Multivariate Analysis: Some Reminiscences on Its Origin and Development

Author(s): C. Radhakrishna Rao

Source: *Sankhyā: The Indian Journal of Statistics, Series B (1960-2002)*, Aug., 1983, Vol. 45, No. 2 (Aug., 1983), pp. 284-299

Published by: Indian Statistical Institute

Stable URL: https://www.jstor.org/stable/25052296

Sankhyā : The Indian Journal of Statistics
1983, Volume 45, Series B, Pt. 2, pp. 284-299.

MULTIVARIATE ANALYSIS*

Some reminiscences on its origin and development

By C. RADHAKRISHNA RAO
University of Pittsburgh
and
Indian Statistical Institute

SUMMARY. The article traces the history of multivariate analysis, the pioneering contributions of R. A. Fisher, the development of multivariate statistical methodology for applications in various fields under the guidance of Fisher, the work of the Indian School of Statisticians under the leadership of P. C. Mahalanobis, the author's experience in the statistical analyses of three large bodies of anthropometric data, and some of the modern trends of research. The origin and development of multivariate analysis clearly show that contact with live problems is essential for worthwhile research in statistical methodology.

It gives me great pleasure in thanking the Pfizer Corporation and the Statistics Department of the University of Connecticut, specially Professor Harry Posten, for kindly inviting me to speak at the Pfizer colloquium for the second time within a short period of time. Today, I would like to talk to you about Multivariate Analysis, the state of the craft as it was when I was a student and the directions in which it is progressing today.

The origin of the multivariate normal distribution can be traced to the writings of Gauss, Bravais, Shols, Galton and Edgeworth during the 19th century. But the key figure whose quest for knowledge on laws of heredity triggered off research on the theory and applications of the multivariate normal distribution is Francis Galton. In a lecture delivered at the Royal Anthropological Institute in 1885**, Galton presented his analysis of data on heights of parents and adult children in the form of a bivariate frequency chart as in Figure 1.

He made the following observations :

(i) The locus of verticals, i.e., the graph joining the conditional mean heights of children given parents' height, is nearly a straight line.

*T.V. talk recorded at the University of Connecticut, Storrs, Connecticut on November 20, 1981 for the archives of the American Statistical Association. The lecture was delivered under the auspices of the Pfizer Colloquim in Statistics.

*This work is sponsored by the Air Force Office of Scientific Research under Contract F49629-82-K-001. Reproduction in whole or in part is permitted for any purpose of the United States Government.

**The lecture was given on September 10, which coincides with my birthday.

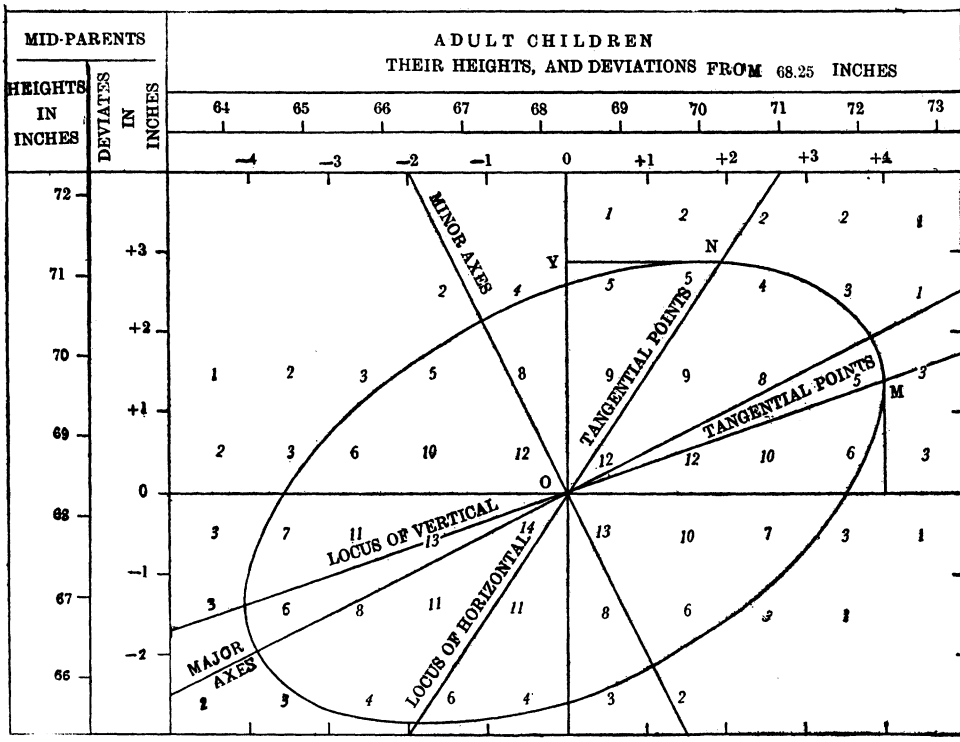

Fig. 1. Diagram based on Table 1. (all female heights are multiplied by 1·08)

(ii) The scatter of points is homoscedastic, i.e., the conditional variance of children's heights given the parental height is the same for all parental heights.

(iii) The equiprobable contours of the bivariate distribution are elliptical.

Having noted these three properties, viz., linearity of regression, homoscedasticity and elliptic form of the equiprobability contours, Galton could have written down the expression for the underlying distribution as bivariate normal. It is surprising that he did not. However, he mentioned these properties to the Cambridge Mathematician J. D. Hamilton Dickson who immediately gave the answer. Once the expression for the bivariate normal probability distribution was known, the generalization to the multivariate case was only a simple step.

Galton also found some interesting features of the height data which led to the development of the modern regression theory. From the locus of

verticals, as shown in Figure 1, he inferred the phenomenon of "reversion" which was later referred to as "regression", viz., the tendency of the conditional means at the extreme values of parental heights "to depart from the parental type and revert to the general mean". He introduced a numerical measure r to quantify the magnitude of reversion. This r is the source of our symbol for the correlation coefficient, which according to Karl Pearson was really the fisrt letter of reversion.

Galton's work and its relation to the bivariate normal distribution won three recruits for the field of correlation, Weldon, Edgeworth and Karl Pearson. Galton himself called his measure r, 'the index of co-relation' which was changed to the more weighty word 'coefficient of correlation' by Edgeworth, while Weldon who computed the values of r for a number of bivariate samples called it the 'Galton function'. Weldon and Edgeworth did considerable spade work which was followed up by Karl Pearson and Sheppard (of Sheppard's correction for the moments). They worked out the expressions for partial, multiple and total correlations as we know them today, and also provided the large sample standard error of r. All this work was done before the end of the last century.

It was recognized that the distribution of r was skew and, for purposes of inference in small and even moderately large samples, its exact distribution was necessary. All efforts to find a breakthrough failed until 1915 when R. A. Fisher emerged on the scene and gave the exact sampling distribution of r. This famous paper marks the beginnings of research in multivariate analysis as a statistical technique for drawing inferences from multivariate data. The form in which the exact distribution of r was expressed was somewhat complicated, but Fisher soon found a simple transformation, $r = \tanh z$ known as Fisher's z-transformation, which considerably simplified the sampling distribution and the inference procedures based on the observed value of r. In the early twenties, Fisher introduced the F distribution which provided tests of null hypotheses for regression coefficients, partial and multiple correlation coefficients. During this period, Fisher also introduced the maximum likelihood method of estimation and developed the modern technique of the design and analysis of experiments. *Statistical Methods for Research Workers* by Fisher, first published in 1925, was essentially a practical handbook on these new methods. This book opened up the doors to new knowledge through appropriate analyses of data, and soon became popular, specially with biologists who were accumulating masses of data. Perhaps, the main drawback of the book from the modern point of view is its heavy emphasis

on tests of null hypotheses and the use of the 5% and 1% levels of significance. I must add that Fisher himself did not use these levels of significance religiously in his own work. But in his wisdom, he must have thought that some safe prescription for interpreting an observed value of a test criterion might be of help to research workers in applied fields, specially when tables of the entire cumulative distribution could not be made available.

Fisher's 1925 book attracted the attention of research workers from all over the world. The next ten years saw some major advances in multivariate analysis mostly under his guidance. In 1928 Wishart extended Fisher's work on the bivariate distribution to the joint distribution of estimated variances and covariances in samples from a general p-variate normal population. Hotelling introduced his T^2 statistic in 1931 as a generalization of Student's t to the multivariate case. He also developed the concepts of principal components in 1933 and canonical correlations in 1936. Wilks started in 1932 his work on likelihood ratio criteria for testing various hypotheses concerning the mean vectors and co-variance matrices of multivariate normal distributions. At the suggestion of Fisher, Martin constructed in 1936 the linear discriminant function for deciding whether a jaw bone found in a grave belonged to a male or a female. Fairfield Smith worked out in the same year the discriminant function for plant selection, which is the forerunner of what is now known as empirical Bayes estimation (cf. Rao, 1953).

While Hotelling and Wilks in the U.S.A. were interested in tests of significance, Mahalanobis in India was concerned with the problem of studying interrelationships among a given set of populations. For this purpose he devised a measure of distance between two populations known as Mahalanobis' D^2 which was an improvement over Pearson's C^2, the coefficient of racial likeness. In 1938, Bose and Roy found the non-null distribution of Mahalanobis D^2, which has the same form as that of the multiple correlation coefficient derived 10 years earlier by Fisher. The analogy was not clear until Bowker found in 1966 an elegant representation of D^2 as the ratio of two independent chi-square variables, one of which is noncentral.

Fisher used geometrical arguments in deriving the distribution of the multiple correlation coefficient and the proof was difficult to understand as many of the intermediate steps in the argument were missing as is usual with Fisher's writings. Mahalanobis in his biographical sketch of Fisher mentions that "he does not attempt to write down the analysis until the problem is solved in his mind, and sometimes, he confesses, after the key to the solution has been forgotten".

Fisher generalized the discriminant function approach to more than two populations which led to the sampling theory of what are called the roots of determinantal equations, which are same as the eigen values of one random matrix with respect to another independent random matrix, both having a Wishart distribution. Fisher found the exact null distribution of these roots in 1939, which is his last major work in deriving sampling distributions. The distribution has the same form in the null case as that of the test criteria introduced by Roy in the same year for testing equality of two variance-covariance matrices.

These were the main lines of development in multivariate analysis up to 1940. There were other types of studies involving multiple measurements such as factor analysis, with fancy titles such as "vectors of the mind", which originated with Spearman in the beginning of the century but did not receive much attention from the statisticians. This might be partly due to the difficulties involved in estimating the large number of unknown parameters in a factor analytic model and partly to the subjective interpretation and doubtful utility of the results arising out of factor analysis. There were no references to any existing factor analytic methods in Fisher's own writings, although he postulated a factor analytic type model in explaining the variation of human measurements in his 1918 classical paper on correlation between relatives. Unlike the factor analysts, Fisher was cautious in making the distinction between the 'number of factors' and the 'effective number of factors' and thought that the latter, though abstract had a conceptual meaning and might be estimable.

In 1940, I graduated in mathematics with statistics as a special subject. The job opportunities for a mathematician were poor during the war time. There were not enough scholarships for pursuing a research career in mathematics. After remaining unemployed for about six months I had to choose between joining the British army or go in for an alternative course of study which provided better prospects. I had heard about the Indian Statistical Institute which was established by Mahalanobis a few years earlier. Mahalanobis was at that time trying to lure all intelligent students to study the new subject of statistics promising them an attractive career. I felt extremely happy when, in response to my application, he admitted me to what was called the training section of the Institute to be trained as a statistician. This was in the beginning of the year 1941 which turned out to be the beginning of my career in statistics and my long association with the Indian Statistical Institute, which continues even today.

Soon after joining the Institute I met Bose, Roy and Mahalanobis. Bose was working on combinatorial problems in Design of Experiments, Roy was continuing his work on the roots of the determinantal equation for testing equality of dispersion matrices and Mahalanobis was developing the theory and methodology of large scale sample surveys. As a part of my training in Statistics, Mahalanobis, who always emphasized the difference between education and training, assigned to me an anthropometric project. I was to be in charge of the analysis of about 12 morphological measurements made on a large number of individuals classified by about 21 caste groups and tribes. In this study, I was supposed to use Mahalanobis D^2 as the main tool, which meant computing a large order covariance matrix, its inverse and a number of quadratic forms. This was an impossible task at that time as we had only the mechanical desk calculators. My first problem was to devise methods for reducing the number of measurements to simplify the computations. This needed a criterion for the selection of variables (cf. Mahalanobis, Mazumdar and Rao, 1949).

As a first step in this process I derived a test criterion called V_r for testing the equality of mean vectors in several populations. V_r is the weighted average of all possible D^2 values based on r specified measurements for different pairs of populations. The distribution of V_r, when the variance-covariance matrix is known or estimated on a large number of degrees of freedom, is reported in my Master's degree thesis in 1943 and later published as a note in 1945*. The statistic V_r which I called the perimeter test is now known as the trace criterion being the sum of the roots of the determinantal equation of Fisher. I used the same criterion by taking a simple average of all possible D^2 values for the selection of r best variables by maximizing V_r over all possible selections of r out of a larger set of p measurements. This was still a difficult problem for the desk calculator and some short cuts had to be made instead of computing all possible V_r values. I am glad to see that this procedure is now available in a computer program for the selection of variables in discriminant analysis.

I wanted to make sure that not much information is lost by reducing the number of measurements. To examine this, I derived a criterion for testing the significance of the extra information gained by augmenting a specified set by an additional set of measurements. This criterion which was described in a paper published in *Sankhyā* in 1946 (Rao, 1946) also provided a

* Page 348, Volume 6 of *Sankhyā* (where the criterion is identified as the sum of the roots of Fisher's determinantal equation).

generalization of the Fisher-Bartlett test for an assigned discriminant function. For instance, we can test whether the coefficient vector of the true discriminant function is in any specified subspace.

To avoid the inversion of a large order covariance matrix, I used a step wise transformation of the correlated variables into an uncorrelated set and computed D^2 as a sum of squares. The transformation was carried out in a triangular scheme listing the measurements in order of their importance in discrimination. Later, I discovered that the method of transformation is the well known Gram-Schmidt orthogonalization process and the method of computation employed is numerically a stable one for which computer programs exist.

Following the work of Fisher and Roy on the roots of determinantal equations, I introduced in a paper published in 1945 (Rao, 1945) what are called familial correlations, which are generalized intraclass correlation coefficients obtained by considering linear combinations of measurements. With multiple measurements on each of k brothers or sisters in several families, the familial correlations are functions of the eigen values of between to within family covariance matrices, as in Fisher's problem. They have the same definition and null distribution as Fisher's statistics but their non null distribution is similar to that of Roy's statistics. They also provide a specialization of Hotelling's canonical correlations when we use the same linear combinations for the two sets of measurements (as on brothers).

In the course of my two years of work on anthropometric data at Calcutta, I was able to establish credentials of my expertise in multivariate analysis. This, I realised when Mahalanobis decided to send me to Cambridge, England, in response to a request from J. C. Trevor, an anthropologist at Cambridge, to depute some one to apply Mahalanobis D^2 on some skeletal data collected by a British expedition from Jebel Moya in Africa.

I arrived in Cambridge in the fall of 1946 and joined the Duckworth Laboratory which housed the anthropological museum. Although I was theoretically oriented and would have preferred to work in a cleaner place and not in the midst of bones and stones, I took my new assignment as a challenge and settled down to work, this time with an electric desk calculator as my companion.

My visit to Cambridge was profitable in many ways. Although Fisher had earlier written a paper criticising Pearson's C^2, the coefficient of racial likeness, and denouncing all research work on skeletal material as futile, he agreed to be my supervisor for a possible Ph.D. thesis arising out of my work.

He was happy to see some extensions of his work I was making on discriminant functions when more than two populations were involved, but thought that it would do me good to spend some time in his mice laboratory where work was being done on the mapping of chromosomes. So I ended up working at both places, the Duckworth Laboratory where I was handling bones of dead people and Fisher's laboratory at Whittingehame Lodge where I was experimenting with living mice, mating different types and studying the progeny.*

In the case of the African skeletal material, all the measurements were made by the British anthropologists at the site of the excavation and only a small portion of the bones, those that were well preserved, was shipped to the Museum at Cambridge. By the time I reached Cambridge, the measurements were already processed on IBM machines. Means and standard deviations were computed for each character after classifying the bones by three levels of depth from which they were excavated.

While working on the anthropometric material in Calcutta, I became deeply aware of the existence of recording errors, investigator differences and outliers in any large body of data collected under field conditions. I, therefore, decided to set aside the nice IBM printouts and look at the original schedules for possible errors. Scrutiny of data or, to use a more appropriate expression due to Fisher, "cross examination of data" is not a routine type of work. A statistician has to look for gross irregularities in the data due to presence of outliers and errors of various kinds and also noticeable regularities due to faking. Thus it may be necessary to subject the data to what may be called "irregularity analysis" to detect errors and "regularity analysis" to detect faking. An example of the latter kind is the analysis of Mendel's data on sweet pea by Fisher which indicated that Mendel's data were possibly adjusted to agree closely with expected values.

In scrutinizing multivariate data, it always pays to examine the distributions of individual measurements and ratios, plot the histograms, compute the first four moments and measures of skewness and kurtosis, and examine the changes in these descriptive constants when extreme values are omitted. The table (Figure 2) provides an example of such a procedure, where the figures with an asterisk refer to high values of skewness and kurtosis computed from the original data. The recomputed values, after omitting outliers, given in the second line are in conformity with the others.

* Fisher assigned to me the task of breeding a quadraple recessive for the characters, undulated tail, shaky, skin color and wellhaarig for linkage studies, which I was able to do with a minimum number of matings.

Fig. 2

TABLE : TEST STATISTICS γ_1 FOR SKEWNESS AND γ_2 FOR KURTOSIS FOR SOME ANTHROPOMETRIC MEASUREMENTS OF SIX MALE TRIBAL POPULATIONS
(From the Thesis of Dr. Urmila Pringle)

	male tribal populations									
Character	KOLAM		KOYA		MANNE		MARIA		RAJ GOND	
	γ_1	γ_2	γ_1	γ_2	γ_1	γ_2	γ_1	γ_2	γ_1	γ_2
H.B.	.15	−.62	.39	.37	1.62* .71*	4.54* .29	−.27	.48	−.30	.23
H.L.	−.14	−.06	.48	1.12	−.05	−.08	.05	−.09	−.32	.28
Bg.B.	.83* −.14	2.93* −.03	.17	.19	1.72* −.40	8.42* .27	−.17	−.63	−.12	−.61
T.F.L.	−.26	−.07	.44	.11	.66*	.32	−.05	−.10	−.04	−.24
U.A.L.	−.05	−.63	−1.95* −.30	6.88* .74	−.01	−.27	.13	.76	.14	−.40
L.A.L.	−2.17* .08	9.98* −.62	−.07	.59	.19	−.67	−.02	.28	−.06	−.67
L.A.L./U.A.L.	−2.10*	11.25*	3.63*	15.62*	.11	−.43	−1.19	4.44*	−.38	1.13

The values in the second line for each character are calculated after omitting extreme observations.

Plotting of principal components (Rao, 1964) or simple bivariate frequency charts can also be of great help in detecting errors and outliers. Such examination of the original data on the African skeletal material led to discarding the whole of the field data and working only with the measurements made at the Duckworth Laboratory on the small sample of material sent from the excavation sites (Rao *et al.*, 1955).

Another difficulty faced was a high degree of incompleteness of the data. Most of the skulls were broken and not all measurements could be taken on every skull. For instance, cranial capacity could be measured only when the whole skull was intact. How does one estimate the mean values and variances and covariances when data are incomplete ? I am surprised to see some recent papers on the subject where maximum likelihood estimates and likelihood ratio criteria are recommended on the assumption that the missing values occur in a random manner. This is not generally true in practical problems. For instance, the skulls that were well preserved had on the whole smaller measurements than those that were broken, which showed that the well-preserved skulls were on the whole smaller in size and therefore cannot be considered as a representative sample from the original population of skulls. Thus pushing all the data through the likelihood mill might give

wrong results. Suitable adjustments had to be made in arriving at plausible estimates of the parameters characterizing the original population of skulls.

When I found this phenomenon of differential preservation of skulls in the graves, I thought it was interesting. So I took the first opportunity of talking about it at a statistical conference. I showed by statistical analysis of several series of skeletal data that the chance of a small skull being preserved is greater than that for a large skull (Rao and Shaw, 1948). At the end of my talk, the chairman of the meeting wanted to comment. He said he was taking a guided tour of a Museum in Paris along with some tourists. The guide stopped near a glass case and said 'The well preserved skull you see in the glass case was Ceasar's'. Someone among the tourists who obviously read about Rao's discovery exlaimed, 'Ceaser had such a small skull. It is no wonder that it was well preserved'. Hearing this, the guide said, 'Actually this was Ceasar's skull when he was sixteen'.

The next problem was the use and interpretation of the computed D^2 values between the groups under comparison. Cluster analysis algorithms were not developed at that time. However, one could attempt to roughly indicate the relative positions of the various groups on a two-dimensional chart using the D^2 values. I tried to make such a representation more objective through what I have called canonical variates (Rao *et al.*, 1949) which are ordered linear combinations $l_1, l_2, \ldots$ of the measurements such that the first k of them, $l_1, \ldots, l_k$, provide the best representation of the groups in a k dimensional subspace of the space of all measurements. These linear combinations are computed using the eigen vectors of the variance covariance matrix between groups relative to within as shown in a paper published in the Journal of the Royal Statistical Society in 1948 (Rao, 1948b). Figure 3 is an example of such a representation based on the first two canonical variates. The names of the groups given in the figure are those of some castes in one of the states of India. In practice, one could make a three dimensional model using the first three canonical variates or a series of three dimensional models from a selected set of canonical variates. It should always be borne in mind that there will be some distortion in representing higher dimensional data in a space of lower dimensions. To guard against this it is necessary to compute the amount of distortion caused in each distance between groups and use this knowledge in interpreting the configuration of points in the three dimensional models (Rao, 1971).

Following the work of Torgerson on multidimensional scaling, I suggested in a paper published in 1964 (Rao, 1964) a method of computing canonical

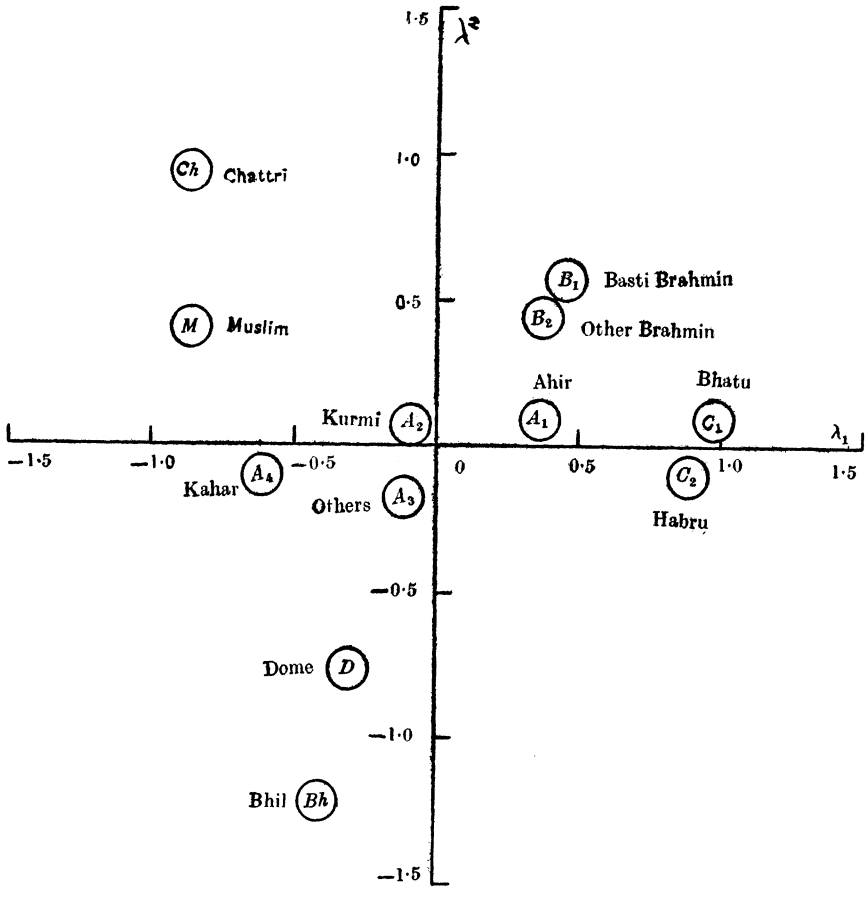

Fig. 3. U.P. anthropometric survey group constellations in the $(\lambda_1 - \lambda_2)$ chart.

variates directly from a table of D^2 values or any dissimilarity measures estimating the underlying distances. I am glad to see that such graphical representations through canonical variates (or coordinates) are now widely used although the caution I have indicated regarding the distortion of distances is usually not considered. I prefer the nemonic CANCOORD for my canonical coordinates, which sounds less guilty than CRIMCOORD suggested by Gnanadesikan.

Since the fifties many methods of cluster analysis have been developed and there are several books describing various algorithms for cluster analysis. However, I still believe in what I have stated in an early publication that no specific rules can be laid down for forming clusters. Different methods may have to be used in different situations. For instance, it is possible that a number of groups may form a chain with each group closely linked to two

others. In such a case, forcing the groups into distinct clusters or a tree pattern may lead to wrong interpretations. I have advocated listing of clusters of groups, which may be overlapping, such that the groups within a cluster have dissimilarity coefficients less than some chosen value. Such a listing can be pictorially represented as a graph with sub-graphs (shown in Figure 4) where the vertices represent the groups and all vertices with distances less than a chosen threshold value are connected. Such a figure has more information than a dendrogram (Rao *et al.*, 1958). For instance, the central position of the group represented by *My* in Figure 4 could not be inferred through a dendrogram, where it would have been classified with some cluster at some level of dissimilarity. This led to an interpretation of the relationship of *My* with others consistent with historical facts.

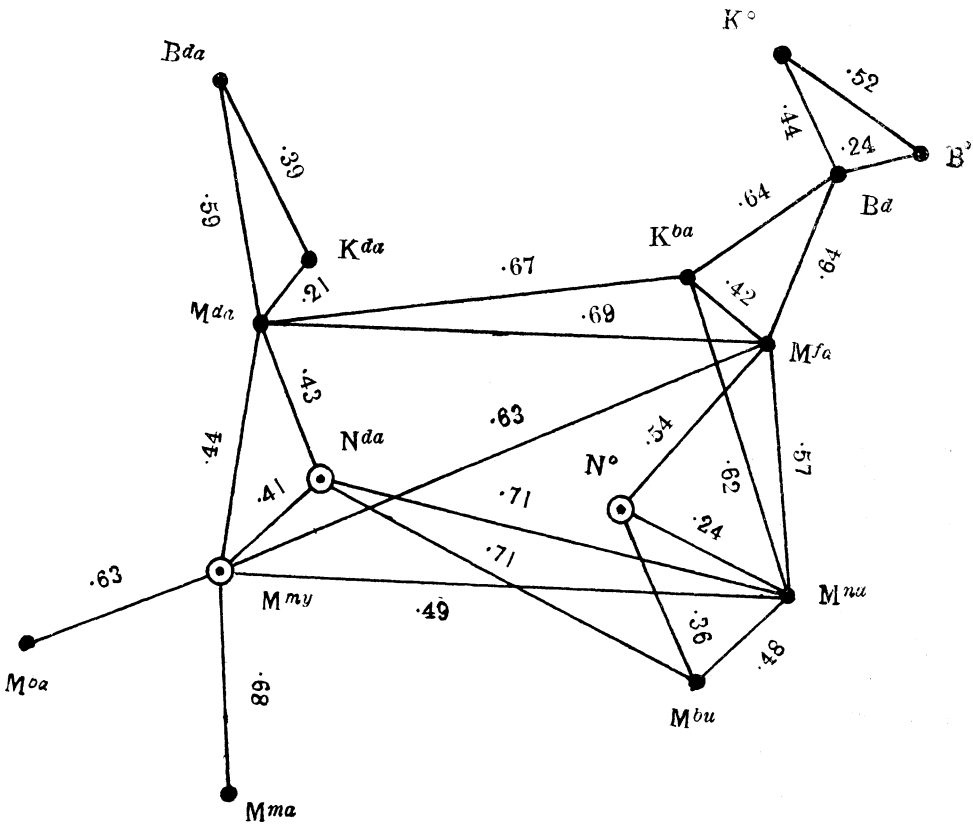

Fig. 4. Graph of groups and maximal subgraphs.

Bartlett was making valuable contributions to multivariate analysis during and immediately after the second world war. In 1947, Bartlett read

a paper at a meeting of the Royal Statistical Society summarizing his work. I had the opportunity to attend this meeting and participate in the discussion. In his paper, Bartlett derived tests of some hypotheses by factorizing Wilk's overall lambda criterion, about which I had some doubts. The resulting test procedures seemed to underestimate significance. I made some remarks at the meeting based on the work I had been doing at that time. My paper published in Biometrika in 1948 (Rao, 1948) clarified some points in Bartlett's paper and also provided a general framework, through what I called analysis of dispersion, a term approved by Fisher, but later christened as MANOVA in the U.S., for testing specified hypotheses such as main effects, interactions, effect of regression, deviation from regression, effectiveness of concomitant variables and additional information contained in some measurements in problems of discrimination.

During my student days, Fisher's Statistical Methods was the only advanced book on statistics, but it was neither prescribed as a text nor recommended for general reading. It is not an easy book to read for those who want to understand the basis of the statistical methodology as the proofs and the assumptions under which the results are derived were not clearly set out. I have even heard remarks such as, "you should not attempt to read it unless you have read it before". However, I found it stimulating and wrestling through it was, indeed, a rewarding experience. I had even organized a study circle to read every page of Fisher's book and discuss its theoretical and practical contents. It occurred to me that it was worthwhile to write a commentary on Fisher's book supplying the missing theory and mathematics and expanding on what I thought was its *tour de force*, analysis and interpretation of small self-contained sets of data. My work in Calcutta and Cambridge on anthropometric data provided excellent material to review Fisher's methods and also discuss the multivariate statistical methodology not covered in Fisher's book. I was also looking for an opportunity to put together the contributions of the Indian School of Statisticians to multivariate analysis for wider dissemination. I started writing the book while I was still in Cambridge. I could not decide on the title for a long time. The book had enough theory and mathematics to be called, "Mathematical Statistics" which would have probably attracted wider attention. But, being somewhat mission oriented at the younger age, I wanted to emphasize the practical aspects of statistics, and so decided to call it "Advanced Statistical Methods in Biometric Research".

I returned to Calcutta with a Ph.D. degree from Cambridge and a first draft of "Advanced Statistical Methods in Biometric Research", and with

the reinforced conviction that contact with live problems is essential for worthwhile research in statistical methodology.

The course of research in multivariate analysis followed in two different directions during the last 25 years. In one direction, there has been vigorous work by Anderson, James, Pillai and others in determining the exact distributions of roots of determinantal equations in the non-null case, devising various test criteria based on them and constructing tables of percentage points for carrying out tests of significance. In multivariate analysis, one is often led to consider several hypotheses or estimate several parameters simultaneously. Appropriate inference procedures necessary in such cases are being worked out under the assumptions of normality by Krishnaiah and his colleagues.

In another direction, breaking away from dependence on the multivariate normal distribution and probability considerations, Kendall, Kruskal and their colleagues are solving interesting problems such as seriation of archaeological material, constructing evolutionary tree structures in historical and other contexts and multidimensional scaling. Cavalli-Sfroza and Nei are estimating tree structures of biological evolution using genetic models.

Research in multivariate analysis is still in a growing stage. There are numerous problems for which we have no answers or which are not satisfactorily solved for practical applications. I am reminded of what Finney once said, that research in multivariate analysis has served only as outlets for the mathematical skills of authors without any practical use, and the skepticism expressed by Tukey that we go multivariate when we do not understand the situation. I hope this situation will change if future research workers address themselves to the solution of problems arising out of practical applications.

I must also refer to the uncritical use of computer package programs. Unfortunately, the tendency has been to let a computer program decide the questions to be raised and solve them in a particular way. Take for instance, the problem of deciding whether *Australopethicus africanus*, a fossil tooth discovered in Africa, is hominid or anthropoid (Rao, 1973, p. 579). The appropriate computer program appeared to be discriminant analysis and the computer printout gave the categorical answer that it is more likely to be a hominid than anthropoid. In fact, it can be seen from Figure 5 that the observed fossil does not fit in with either hominid or anthropoid samples. The possibility that a specimen to be classified may not belong to any of the specified groups was not considered. This led to the computation of Mahalanobis D^2 values of the observed fossil not from its actual position but from

its projection P in the direction of the discriminant function. Thus we find from Figure 5 that its D^2 value from the Human is 0·3 which is smaller than the D^2 value 3·8 from the Chimpanzee. This led to the wrong conclusion.

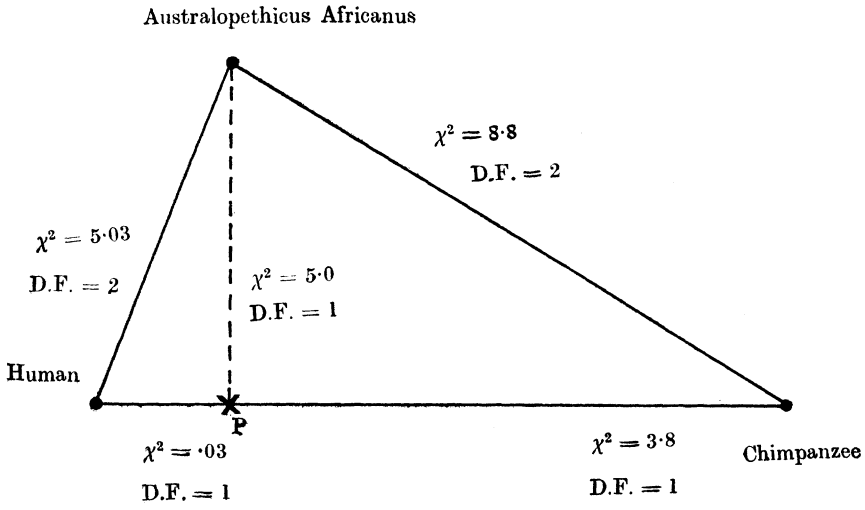

Fig. 5. Is A. Africanus more homonid than ape like ?

There cannot be tailor-made computer programs for every problem, but the ready-made loose fits can be of great help in determining the shape of the final analysis.

I must also refer to graphical methods which are being developed to provide a better comprehension of the underlying structure of multiple measurements, to detect outliers and to help in the interpretation of statistical results. Reference may be made to a recent book on this subject by Gnanadesikan and the pictorial representation of multivariate data suggested by Chernoff.

I thank you for your patience in listening to my remininscences and personal comments on certain aspects of the developments in multivariate analysis.

References

BARTLETT, M. S. (1947): Multivariate analysis. *J. Roy. Statist. Soc.*, B9, 176-197.

MAHALANOBIS, P. C., MAJUMDAR, D. N. and RAO, C. R. (1949): Anthropometric survey of the United Provinces : A statistical study. *Sankhyā*, **9**, 90-324.

MAJUMDAR, D. N. and RAO, C. R. (1958): Bengal Anthropometric Survey, 1945 : A statistical study. *Sankhyā*, **19**, 201-408.

MUKHERJEE R. K., TREVOR, J. C. and RAO, C. R. (1955): *The Ancient Inhabitants of Jebel Moya*, Cambridge University Press.

RAO, C. R. (1945): Familial correlations or the multivariate generalization of the intraclass correlation. *Current Science*, **14**, 66.

——— (1946): Tests with discriminant functions in multivariate analysis. *Sankhyā*, **7**, 407-414.

——— (1948a): Tests of significance in multivariate analysis. *Biometrika*, **35**, 58-79.

——— (1948b): The utilization of multiple measurements in problems of biological classification. *J. Roy. Statist. Soc.*, **B10**, 159-203.

——— (1953): Discriminant function for genetic differentiation and selection, *Sankhyā*, **12**, 229-246.

——— (1964): The use and interpretation of principal components analysis in applied research. *Sankhyā*, **26A**, 329-358.

——— (1971): Taxonomy in anthropology, in *Mathematics in Archaelogical and Historical Sciences*, 19-29, Edin. Univ. Press.

——— (1973): *Linear Statistical Inference and its Applications*, Second edition, John Wiley.

RAO, C. R. and SHAW, D. C. (1948): On a formula for the prediction of cranial capacity. *Biometrics*, **4**, 247-253.

Paper received : *February*, 1983.