H. Freeman & Co., San Francisco.
[29] Solomon, H. (1955). "Probability and Statistics in Psychometric Research: *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability*" Vol. V.
[30] Torgerson, W. S. (1965). Multidimensional Scaling of Similarity. *Psychometrika*, Volume 30, No. 4, Dec., pp. 379–93.
[31] Weiner, J. M. and Dunn, O. J.: Elimination of Variables in Linear Discrimination Problems. *Biometrics*, Vol. 22, No. 2, June 1966, pp. 268–75.
[32] Wilk, M. and R. Gnanadesikan (1965) et al. "Comparison of Some Statistical Distance Measures for Talker Identification." Unpublished manuscript. Bell Telephone Laboratories, Murray Hill, N. J.
[33] Wilks, S. (1960). "Multidimensional Statistical Scatter" *Contributions to Probability and Statistics*, Edited by I. Olkin. Stanford University Press.
[34] Wilks, S. (1962). *Mathematical Statistics*. John Wiley & Sons, N. Y.

This content downloaded from
86.59.13.237 on Wed, 27 Jan 2021 17:36:17 UTC
All use subject to https://about.jstor.org/terms