MULTIVARIATE DATA ANALYSIS

Herbert Solomon

Stanford University

Prepared for:

Office of Naval Research
Army Research Office
Air Force Office of Scientific Research

3 February 1975

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE		READ INSTRUCTIONS BEFORE COMPLETING FORM
1 REPORT NUMBER: 216	2. GOVT ACCESSION NO.	3 RECIPIENT'S CATALOG NUMBER
4 TITLE (and Subtitle): MULTIVARIATE DATA ANALYSIS		5. TYPE OF REPORT & PERIOD COVERED: Technical Report
		6. PERFORMING ORG. REPORT NUMBER
7. AUTHOR(s): HERBERT SOLOMON		8. CONTRACT OR GRANT NUMBER(s): N00014-67-A-0112-0085
9 PERFORMING ORGANIZATION NAME AND ADDRESS: Department of Statistics Stanford University Stanford, California 94305		10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS: (NR-042-267)
11 CONTROLLING OFFICE NAME AND ADDRESS: Office of Naval Research Statistics & Probability Program Code 436 Arlington, Virginia 22217		12. REPORT DATE: February 3, 1975 13. NUMBER OF PAGES: 40
14 MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office)		15. SECURITY CLASS. (of this report): Unclassified 15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16 DISTRIBUTION STATEMENT (of this Report)

Approved for public release: distribution unlimited

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

19 KEY WORDS (Continue on reverse side if necessary and identify by block number)

multivariate data analysis; clustering; multidimensional contingency tables; factor analysis

20 ABSTRACT (Continue on reverse side if necessary and identify by block number)

This paper contains an account of several techniques in multivariate data analysis. Included among these techniques are classification and clustering procedures, multidimensional contingency table analysis, and some graphical representation techniques. Some data bases are employed to illustrate the techniques.

DD FORM 1 JAN 73 1473 EDITION S/N 010:

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

MULTIVARIATE DATA ANALYSIS

by

HERBERT SOLOMON

TECHNICAL REPORT NO. 216
FEBRUARY 3, 1975

PREPARED UNDER CONTRACT N00014-67-A-0112-0085
(NR-042-267)
OFFICE OF NAVAL RESEARCH

Herbert Solomon, Project Director

DEPARTMENT OF STATISTICS
STANFORD UNIVERSITY
STANFORD, CALIFORNIA

MULTIVARIATE DATA ANALYSIS

Herbert Solomon
Stanford University

Table of Contents

MULTIVARIATE DATA ANALYSIS*

by

Herbert Solomon
Stanford University

1. Introduction

There has always been a need to achieve parsimonious yet operationally meaningful accounts of what is going on in nature and in human behavior. We are aware of attempts by biologists to classify flora and fauna, and even that dichotomy was a major step forward. It is in the physical and life sciences that we find the first quantifiers at work on such matters. Later we find social anthropologists and psychologists engaging in studies on how groupings can be accomplished. Today we find numerical taxonomy pervasive in practically every field of study. This has been spurred by increased activity in data collection and developments in computer technology. Multiple measurements on elements, individuals, or variables abound nowadays, and one sees investigators scurrying about to apply discriminant analysis, classification or clustering techniques, multidimensional contingency table analysis, factor analysis, and with good reason. We will return to these topics.

Even though we regard classification in social sciences as rather new, it is difficult to think of its counterpart in physical sciences as very old unless one thinks of a few hundred years in the course of mankind as a very long step. It was just two or three hundred years ago that many physical ailments were labeled "consumption", because they were characterized by a "wasting away of the tissues". Under this were

---

*This is an extended version of an invited talk given at the 20th Annual Army Design of Experiments Conference, Ft. Belvoir, Virginia, October 1974.

lumped such diseases as leprosy, tuberculosis, diabetes, and others. It was not until some time later that someone noted that the urine of some of these sufferers was sweet and that of others was not. Of course, the subsequent discoveries of two different bacilli for leprosy and tuberculosis suggested finer groupings that obviously were more meaningful in connection with specific treatments.

There is a lesson here for all of us, namely that the classification and grouping of individuals or elements based on data analyses of sets of variables can lead to man-made group concoctions that are artificial and sometimes misleading. What should be kept in mind is that when this is done, a grouping has some meaning to the investigator. For the last forty years or so, aberrant mental behavior has been subjected to classification and groupings produced on the basis of observations made on any number of variables. For an individual placed in one of these groupings, some treatment is suggested. I imagine one does not feel as comfortable here in a diagnosis as in the case of diabetes or tuberculosis groupings at present; and rightfully so. Yet treatment will be undertaken based on a diagnostic category to which an individual is assigned. This should give us pause when classification is attempted by data analysis in the newer investigations such as those that occur, for example, in the reenlistment decision in the armed services.

2. History

It is in the late 19th century that we find a blossoming of inquiries into classification through the selection and appropriate use of manifest variables. Quite often a one-dimensional index that

incorporates all pertinent variables was sought so that a technician could assign an individual to one of several groups based on his responses to the variables employed. For example, the coefficient of racial likeness was an index developed at the turn of the century to distinguish different national or tribal groups on the basis of a set of physical measurements. Inquiries on association of criminal types with physical measurements of individuals also received attention in this period by such investigators as Lombroso.

Much of this inquiry took place in the British community of scholars. In a way it might be viewed to have begun at least in a larger sense with Charles Darwin's vast collection of data arising from his travels around the world. His diaries presented many observations on the animal kingdom and served as a base for study by many who came later in the 19th century.

It was with these investigators in the last quarter of the 19th century that we have the beginnings of statistical contributions to classification. In fact, it is the classification problem that in a way motivated and created statistical inference as an area of scientific inquiry. The modern discipline we now call statistics was brought about by the anthropometrists, biologists, and psychologists of that era. Such initial contributors to modern statistics as Francis Galton and Karl Pearson stem from that period.

Galton seemed to be perpetually engaged in data analysis. He and his cousin, Darwin, and others revolved in an age of scientific inquiry that emphasized empiricism. Pearson, along with others, later attempted quantification and mathematization from the empirical analyses provided

by their colleagues. Galton, whom we regard as the founder of regression analysis through his study on relationships between children's heights and parents' heights, also initiated and developed the notion of correlation prior to 1885. The correlation coefficient serves as a basic summarization in multivariate data analysis and consequently in studies that go into techniques of grouping. From its very nature, obviously a high correlation coefficient would indicate that the two variables belong in a group and a low correlation would suggest that they do not.

In one of his papers in 1888, Galton became interested in the classification problem. He pointed out that 12 measures proposed by Bertillon to be used for classification of criminals were not independent and suggested that the observed measurements be transformed into a set of independent measures. He also suggested the method of transformation, which we can now view as simple or unweighted summation in factor analysis. Thus quite early we see the intermingling of classification analysis and factor analysis - and of course this is still quite current. We will return to factor analysis and its place in classification analysis.

Pearson was engaged in studies that were obviously related to classification. In an interesting paper in 1901, he discussed mathematical representations of lines and planes of closest fit to systems of points in space. This geometrical way of looking at the classification problem may present a neater view of the problem to some. In effect, the multi-dimensional observations at hand, e.g., age, IQ, schooling, number of dependents, rank, length of enlistment, etc., for each member of a

population of N members up for reenlistment decision can be viewed as N points in a 7-dimensional space. Moreover, each point cannot be reached by traveling along 7 perpendicular axes, for the 7 variables can and usually have degrees of association which must be taken into account.

This effort is a fundamental problem in multivariate data analysis, namely finding a grid of orthogonal axes to replace the grid of correlated axes (naturally the points remain where they are). If the number of dimensions can be reduced to two or three, some ease is achieved since elements can be grouped by eye. In fact, this is related to one of the central problems in factor analysis and is pertinent to the use of factor analysis as a classification technique.

3. Assignment Procedures and Discriminant Analysis

It is now important to be specific about the term "classification". For our purposes, we will assume that the term comprises both the clustering of data into groups and the assignment of data to previously specified groups. Actually, the latter can be valued as a subset of the former. In the former category, we require the data to produce both the number of groupings or clusters and the assignment of each element or individual to these groupings. In the latter category, the number of groups or clusters is predetermined. Each group is labeled, and rules are designed on the basis of which an assignment of each element is made to one of the fixed groups.

We do not wish to convey a sharp distinction between clustering and assignment procedures. If a classification procedure is not producing meaningful groups through the assignments that are made, then changes are called for, namely revising the predetermined groupings either in

number or in shape or in both on the basis of the new information. This sequential revision of groups on the basis of the data available at different times suggests that one is indirectly engaging in clustering procedures. On the other hand, it is wise to keep in mind the conceptual differences just mentioned between attempts at clustering and attempts at assignment.

An essential step in classification procedures is the representation of the relationships among the variables on which data has been collected. Among other important and prior steps, there are the processes of developing numbers to measure phenomena, making decisions on the employment of nominal, ordinal or continuous data, and subsequent coding of this data for analysis. In this paper, we do not review these issues, but we are mindful of their impact on the data analysis that will undergo investigation. Thus, we return quickly to clustering and assignment techniques and the basic summarizations of data for these purposes.

The clustering and assignment problems, even though they were recognized for some time, did not possess any techniques until rather recently. The assignment problem received the first thrust. The analysis was provided by one of the great savants of modern statistical inference, namely R. A. Fisher. In a paper in 1936, we find what is now Fisher's classic work on discriminant analysis. It is entitled "The Use of Multiple Measurements in Taxonomic Problems" and was published in The Annals of Eugenics. The author was to say somewhat later that the paper was written to embody the working of a practical numerical example arising in plant taxonomy in which the concept of a discriminant function seems to be of immediate service. This is a simple but fascinating statement, because

it demonstrates once again that when there is a problem requiring solution some strides can be made. Too often we find solutions looking for a problem, and this is something we should be especially concerned with in classification problems.

In his paper, Fisher also listed the basic data he analyzed. This is rarely done by authors, and so we find the Fisher data and just a few other data bases referred to time and time again by subsequent authors who are experimenting with new assignment or clustering techniques. In this way, an anchor is provided against which the results of other techniques can be assessed.

The data employed by Fisher was supplied by a botanist, and it represented measurements on the irises of the Gaspe Peninsula. This data was previously published in the Bulletin of the American Iris Society and was therefore not a likely contender for a best seller. Since it is a classical piece in the statistical literature, let us look at it in some detail. Four measurements on each of fifty plants in each of three iris categories were obtained. The categories are: Iris Virginica, Iris Versicolor, and Iris Setosa. For each of the 150 plants already assigned to one of three categories, there are measurements of sepal length, sepal breadth, petal length, and petal breadth.

If we refer back to our geometrical representation, we have 150 points scattered in a four-dimensional space, except that each point is already labeled as belonging to one of three groups. The question is whether in some neat and simple way we can separate the 50 points belonging to any one group from the other two sets. This is compounded by the fact, in this case, that two of the irises, namely Versicolor and Virginica,

actually have a specific genetic relationship and obviously, then, do have some overlap. In other words, Fisher is looking for hyperplanes that partition the four-dimensional space, and after partitioning, hopefully leave each group inviolate. Algebraically, he is asking for a linear function of the four measurements (later called the discriminant function) that accomplishes this. As a reasonable index for determining the coefficients of the linear function, he suggests one that will maximize the ratio of the difference between the means to the standard deviations within species. To be specific, let $d_p, p = 1,2,3,4$ represent the difference in the observed means.

Then for any linear function, X, of the measurements, namely

$$X = \lambda_1 x_1 + \lambda_2 x_2 + \lambda_3 x_3 + \lambda_4 x_4$$

the difference between the means of X in the two species is

$$D = \lambda_1 d_1 + \lambda_2 d_2 + \lambda_3 d_3 + \lambda_4 d_4$$

while the variance of X within species is proportional to

$$S = \sum_{p=1}^{4} \sum_{q=1}^{4} \lambda_p \lambda_q S_{pq}$$

where S_{pq} is the sum of squares or products in X_p and X_q.

The particular linear function that best discriminates the two species will be one for which the ratio D^2/S is greatest, by variation of the four coefficients $\lambda_1, \lambda_2, \lambda_3, \lambda_4$. Geometrically we are locating the hyperplane that best separates two groups of points in the sense that the distance between the four-dimensional centroids is greatest. Even though there are three groups of irises, in effect Fisher acts as

if there are two groups, since Iris Versicolor and Iris Virginica are genetically tied together. Note that the variations within species is assumed to be the same in this development.

The index that is employed to provide the delineation is tied at first to the multivariate normal structure assumed for each species. Yet it is very similar to the indexes suggested by strict multivariate data analysis as we will see in the next section. Here we are maximizing the difference between the centroids of the two species of irises, or, in other words, maximizing heterogeneity between groups. This theme will carry through all of our attempts of classification. Either we will maximize heterogeneity between groups or minimize the scatter (i.e., seek homogeneity) within groups.

As a result of the analysis, Fisher arrives at a linear discriminant function that accomplishes a nice separation. For example, Iris Setosa is separated completely from Versicolor and Virginica. It turns out that only one of the four measurements is really necessary to do this, namely petal length, and this can probably be seen by just looking at the 150 sets of measurements. This should be something for us to highlight, especially when we get into data sets for which meanings are not so specific and measurements are not so commensurate. This will obviously be so in any number of studies in criminal justice.

Fisher's work has been extended to assign an element to any one of k groups, and computer programs exist in Computer Center libraries to accomplish multiple linear discriminant analysis. Attached to this subject is the question of how many variables should be used in a discriminant function. It is obvious that the more variables one uses, the better the discrimination should be, but it is also obvious that the

marginal gain in using additional variables can decrease sharply and therefore some variables can best be omitted in the interests of parsimony. Thus we seek the best discriminating variables.

We might also ask what one would do if one were faced with the 150 irises and did not know their groupings; that is, if we had only the four measurements on each, and we wished to see what number of groupings as well as assignments could be made. Here we are no longer faced with the assignment problem alone, but with the clustering problem or grouping problem, which of course subsumes an assignment problem. It is to this topic that we now turn.

4. Data Summarization

It is important in talking about grouping to consider whether we are grouping measurement variables or individuals or elements of a population. For the iris data, we are grouping elements of a population. Quite often, one is interested in grouping measurement or test variables. The basic data summarization in multivariate data analysis will depend on whether we are grouping variables or elements. We will resolve this in subsequent discussion by first going in some detail into the data summarization question.

There are several ways to begin the data summarization. All give a picture of data interrelationship, but each has special reasons for its employment by an investigator. One representation is that of the scatter matrix. Here we portray the total scatter or dispersion displayed by n individuals or elements each measured on p variables (n points in a p-dimensional space) by a matrix with p rows and p columns where an element in the i^{th} row and j^{th} column, say t_{ij}, is

the sum of the n cross products of measurements (taken around the mean) on variable x_i with measurements (taken around the mean) on variable x_j. In brief,

$$t_{ij} = \sum_{i<j=2}^{p} \sum_{k=1}^{p} (x_{ik}-\bar{x}_i)(x_{jk}-\bar{x}_j),\ t_{ij} = t_{ji},\ \bar{x}_i = \frac{\sum_{k=1}^{n} x_{ik}}{n} .$$

Let us label this matrix T. Naturally an element in the main diagonal, say i^{th} row and i^{th} column, is the sum of the squares of the deviations of x_i from its mean. If $p = 1$, then T is a scalar, namely

$$\sum_{k=1}^{n} (x_k-C)^2 \quad \text{where} \quad C = \frac{\sum_{k=1}^{n} x_k}{n} .$$

If each element in the scatter matrix T is divided by n, the resulting matrix is the covariance matrix with cell entries s_{ij} and we label this K. Now if we also divide each element, s_{ij}, in K by the standard deviations of x_i and x_j, the resulting element $r_{ij} = s_{ij}/s_i s_j$ is the correlation coefficient between x_i and x_j and the resulting matrix is now the correlation matrix which we label R.

An important advantage of T is the manner in which it can be decomposed into two matrices that are especially pertinent in clustering and classification studies. In a classification study, the n elements will be assigned to k predetermined groups. Each group with, say, n_i elements can be viewed as a universe with its own scatter matrix formed as before and labeled W_i. If we sum all the W_i scatter matrices, we get $W = \sum_{i=1}^{k} W_i$ and let this represent the within scatter

or homogeneity of the groupings. Likewise, if for each of the k groups, we compute the group mean (a p-dimensional vector where the r^{th} coordinate is the mean value based on the n_r observations for x_r) and then produce the $(p \times p)$ matrix that we label B, for it expresses a measure of the "betweenness" or heterogeneity of the k groups. The central point in this development is the existence of the fundamental matrix equation

$$T = W + B \quad .$$

This result suggests immediately an index by which classification (predetermined number of groups) can be evaluated and, by extension, how clustering can be terminated at some cluster size. For any given data set T is fixed. Thus measures of "groupiness" or "clusteriness" as functions of W and B are thrust forth for examination.

For $p = 1$, the matrix equation reduces to an equation about scalars. Thus a good grouping index is one which minimizes W or equivalently maximizes B. We may also consider maximizing either the ratio B/W or $T/W = 1 + B/W$. An added benefit is that this ratio is invariant under linear transformations of the data. Statisticians have long exploited this fact, for B/W multiplied by an appropriate constant is the familiar F ratio in the analysis of variance.

When the number of measurements per element is two or more $(p > 1)$, grouping criteria are not so straightforward. Several possibilities suggest themselves and have been developed and studied by investigators. One criterion suggested by several authors that is a quite natural index is the minimization of the trace of W (sum of all elements in

the main diagonal of the matrix) over all possible partitions into k groups. This is equivalent to maximizing Trace B because

$$\text{Trace } T = \text{Trace } W + \text{Trace } B \quad .$$

However, Trace W is invariant only under an orthogonal transformation and not under non-singular linear transformations.

Another criterion that may be employed for $p > 1$ is the ratio of the determinants

$$|T|/|W| = |1 + W^{-1}B| \quad .$$

We can use $|T|/|W|$ as a criterion for grouping and select that grouping for which this index is maximized, or equivalently $|W|$ is minimized. Also we may employ $\log(|T|/|W|)$ since it is a monotonic function.

Another criterion for grouping is the trace of $W^{-1}B$ and we select the grouping that maximizes this index. This index has been used as a test statistic in multivariate statistical analysis as has the ratio $|W|/|T|$. The latter was employed by Wilks to test whether groups differ in mean values, and the former has been put forth by Hotelling in some situations and by Rao as a generalization of the Mahalanobis distance between two groups for $k > 2$ groups. We will shortly define and discuss the implications and uses of the Mahalanobis distance in clustering procedures.

Both Trace $(W^{-1}B)$ and $|T|/|W|$ may be expressed in terms of the eigenvalues, λ_i, of the matrix $W^{-1}B$. We write

$$|T|/|W| = \prod_{i=1}^{p} (1 + \lambda_i)$$

and

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

where λ_i are the roots of the determinantal equation, $|B-\lambda W| = 0$. The characterization of these ratios in terms of eigenvalues is helpful in data representation especially when the effects of some reduction in dimensionality is desired. All the eigenvalues of this equation are invariant under non-singular linear transformations of the data. It can be proved that these eigenvalues are the only invariants of W and B under non-singular linear transformations.

5. Distance Matrix

Thus far we have discussed some summarizations of multivariate data in matrix form, either T (scatter), K (covariance), or R (correlation) and the kinds of grouping criteria that are suggested by the T format. Intuitively, we see that any grouping criterion is a function of homogeneity within groups and heterogeneity between groups and the indexes already described are specific quantities embodying these notions. We shall discuss other indexes as we proceed, but each will be a function of homogeneity within groups and heterogeneity between groups in which attempts will be made to minimize the former, maximize the latter, or in effect do both. For the correlation coefficient index, large values indicate homogeneity; small values indicate heterogeneity.

Another method of summarizing data that is more appropriate on occasion is to find the distance between each pair of the n points in the p-dimensional space. This leads to a representation in matrix

form of an $n \times n$ matrix where each element, in the i^{th} row and the j^{th} column, say d_{ij}, is the distance in the p-dimensional space between the i^{th} element or individual and the j^{th} element or individual. All the elements in the main diagonal are zero. The distance matrix is akin to the correlation matrix in that both may be viewed as similarity matrices - the jumping-off place for clustering attempts.

The decision as to whether correlation matrices or distance matrices are to be employed is usually determined by the problem at hand. If n individuals or n elements are to be grouped on the basis of p measurements on each, then the $n \times n$ distance matrix is the natural summarization; if the p measurement variables are to be grouped on the basis of the measurements on n individuals or n elements, then the $p \times p$ correlation matrix is the natural summarization of the data. This latter matrix is the natural beginning point in factor analysis where parsimony in the number of latent measurement variables is a desired goal. We will return to factor analysis and its place in clustering in subsequent sections. In some taxonomic situations the question of which measure of similarity to employ, whether it is of the association or distance type, will require some thought. While we will touch on these points, these inquiries will not be featured in this exposition.

The notion of a distance matrix will be placed in sharper focus, and this will be done by some discussion of appropriate distance measures. Because we will normally think of our data bases for clustering individuals or elements as n points in a p-dimensional space, the distance measures usually appropriate and available are Euclidean distance and Mahalanobis distance. The Euclidean distance between individuals or elements with

respect to all p measurement variables may be written in vector notation

$$d_{ij}^2 = (P_i - P_j)'(P_i - P_j)$$

where d_{ij} is the Euclidean distance between individual i and individual j, P_i and P_j are column vectors each with p rows listing the p measurements on the i^{th} and j^{th} individuals respectively. The product of the difference row vector $(P_i - P_j)'$ by its transpose is a scalar. This is the distance function with which most of us are familiar. The Mahalanobis distance may be written as in the notation above as

$$_m d_{ij}^2 = (P_i - P_j)'W^{-1}(P_i - P_j)$$

where W^{-1} is the inverse matrix of $W = \sum_{i=1}^{k} W_i$ and W_i is obtained for each of the $i = 1,3,\ldots,k$ groups by

$$W_i = \sum_{m=1}^{m} (P_{mi} - C_i)(P_{mi} - C_i)' \quad .$$

Note that a grouping of elements is necessary to compute W_i and consequently W. Thus the Mahalanobis distance takes into account the associations or interrelationships in the measurement variables. If two measurement variables are highly correlated, the Euclidean distance can be misleading because of the equal weight it imposes inaccurately on each measurement variable, but this will not be so with the Mahalanobis distance. The Mahalanobis distance is more tedious to compute and for

a long time it was avoided for this reason alone, but the computer has brought it within reach. Actually if each of the correlations between the measurement variables is low, the error in employing the Euclidean distance is not damaging. As a rule of thumb, correlations as high as 0 • 5 will not produce Euclidean distances that lead to operational difficulties.

Other distance measures appear in the literature. The Minkowski distance is the name applied to all distance measures that are of the form

$$d(i,j) = \left\{ \sum_{m=1}^{p} |x_{im}-x_{jm}|^{n} \right\}^{1/n} .$$

We have discussed the case $n = 2$. When $n = 1$, the label "city-block" distance is sometimes employed and it may be relevant for some distance situations.

6. Clustering

We now look at the clustering side of classification analysis. Our main emphasis will be on clustering as an exploratory device. Development of assignment procedures is for those who already enjoy the luxury of knowing the groups that exist. We will place ourselves in the situation where a body of multidimensional data has been collected by some investigator and he wishes to decipher what kind of structure, if any, underlies the data collected. A wide variety of techniques have been suggested and attempted. They run the gamut from looking at all possible partitionings of the data to erving to zero in on an optimal partitioning without having to look at too much of all the possibilities. The former method is a "dumb" procedure which

is workable if the computer can quickly look at everything, and of course this is not so even for a small number of observations in a small number of dimensions. Thus we sacrifice optimal partitioning for what we hope are suboptimal partitions that can be achieved much more cheaply.

Let us consider one general way of looking at the problem considered by several authors. 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 a clustering criterion is increased, leave the object where it is. Otherwise, move it so that the maximum increase in the criterion occurs. Naturally, we are assuming the existence of a reasonable criterion. Using the partition thus created, we process the second object in the same way, then the third, etc. After several passes, one will 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 "local maximum" of our criterion function. This rarely takes more than a reasonable time on a computer. This has been labeled the "hill-climbing" pass algorithm by Friedman and Rubin.

They and others have suggested modifications. For example, we start with the best partition yet known. Then process one group at a time, in sequence, by placing each object of the group being processed into the outside group with nearest center of gravity, recalculating the criterion function after each move. This is done in 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. This has been labeled a "forcing pass" algorithm. It 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.

Still another procedure proposed by Friedman and Rubin and others 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. With either of the other two criteria just discussed, we use the metric defined by the matrix W^{-1} computed from the partition P -- i.e., $d(P,C_k) = (P-C_k)W^{-1}(P-C_i)^T$. 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 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 then hill-climbing is resorted to for a new local maximum. Other modifications are also applied, but when it proves impossible to reach a higher local maximum, the procedure is terminated. If one is willing and

financially able to spend the computer time, one can repeat the entire procedure using another starting partition chosen at random or, as we will soon see, obtained by a quick step-wise method. The forcing and reassignment passes are fast, but only occasionally helpful. Restarting from each of several random partitions or the step-wise solution is slow but provides more confidence in the result.

7. Initial Partitioning

There is a much simpler way of initiating clustering. It was proposed by King and in effect gives a quick initial partitioning of the data whether it be measurement variable groupings or delineation of individuals in a population. Either something of interest and use to the investigator appears quickly, or what does emerge can serve as the first step for those algorithms that require a start upon which various kinds of iterations are attempted. These were just described in the previous section.

The procedure proposed by King is a step-wise clustering procedure. This is its principal asset because it leads to a simple and quick algorithm that involves $(n-1)$ scannings of a correlation matrix based on n variables. At each scanning or pass, the variables are sorted into a number of groups that is one less than at the previous pass. In this way, we obtain $(n-k)$ groups of variables at the k^{th} scanning. The $(n \times n)$ matrix can also be a distance matrix. In that case, we sort individuals or elements into groups.

The procedure operates as follows. We will employ the correlation matrix as our similarity matrix for expository purposes, and bring in the distance matrix when appropriate to highlight differences.

As a start, we can view the n variables as n groups, one variable to each group. Now scan the correlation matrix for the maximum cell entry (naturally without regard to sign). In a distance matrix we would seek the minimum distance cell entry. Suppose the maximum correlation is between variables X_i and X_j. Label it $r_{i,j}$. We place X_i and X_j in the same group, and we now have $(n-1)$ groups $X_1X_2,\ldots,(X_i,X_j),\ldots,X_{n-1},X_n$. This produces an $(n-1)\times(n-1)$ correlation matrix, all pairs of correlation coefficients over the original $(n-2)$ variables plus the correlations obtained by pairing each of these with the concocted variable $X_i + X_j = Y_{ij}$. Essentially, we are representing the group of two elements by its centroid.

On the second pass of what is now an $(n-1)\times(n-1)$ correlation matrix, a third variable may join the group of two variables formed on the first pass if the correlation between it and Y_{ij} is maximum, or the maximum correlation value in the reduced correlation matrix may again involve two individual variables. Thus we would get either one group of three variables and $(n-3)$ groups each containing one variable, or two groups each containing two variables and $(n-4)$ groups each containing one variable. In either situation we merge variables and revise the correlation matrix as on the first pass. In the former case, the centroid of the group of three variables represents its group, and in the latter case, each group with two variables is represented by its centroid. Recall that we do not have to divide the sum of the variables by the number of variables to obtain the centroid because the correlation coefficient is invariant when one variable of the pair is always multiplied by the same constant.

Thus, at each pass, the two groups with the highest correlations are merged and the total number of groups to that point is reduced by one. After a variable has joined a group of variables, it cannot be removed from that group. In this way it is possible to miss an optimal grouping. This is very similar to selection of predictors in step-wise linear regression. It should also be mentioned that a group can lose its identity by merging with another group on a later pass. By the time all the scanning is completed we have produced successively $(n-1)$, $(n-2)$, $(n-3),\ldots,3,2$ groupings.

The clustering index employed by King for measuring the worth of the grouping is that of minimum correlation (or maximal distance) between the group centroids when the scanning has placed the variables into two groups. This leaves something to be desired because it does not look at the effectiveness of the grouping when more than two groups are involved. He also reviews another index, suggested originally by Wilks for testing the mutual independence of k subsets of n multivariate normal random variables. In terms of what we described earlier in the paper, the index is the ratio of the determinants

$$Z = \frac{|T|}{\prod_{i=1}^{k} |W_i|}$$

where T is the scatter matrix defined previously and each W_1 is the scatter matrix for each of the k groups.

This index has some nice geometrical and statistical properties. For example, when $k=2$,

$$Z = \frac{|T|}{|W_1| \cdot |W_2|} = \Pi(1-r_i^2)$$

where r_i is the i^{th} canonical correlation between the two sets of variables. This index may be viewed as a "generalized alienation coefficient" since it is an extension of $1-R^2$, where R is the multiple correlation coefficient occurring when two groups have one variable in one group and $(n-1)$ in the other. However, it is not too useful in some data analyses, especially in social science, because a number of data sets lead to quasi-singular correlation matrices and truncation error can give ridiculous results. For this reason, and possibly others, negative determinants appear and make it impossible to employ the Wilks index.

Let us look at the King method for two particular data bases. The first is in connection with a penalty jury decision in California, and the second is the iris data we discussed previously.

Individuals convicted of murder: 238 individuals convicted of first-degree murder in California over a recent ten-year period were studied on the basis of 25 measurements each as to whether an association existed between their 25-dimensional descriptions and the penalty decision that resulted in life imprisonment for 135 and capital punishment for 103. These 25 variables consisted of biographical information on the individual, description of the crime, information on defense counsel, the prosecution, and the judge. A King step-wise clustering procedure was employed to cluster the 238 individuals and then seek a substantive association, if any, between the characteristics of the individual, characteristics of the crime, judicial process, and the penalty decision. My thanks for the data under analysis go to several Law Review students at Stanford with whom I worked on this study. One of their major concerns was to see if there were any association between the penalty decided upon by a jury, which

under the law is given no instruction on standards to be employed in arriving at a decision, and socio-economic characteristics or racial and ethnic background of the individual. The clustering printout did not reveal any significant associations between penalty and whether the defendant was black, Mexican-American, or white; or whether the defendant was a blue-collar worker or not. At the 58th pass, there was one significant group that contained 18 members, all of whom had received the life penalty. As the number of passes increased, this group remained the principal group until the last few passes. At the 75th step the group contained 34 members, of whom 30 received life imprisonment. At the 100th step the group contained 42 life cases out of 62 members, and at the 125th step, the group contained 63 life cases out of 102 members--a 62 to 38 percent mixture for all 238 cases. What we seem to be getting is clustering indicating very little or no association of penalty with defendant and judicial characteristics. This may also have judicial implications; for a penalty jury is, in effect, tossing for each defendant a coin which lands head or tail in a 55 to 45 percent ratio.

Irises: In Fisher's well-known paper on the linear discriminant function, he employed three groups of irises, each containing 50 members. Sepal width and length, petal width and length were obtained for each of the 150 irises--50 Iris Setosa, 50 Iris Virginica, 50 Iris Versicolor. We will assume only that we have 150 irises represented as points in a four-dimensional space which we wish to cluster by the King step-wise clustering scheme. The results are interesting. The Iris Setosa are quite different from the other two, which overlap a great deal. Thus we find at the 137th pass that there is a cluster of 48 members, each an Iris Setosa; there are

four clusters containing 23, 24, 17, and 24 members respectively, with 12, 4, 16, and 18 Iris Versicolor respectively, all demonstrating the natural overlap between Iris Versicolor and Iris Virginica. At the very next pass (138th) the two groups with 24 members each merge into a group with 48 members, 22 Iris Versicolor and 26 Iris Virginica. Thus when there is real and decided overlap the step-wise clustering scheme reflects it; but if we did not know of the original three groups, we would be hard pressed for a decision, and obviously would have to resort to additional techniques, or expertise, or both.

These data bases and several others are discussed in a paper by Solomon [11]. In that paper some computer printouts for the King procedure are displayed.

8. Data Representation Techniques

An interesting idea in multivariate data analysis has been proposed by Chernoff [1]. It is a graphical data representation technique. In his procedure Chernoff transforms multidimensional vectors into human faces. Thus, for example, several hundred vectors are transformed into several hundred faces and the faces are then classified into groups according to the similarity perceived by the classifier. The theme here is that we are very familiar through experiences in life in classifying facial characteristics. In his paper Chernoff presents a computer program which handles up to 18-dimensional vectors. The reader is referred to his paper for more details.

Up to this point, we have mentioned factor analysis but not said much about it. There is an extensive literature on this subject. Its current use in multivariate data analysis is from the representation point

of view. Computer libraries have factor analysis programs which can take large order correlation matrices and obtain principal component solutions. In this way a large number of measurement variables, say 50 to 100, can be transformed into many fewer variables, say on the order of 5 to 10. Classification and clustering can then be applied to multi-dimensional vectors of very small order. A real payoff occurs when the largest two or three factors are employed, because a graphical display can then be arranged. When this occurs, clustering or classification of the data points can be achieved by eye. See Solomon [11] for more details.

9. Multidimensional Contingency Table Analysis

A multivariate data analysis technique which is receiving more attention these days is that of multidimensional contingency table analysis (logistic response analysis). A number of authors (e.g., Kullback [8,9] and Goodman [6], among others) have done fundamental work on this technique. We will discuss this model by illustrating its use to study reenlistment decision in the armed services. The data stems from some recent Marine Corps analyses.

In this section the structure underlying contingency table analysis is discussed, and the mechanics of obtaining odds and probabilities for the reenlistment decision are illustrated. The reenlistment analysis is based on a large number of categorical variables. Regression analysis and similar multivariate techniques for continuous variables become inefficient and inappropriate for this situation. Multidimensional contingency table analysis, which we now explore, is more suitable.

We are interested in accounting for the variation in reenlistments in a parsimonious way and with meaningful factors. Consider a simple example with two factors, reenlistment decision and rank. Assume rank is categorized into two levels, i.e., high rank or low rank. The reenlistment decision and rank of forty individuals might produce the table

	High Rank	Low Rank
Reenlistment	10	10
No Reenlistment	10	10

which yields probability estimates

	High Rank	Low Rank
Reenlistment	.25	.25
No Reenlistment	.25	.25

or more generally

	High Rank	Low Rank
Reenlistment	p_{11}	p_{12}
No Reenlistment	p_{21}	p_{22}

The overall probability that a person reenlists is $p_{11} + p_{12} = .5$. The probability that a reenlistment is of high rank is also .5 for

$$\frac{p_{11}}{p_{11} + p_{21}} = \frac{.25}{.25 + .25} = .5 .$$

In this example, the probabilities of reenlistment are the same regardless of rank. This table suggests reenlistment decision and rank are independent.

A related measure denoted as an "odds" measure has an interpretation well known to bettors. In the above example, if one wagers that a person selected at random reenlists, the overall odds, i.e., the odds of reenlistment regardless of rank are one to one or even. Knowledge that the bet is on the high rank group or low rank group does not change the odds. Realistically, however, the probability and odds that a high rank and a low rank will reenlist are not the same. As an illustration, consider the table

	High Rank	Low Rank
Reenlistment	15	5
No Reenlistment	5	15

This gives probability estimates

	High Rank	Low Rank
Reenlistment	.375	.125
No Reenlistment	.125	.375

From this table the overall probability of a person reenlisting, $.375 + .125 = .5$, remains the same but the probability that a high rank reenlists is

$$\frac{.357}{.375 + .125} = .75 .$$

This differs substantially from the overall probability of 0.5 which no longer summarizes the data. The odds will change as well, being three to one for high rank, one to three for low rank. The information contained in this and the preceding table is described in terms of three

characteristics: the overall probability that a person will reenlist, the probability that a low rank will reenlist, and the probability that a high rank will reenlist.

The basic objective in a more complex table is to identify the minimum number of probabilities that must be specified to adequately describe the table. The specification of probabilities given in the last example can be used. However, recent research has developed a more formal descriptive model similar to analysis of variance or regression models. Instead of dealing directly with cell probabilities, it is convenient to deal with their logarithms. These new variables, the logarithms of the cell probabilities, have characteristics similar to measurement data, and they can be incorporated into a linear model whose parameters indicate the contribution of the various factors and their interactions to the cell probability.

The linear model for estimating logarithms of p_{tk} (for our analysis where we fix and employ only the marginals) is

$$(9.1) \qquad \ell np_{tk} = \mu + \alpha_t^T + \alpha_k^K + \alpha_{tk}^{TK}, \qquad t = 1, 2, \qquad k = 1, 2$$

where ℓnp_{tk} is the natural logarithm of p_{tk}. The constant μ is a general mean indicating the average value of ℓnp_{tk}. The parameter α^T indicates the "effect" of reenlistment decision on ℓnp_{tk} independent of rank; α^K measures the effect of rank on ℓnp_{tk} independent of reenlistment decision. The parameter α^{TK} measures the interaction effect of reenlistment decision and rank on ℓnp_{tk}. For the first example cited, where all the p_{tk} (and consequently all the ℓnp_{tk}) are equal, α^T and α^K are zero since ℓnp_{tk} does not vary with either

reenlistment decision or rank; and for this reason, too, α^{TK} is zero. Hence, p_{tk} is equal to the anti-log of μ, which in this case is the overall probability that a person reenlists.

The model in (9.1) allows the step-by-step computation of cell probabilities similar to regression analysis. For example, if reenlistment decision is considered as a function of rank, the odds of reenlistment $(t = 1)$ to non-reenlistment $(t = 2)$ for a given rank are

$$\frac{p_{1k}}{p_{2k}}, \quad \text{say} \quad k = 1 \quad \text{for high rank,} \quad k = 2 \quad \text{for low rank.}$$

Using the model in (9.1) to obtain these odds in logarithmic form (denoted hereafter as the log odds), we get

$$(9.2) \quad \ell n \frac{p_{1k}}{p_{2k}} = (\mu + \alpha_1^T + \alpha_k^K + \alpha_{1k}^{TK}) - (\mu + \alpha_2^T + \alpha_k^K + \alpha_{2k}^{TK}) = 2\alpha_1^T + 2\alpha_{1k}^{TK}$$

where $\alpha_1^T = -\alpha_2^T$ and $\alpha_{1k}^{TK} = -\alpha_{2k}^{TK}$.

Since the α parameters measure deviations from a general mean, a deviation from the mean at one level leads to a deviation in the opposite direction at the other level. Replacing $2\alpha_1^T$ and $2\alpha_{1k}^{TK}$ by β^T and β_k^{TK} to simplify the notation in (9.2) yields

$$(9.3) \quad \ell n \frac{p_{1k}}{p_{2k}} = \beta^T + \beta_k^{TK}, \quad k = 1 \quad \text{for high rank,} \quad k = 2 \quad \text{for low rank.}$$

From (9.3) the log odds of reenlistment to non-reenlistment are seen to depend on β^T, the general mean for the log odds, and β_k^{TK}, the relationship between rank and reenlistment decision.

To further illustrate these ideas, let us consider another example. Assume that reenlistment is dependent on two variables: length of enlistment, L, and the presence of absence of dependents, D. Then $p_{t\ell d}$ represents the probability that a specified reenlistment decision is made given an individual's length of enlistment and dependency status. Following the previous example, the logarithm of the odds of reenlisting to not reenlisting as a function of the predictor variables can be written as

$$\ell n \frac{p_{1\ell d}}{p_{2\ell d}} = \beta^{T} + \beta_{\ell}^{TL} + \beta_{d}^{TD} + \beta_{\ell d}^{TLD} \quad . \tag{9.4}$$

Each one of the β parameters has the same interpretation given previously. β^{T} is a general mean for the log odds. The β_{ℓ}^{TL}, $\ell = 1$ (two year enlistment), $\ell = 2$ (three year enlistment), $\ell = 3$ (enlistment of four or more years) are numerical measures of the impact on reenlistment of enlistment length. Similarly, the β_{d}^{TD} are numerical measures of the impact of dependents on reenlistment where the subscript d identifies the number of dependents, $d = 1$ (no dependents), $d = 2$ (one or more dependents). The parameters $\beta_{\ell d}^{TLD}$ are interaction terms. It may be, for example, that the presence of dependents may influence the reenlistment decision of four year enlistees differently than that of three or two year enlistees. First, dependents are more common among four year enlistees and they tend to have more of them. Second, four year enlistees who serve to end of term tend to be older at the time they must decide whether to reenlist. Hence the impetus to reenlist may be greater among members of this group than would be indicated by adding

the separate effects of dependency status and length of enlistment. The presence of a joint interaction effect of length of enlistment and dependency status on reenlistment implies a non-zero β_{32}^{TLD} .

By exponentiation of each side of the log-linear model (9.4), the odds of reenlisting to not reenlisting (hereafter referred to simply as the odds of reenlistment) can be written in the form

$$\frac{P_{1\ell d}}{P_{2\ell d}} = \delta^{T} \delta_{\ell}^{TL} \delta_{d}^{TD} \delta_{\ell d}^{TLD} \tag{9.5}$$

where the δ's are the anti-logs of the β's . In this form of the model, δ^{T} can be interpreted as the overall mean odds of reenlistment which is modified by more detailed information about the levels or values of the predictor variables and their interactions.

For the full model, the overall odds δ^{T} is estimated as

$$\hat{\delta}^{T} = e^{\hat{\beta}^{T}} = e^{-2.60} = .074 ,$$

that is, the odds are .074 to one in favor of reenlistment.* If the odds of reenlistment are desired for Marines who enlist for four years, we need to compute

$$\hat{\delta}^{T} \hat{\delta}_{3}^{TL} = (.074) (2.46) = .182 .$$

---

*Note that this is not the odds that would be computed directly from the observations, but rather from their logarithmic transforms, then averaging, then transforming back to the odds domain. Thus, this "mean odds" is a multiplicative mean, not an additive mean.

Thus, the odds of reenlistment increase from .074 to .182 for Marines who enlist for four years.

The calculation can be extended, for example, to Marines who enlist for four years who have one or more dependents by the end of their enlistment period. If these independent variables entered linearly in the model, the estimated odds for reenlistment would be given by $\hat{\delta}^{T}\ \hat{\delta}_{3}^{TL}\ \hat{\delta}_{2}^{TD}$, but since dependency status and length of enlistment are found to interact jointly on enlistment, the odds of enlistment for this group of individuals are given by

$$(9.6) \qquad \hat{\delta}^{T}\ \hat{\delta}_{3}^{TL}\ \hat{\delta}_{2}^{TD}\ \hat{\delta}_{32}^{TLD} = (.074)\ (2.46)\ (1.72)\ (1.46) = .457\ ,$$

where the last term measures the interaction effect of L and D. Note, the odds of reenlistment for four year enlistees with one or more dependents would have been substantially underestimated if the first order interaction effect had been omitted from the calculation.

As can be seen from this example, the estimation of a small number of δ's permits the computation of odds of reenlistment for individuals having very diverse characteristics. It should be noted that as in the case of regression analysis, the coefficients of the linear model (9.4) (and consequently the δ's in (9.6)) show the effect of a change in a variable holding all the other variables constant. Thus $\hat{\delta}_{\ell}^{TL}$ measures the direct effect of length of enlistment on the odds of reenlistment. If an indirect effect with dependency status is also present, this is measured by $\hat{\delta}_{\ell d}^{TLD}$. Both the direct and indirect effects of length of enlistment are net of the effects of other variables such as rank,

education, race, etc. That is, the effects of variation in the latter variables on the odds of reenlistment are taken into account in the computation of $\hat{\delta}^{TL}_{\ell}$ and $\hat{\delta}^{TLD}_{\ell d}$.

Given the odds of reenlistment for individuals with a given set of characteristics, it is a simple matter to compute the probability of reenlistment for the group from the relationship

$$(9.7) \quad \text{Odds of reenlistment} = \frac{\text{probability of reenlisting}}{\text{probability of not reenlisting}} \ .$$

For example, if the probability of reenlisting, p , is .07 , then the probability of not reenlisting, 1-p, is .93 , and the odds of reenlistment are .074 to one. Solving for p in (9.6) yields

$$(9.8) \quad \text{Probability of reenlisting} = \frac{\text{odds of reenlistment}}{1 + \text{odds of reenlistment}} \ .$$

In these calculations it is important to distinguish between individual δ's referred to as "odds factors" (e.g., δ^{TL}, δ^{TD}, δ^{TLD}) which indicate how the overall mean reenlistment odds, δ^{T} , is modified and the <u>product</u> of δ's (e.g., $\delta^{T}\,\delta^{TL}\,\delta^{TD}\,\delta^{TLD}$) which measures the odds of reenlistment for individuals with a specified set of characteristics. Since (9.7) converts the odds of reenlistment for a given group of individuals to the probability of reenlistment for that group, it cannot be applied to the individual δ's .

The above discussion makes clear that a large number of parameters may enter the contingency table model, thus raising the problem of identifying which parameters are to be included in a model and which are to

be excluded. Statistical distribution theory and a measure I^*, which is similar to R^2, the multiple correlation coefficient in regression analysis, is used to resolve this problem.

In regression analysis the explanatory value of a set of predictor variables is measured by the percentage of variation in the dependent variable explained by the predictor variables. The base measure of variation in regression analysis is the sum of squares about the mean of the dependent variable, i.e., $\Sigma(Y_i - \bar{Y})^2$. As predictor variables are added to the model, the predicted values of the dependent variable, $\hat{Y}_i$, are used to measure the amount of variation, $\Sigma(Y_i - \bar{Y})^2$, explained. The percentage of base variation explained is then

$$100\ R^2 = 100\ \frac{\Sigma(Y_i - \bar{Y})^2 - \Sigma(Y_i - \hat{Y}_i)^2}{\Sigma(Y_i - \bar{Y})^2} .$$

One method of measuring the contribution of any particular variable is the change in R^2 when that predictor variable is added to the model.

For contingency tables, the base measure of variation is computed either as the chi-square statistic*

$$\Sigma\ \frac{(O - E)^2}{E}$$

or the information measure

$$2\ \Sigma\ O\ \ln\ \frac{O}{E}$$

---

* The symbol O stands for the observed cell count and E the estimated cell count. The summation is over all cells in a table.

under the hypothesis that all β parameters in (9.4) except the general mean are zero. I^* is then the percentage of base variation explained by the introduction of some collection of β parameters into the model, i.e.,

$$I^* = \frac{\left(\Sigma\ 0\ \ell n\ \frac{0}{E}\right)_{\text{Base}} - \left(\Sigma\ 0\ \ell n\ \frac{0}{E}\right)_{\text{Model}}}{\left(\Sigma\ 0\ \ell n\ \frac{0}{E}\right)_{\text{Base}}} .$$

In practice, an I^* of 70 percent or better is desired. Sometimes a lower value is acceptable because increasing I^* requires the addition of many interaction parameters with the consequent difficulty of interpretation. The prime objective is to find the most important parameters. When the number of observations is large, parameters signifying marginal impact will be statistically significant. Thus we may adopt a convention, say, of excluding parameters when they increase I^* by less than two percentage points.

Bibliography

[1] Chernoff, H. (1973), "The Use of Faces to Represent Points in k-Dimensional Space Graphically," J. Amer. Statist. Assoc., 68, pp. 361-8.

[2] Fisher, R. A. (1936), "The Use of Multiple Measurements in Taxonomic Problems," Annals of Eugenics, Vol. VII, Pt. II, pp. 179-88.

[3] Fortier, J. J. and Solomon, H. (1966), "Clustering Procedures," Multivariate Analysis, pp. 493-506 (ed. Krishnaiah, P. R.), New York: Academic Press.

[4] Friedman, H. P. and Rubin, J. (1967), "On Some Invariant Criteria for Grouping Data," J. Amer. Statist. Assoc., 62, 1159-78.

[5] Galton, Francis (1888), "Co-relations and Their Measurements, Chiefly from Anthropometric Data," Proceedings of the Royal Society, Vol. 45, pp. 135-40.

[6] Goodman, L. A. (1971), "The Analysis of Multidimensional Contingency Tables: Stepwise Procedures and Direct Estimation Methods for Building Models for Multiple Classifications," Technometrics, 13, pp. 33-61.

[7] King, B. F. (1967), "Step-wise Clustering Procedures," J. Amer. Statist. Assoc., 62, pp. 86-101.

[8] Ku, H. H. and Kullback, S. (1968), "Interaction in Multidimensional Contingency Tables: An Information Theoretic Approach," Journal of Research of the National Bureau of Standards--Mathematical Sciences, 72B, pp. 159-99.

[9] Kullback, S., Kupperman, M., and Ku, H. H. (1962), "An Application of Information Theory to the Analysis of Contingency Tables, with a Table of 2n ln n, n = 1(1)10,000," Journal of Research of the National Bureau of Standards--B. Mathematics and Mathematical Physics, 66B, pp. 217-43.

[10] Pearson, Karl (1901), "On Lines and Planes of Closest Fit to Systems of Points in Space," The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science, Vol. 2 (6th Ser.), pp. 559-72.

[11] Solomon, H. (1971), "Cluster Analysis," Mathematics in the Archaeological and Historical Sciences, Edinburgh University Press, pp. 62-81.