# PROBLEMS AND PITFALLS IN COMPUTER STATISTICS

R. A. Reyment

(Seminar held at The University of Kansas,
Geological Survey of Kansas, January, 1967.)

## TABLE OF CONTENTS

# PROBLEMS AND PITFALLS IN COMPUTER STATISTICS

## INTRODUCTION

The object of this course is to present a number of fairly well known statistical techniques, presently rather generally available in computer libraries, and to examine them critically. In particular, it is proposed to discuss statistical techniques used in geology and paleontology. This is certainly not a new approach per se, but the present analysis attempts to outline possible pitfalls, and in most cases suggestions are made with respect to solving difficulties which arise.

It is reasonably well know, that the main body of statistical theory based on the properties of the univariate and multivariate normal distributions is, for practical applications, often a somewhat idealized representation of a situation. In many problems, data will approximate to the normal distribution. However, one of the sciences which deviates rather frequently is, geology. By the intrinsic nature of these data, the way in which they were and are collected and, for example, the complications occasioned by geological processes, data are frequently of such a kind as to deviate significantly from the normal distribution. As a concrete example hereof may be mentioned the effects of sorting on fossils. In my analysis of Russian Cretaceous belemnites with Dimitri Naidin this kind of disturbance factor was found to be of very considerable importance (Reyment and Naidin, 1962).

We have just reviewed the first kind of disturbance in the normal course of study of geologic materials. A second, equally as important, problem is posed by the following: Statistical techniques are often devised in order to solve a particular technique to solve other problems that may or may not be related to the original model. When the application is made by a competent mathematical statistician, the chances of going astray may not be overly great. Where, however,

the application is made by a nonstatistician, the pitfalls are oft-times fearful. I have, in particular, one statistical method in mind, that of factor analysis. This was originally conceived and developed by nonmathematicians, although the basic model of what is now called principal component analysis (often confused with factor analysis) is to be found in the writings of Karl Pearson. The factor analysis model thus arose nonmathematically. Subsequently, mathematicians have been called in to bolster up the original concept with acceptable theory. This has, I feel, led to a patchwork quilt model; it is especially padded to fit a particular set of concepts occurring in psychology, and closely related topics, hence, its application to nonpsychologic problems requires a great amount of forethought and care. Interpretation of the results of factor analysis is open to considerable subjectivity.

However, as will become apparent further on, I do not advocate the annihilation of Factor Analysis in relation to geostatistical problems, but rather am sure that it may be useful as a techbique for offering an extra way of regarding a set of data. The moral of the story is thus that care and afterthought should be adequately exercised when this method is employed.

# CHAPTER II

## THE COVARIANCE MATRIX HOMOGENETITY PROBLEM

If we are required to perform tests involving two samples from two populations in which the variances, in the univariate situation, play a part, it is necessary to take cognizance of the homogeneity of these variances. Whether or not the variances are homogeneous is very often of considerable importance in many statistical computations. The question is not only one of importance in univariate statistical analysis but also a vital one in multivariate statistical analysis.

In this section $\Sigma_i$ denotes a covariance matrix for the ith population, $\pi_i$, with population mean vector $\mu_i$. The populations will be taken to be multivariate normally distributed. The corresponding sample quantities to the foregoing will be written $S_i$, $\bar{x}_i$.

Although the main discussions here will be concerned with multivariate statistics, for computer applications, the discussion will be initiated via the appropriate univariate situation where enlightening.

(1) Statement of the homogeneity problem:

Consider k, p-variate multivariate normal distributions each of which has been sampled. It is desired to test the hypothesis that the covariance matrices of these populations are equal. Let $x_\alpha^g (\alpha = 1, N_g;\ g = 1, k)$ be an observation form the gth population $N(\mu^{(g)}, \Sigma_g)$. The hypothesis desired to be tested is:

$$H_1 : \Sigma_1 = \ldots = \Sigma_k . \qquad (II:1)$$

That is, we wish to test the equality of the k matrices. We shall first pause to see how this is done for two variances in the univariate case:

Here, one computes the variance ratio for $s_1^2$ and $s_2^2$, the usual unbiased estimates

of $\sigma_1^2$ and $\sigma_2^2$ (the two population variances), which is given by:

$$F = s_1^2/s_2^2 \qquad (II:2)$$

This has degrees of freedom $n_1$ and $n_2$ .

For more than two populations, M.S. Bartlett developed a special test. This test is, unfortunately, not robust, which means that it is sensitive to departures from the normal distribution and certain other kinds of "nonstandardness". The univariate test for homogeneity of variances is as follows:

(Here for two samples)

$$\frac{(n_1)^{1/2\, n_1}(n_2)^{1/2\, n_2} F^{1/2\, n_1}}{(n_1 F + n_2)^{1/2(n_1+n_2)}}, \qquad (II:3)$$

where F is as in (II:2).

Many assumptions of the analysis of variance (ANOVA) include the prerequisite of equal variances. An approximate test for variance equality, suggested by Scheffe, has been put forward in order to bypass the sensitivity of the Bartlett test. This test is based on the analysis of variance of the logarithms of the sample variances, thus transforming the problem to one of the comparison of means. This is useful, as the analysis of variance is fairly <u>insensitive</u> to the shape of the distributions of the estimated means:

Let $s^2$ denote the sample variance of a random sample of size $\underline{n}$ , drawn from a population with variance $\sigma^2$:

$$s^2 = \sum_{i=1}^{n} (x_i - x.)^2/(n-1), \qquad (II:4)$$

and thus, $E(s^2) = \sigma^2$.

And,

$$Var(s^2) = \sigma^4(2/(n-1) + \lambda_2/n), \qquad (II:5)$$

where, $\lambda_2$ is a measure of kurtosis. This is defined as

$$\lambda_2 = \sigma^{-4}\mu 4 - 3.$$

Here, $\mu_4$ is the fourth central moment of the population. For a normally distributed population, $\lambda_2 = 0$.

Let $y = \log_e s^2$, then,

$$E(y) \approx \log\sigma^2,$$

and,

$$\text{Var}(y) \approx 2/(n-1) + \lambda_2/n. \qquad (11:6)$$

Consider now I sets. Assume that populations falling in the same set have the same variance. If there are J populations in the ith set and $s^2_{ij}$ is the sample variance for the sample, from the jth population in the ith set, then under the fundamental assumption,

$$E(s2_{ij}) = \sigma_i^2 \quad (J = 1, 2, \ldots, J_i).$$

One requires to test the hypothesis:

$$H: \sigma_1^2 = \sigma_2^2 = \ldots = \sigma_I^2 .$$

Let

$$y_{ij} = \log_e s^2_{ij} .$$

Then under the fundamental assumption,

$$E(y_{ij}) \approx \log_e \sigma_i^2 ,$$

and,

$$\text{Var}(y_{ij}) \approx 2/(n_{ij}-1) + \lambda_{2,\,ij}/n_{ij} . \qquad (11:7)$$

Here, $n_{ij}$ is the sample size for $s^2_{ij}$ and $\lambda_{2,\,ij}$ denotes the corresponding kurtosis population measure.

The above formulated hypothesis is equivalent to

$$H: \log\sigma_1^2 = \ldots = \log\sigma_I^2 .$$

If the sample sizes, $n_{ij}$ , are all equal, then the $y_{ij}$ all have approximately the same variance. The variance ratio statistic for testing the hypothesis for unequal sample sizes is:

$$F = \frac{\sum_i (J_i - 1) \cdot \sum_i \sum_j (n_{ij} - 1) [\nu_i - \nu]^2}{(I - 1) \sum_i \sum_j (n_{ij} - 1)[y_{ij} - \nu_i]^2} \qquad (II:8)$$

with $I - 1$ and $J_i - 1$ degrees of freedom. In the above equation,

$$\nu_i = \sum_j (n_{ij} - 1) y_{ij} / \sum (n_{ij} - 1) \text{ and,}$$

$$\nu = \sum_i \sum_j (n_{ij} - 1) (\nu) / \sum_i \sum_j (n_{ij} - 1). \qquad (II:9)$$

## THE MULTIVARIATE STATISTICAL PROBLEM

We shall now consider the generalization of the univariate problem of testing the homogeneity of variances to the multivariate case. I referred to a test as the BARTLETT test for homogeneity of variances in the beginning of this chapter. This test has, in various forms, been generalized to multivariate form by several statisticians, among them G. BOX, S. KULLBACK and T. ANDERSON. It has come to be known as the generalized BARTLETT test, although it should be pointed out that Professor BARTLETT disclaims it as being his product (personal communication). This is not because the test is excessively bad, but rather because it is not particularly good. As far as I am aware, the procedure just outlined, in which an ANOVA model is used, has not yet been given generalization, although it is not difficult to write down an approximation without theoretical justification. Up to now, the only usable procedure seems to be the multivariate procedure of BOX et al. , and providing one is aware of its shortcomings and limitations, there is no particular reason why it may not be employed. Its application is best shown by means of a simple numerical illustration.

EXAMPLE:

Consider the sample covariance matrices, each based on 51 degrees of freedom,

$$S_1 = \begin{bmatrix} 2 & -1 \\ -1 & 1 \end{bmatrix}, \quad S_2 = \begin{bmatrix} 2 & 1 \\ 1 & 1 \end{bmatrix}, \text{ and,}$$

$$S_1 + S_2 = S = \begin{bmatrix} 2 & 0 \\ 0 & 1 \end{bmatrix}.$$

Both of these matrices are associated with identical mean vectors, notably,

$$\bar{x}_1 = \bar{x}_2 = (1,1)'.$$

The determinants of these matrices are, $|S_1| = 1$, $|S_2| = 1$ and $|S| = 2$.

The generalized test for homogeneity of covariance matrices, in the form presented in KULLBACK (1959) is:

$$2I(H_1:H_2(\cdot)) = N_1 \log_e \left[\frac{|S|}{|S_1|}\right] + N_2 \log_e \left[\frac{|S|}{|S_2|}\right]. \qquad (11:10)$$

Here, $NS = N_1S_1 + N_2S_2$ and $N = N_1 + N_2$. This is distributed as $\chi^2$ with $k(k+1)/2$ degrees of freedom, where k is the number of variables. In the present example, k=2. A better approximation than $\chi^2$ is that of $B^2$*, which, however, has only been tabulated for a few dimensions.

Applying (11:9) to the present example, one finds

$$B^2 = 50\log_e(2/1) + 50\log_e(2/1) = 69.315,$$

which is significant.

Inasmuch as the determinants of the covariance matrices, $S_1$ and $S_2$, are the same, both of these must have the same volume, since $|A| = (n-1)^p |S|$ is the squared volume of a parallelotope. (Here, A is the matrix of sums and squares and cross products, S the sample covariance matrix and n the sample size, with p the number of variables.)

---

*Read "Capital Beta".

In the above example, matrix $S_1$ has the eigenvalues $d_1 = 2.618$, and $d_2 = 0.382$. The corresponding, normalized eigenvectors are:

$$b_1 = \begin{bmatrix} .8507 \\ \\ -.5257 \end{bmatrix}, \qquad b_2 = \begin{bmatrix} .5257 \\ \\ .8507 \end{bmatrix} .$$

Matrix $S_2$ has the eigenvalues $g_1 = 2.618$, and $g_2 = 0.382$. The corresponding normalized eigenvectors are:

$$c_1 = \begin{bmatrix} .8507 \\ \\ .5257 \end{bmatrix}, \qquad c_2 = \begin{bmatrix} -.5257 \\ \\ .8507 \end{bmatrix} .$$

The ellipsoids of scatter of these two matrices have thus the same shape, but they are differently oriented. The angle between the axes is:

$\cos\theta = 0.4473$, hence, $\theta = 63^{\circ}25'$.

It is possible to compute the significance of the differences in the orientations of the ellipsoid axes by means of a statistical test adapted from Anderson (1963). The problem may be formulated as the hypothesis of co-linearity of eigenvectors of a covariance matrix. This may be discussed in the following terms.

The components of any eigenvector $a_i'$ are actually the direction cosines of the ith principal axis, namely the cosines of the p angles this axis makes with the p original axes taken in order. For example, the elements of $a_i'$ are $\cos\alpha$ and $\cos(90^{\circ} - \alpha) = \sin\alpha$, respectively. In some cases biological theory may prescribe values that these cosines should assume. Thus if the x-values are the logarithms of measurements of length made on an animal the hypothesis of equal relative growth rates of every part measured would imply that each of the components of $a_1'$ would equal $p^{-1/2}$. Certain differential growth rate hypotheses would also ascribe values to the components of $a_1'$.

An approximate (i.e., large sample) test of the appropriateness of a given (i.e., not derived from the data) eigenvector $\alpha_i'$, where $\alpha_i'\alpha_i = 1$ has been derived by Anderson (1963). He shows that

$$N\{\hat{\lambda}_i \alpha_i' \hat{\Sigma}^{-1} \alpha_i + \hat{\lambda}_i^{-1} \alpha_i' \hat{\Sigma} \alpha_i - 2\} \qquad \text{(II:11)}$$

is distributed as $\chi^2$ with p - 1 degrees of freedom. Note that if $\Lambda$ and A have been calculated the inverse of $\Sigma$ is easily obtained from the relation

$$\Sigma^{-1} = A' \Lambda^{-1} A \qquad \text{(II:12)}$$

where the inverse of $\Lambda$ is a diagonal matrix with typical element $\lambda_i^{-1}$.

Soley to illustrate the application of this test, we will assume that theory suggests that the first eigenvector of the female painted turtle measurements should be $\alpha_1' = (3^{-1/2}\ 3^{-1/2})$. Using relation (II:12).

$$\hat{\Sigma}^{-1} = \begin{bmatrix} 0{\cdot}06148 & -0{\cdot}05134 & -0{\cdot}07605 \\ & 0{\cdot}12858 & -0{\cdot}06932 \\ & & 0{\cdot}31490 \end{bmatrix}$$

and with N = 24 and $\hat{\lambda}_1 = 680.40$ the criterion (17) becomes

$$24 \times \{680{\cdot}40 \times 3^{-1} \times 0{\cdot}11154 + 0{\cdot}0014697 \times 3^{-1} \times 1776{\cdot}09 - 2\}$$

where 0.11154 and 1776.09 are the sums of all the elements of $\hat{\Sigma}^{-1}$ and $\hat{\Sigma}$, respectively. The criterion is thus 24.167 and, on the basis of the hypothetical theory, is a value of $\chi^2$ with p - 1 = 2 degrees of freedom. This is a highly improbable value to have obtained at random, and the hypothesis is accordingly rejected.

This theme has been further developed by me in a paper concerned with the analysis of growth patterns.

## CHAPTER III

## PRINCIPAL COMPONENT ANALYSIS

We shall introduce the subject of principal component analysis by means of the following geometric presentation. We shall consider the ellipsoid of scatter,

shown in Fig. 1.

What causes the elliptical shape of the probability (frequency) contours of Fig. 1 is that $\sigma_1$ and $\sigma_2$ have different values. These ellipses become circles when the units of measurement for $x_1$ and $x_2$ and $\sigma_1$ and $\sigma_2$, respectively. This is indicated in the first figure where $\sigma_1 = \sigma_2$.

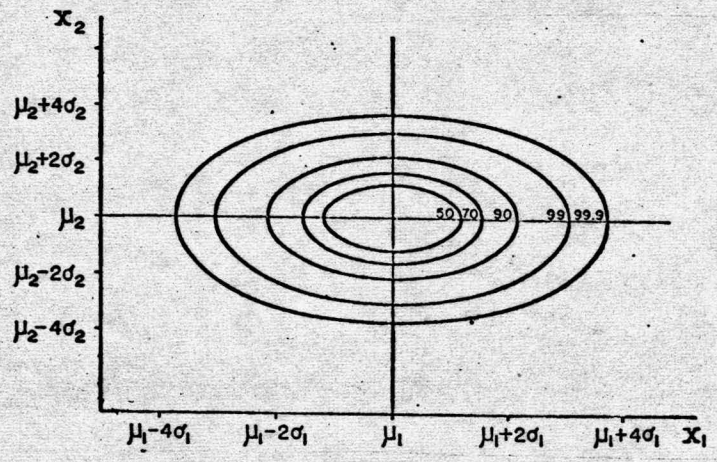



Fig. 1

What is fairly obvious about Fig. 1 is that the distribution of the $x_2$'s for a given $x_1$ is the same, namely $N(\mu_2, \sigma_2)$, whatever $x_1$ is. This expresses the fact that $x_1$ and $x_2$ are uncorrelated (which means 'independent' in the case of the bivariate Normal distribution). What happens when $x_1$ and $x_2$ are correlated is that the two axes of the ellipses in the second figure are rotated rigidly into a position depending on the covariance of $x_1$ and $x_2$. The result - with only one of the concentric equiprobability ellipses shown - is given in Fig. 2. Note that the actual pairs of correlated observations are measured along the horizontal and vertical axes, $X_1$ and $X_2$, respectively.

The original correlated measures $x_1$ and $x_2$ have been transformed into new uncorrelated measures $y_1$ and $y_2$, respectively. Figure 2 shows that the point $(x_1, x_2)$ on the ellipse becomes a point $(y_1, y_2)$ where

$$y_1 = (x_1-\mu_1)\cos\alpha + (x_2-\mu_2)\sin\alpha$$
$$y_2 = -(x_1-\mu_1)\sin\alpha + (x_2-\mu_2)\cos\alpha$$

These relations should be clear from the figure. Thus, for example, O'Q of length $y_1$ is made up of two parts:

(i) OM which is the projection of O'N, of length $x_1 - \mu_1$, on $O'Y_1$; and

(ii) MQ which is the projection of PN, of length $x_2 - \mu_2$, on $O'Y_1$.

The first of these projections is of length

$$(x_1 - \mu_1) \cos \alpha$$

and the second projection is of length

$$(x_2 - \mu_2) \sin \alpha$$

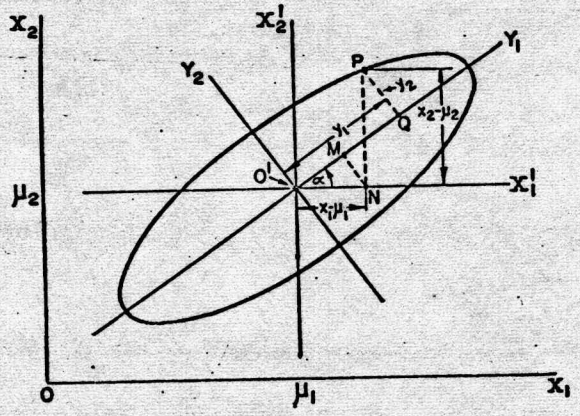



Fig. 2

Figure 2 shows the angle $\alpha$ as that particular angle through which $O'X_1'$ must be rotated in order that its new position $O'Y_1$ coincides (in direction) with the major axis of the family of ellipses generated by the given bivariate Normal distribution. $O'Y_2$ then coincides with the minor axis of this family. Let us ignore this for the moment and discover some of the properties of a rigid rotation of the $(x_1, x_2)$ axes through an arbitrary angle $\alpha$.

In matrix notation this general linear transformation of the x's may be written

$$y = A(x - \mu)$$

where

$$y = \begin{bmatrix} y_1 \\ y_2 \end{bmatrix} \qquad x - \mu = \begin{bmatrix} x_1 - \mu_1 \\ x_2 - \mu_2 \end{bmatrix}$$

and $$A = \begin{bmatrix} \cos\alpha & \sin\alpha \\ -\sin\alpha & \cos\alpha \end{bmatrix} .$$

We notice that

$$AA' = \begin{bmatrix} \cos\alpha & \sin\alpha \\ -\sin\alpha & \cos\alpha \end{bmatrix} \begin{bmatrix} \cos\alpha & -\sin\alpha \\ \sin\alpha & \cos\alpha \end{bmatrix}$$

$$= \begin{bmatrix} \cos^2\alpha + \sin^2\alpha & -\cos\alpha\sin\alpha + \sin\alpha\cos\alpha \\ -\sin\alpha\cos\alpha + \cos\alpha\sin\alpha & \sin^2\alpha + \cos^2\alpha \end{bmatrix}$$

$$= \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} = I \qquad (II:13)$$

On premultiplying by $A^{-1}$ we obtain

$$A' = A^{-1} \qquad (II:14)$$

It may also be confirmed that

$$A'A = I \qquad (II:15)$$

so that

$$A = (A')^{-1} \qquad (II:16)$$

These four relations characterize a so-called orthogonal transformation of x's into y's. Note that by premultiplying the transformation by $A^{-1}$ we obtain

$$A^{-1} y = I(x - \mu)$$

or, by relation (II:14),

$$A'y = x - \mu$$

This shows that the x's can be obtained from the y's just as easily as the y's from the x's.

Observe that although the trigonometrical functions of A emerge naturally from the axis-rotation of Fig. 2 we can write the orthogonality conditions in algebraic symbols. In this case we require

$$A \quad \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix}$$

to satisfy

$$AA' = I$$

Notice that these relations leave one element, say $a_{11}$, of A undetermined. This corresponds to the arbitrariness of $\alpha$ in the trigonometric transformation.

When we move to $p = 3$ we will abandon the trigonometric functions and consider transformations based on matrices with elements such that

$$AA' = I$$

As we have seen, this condition of orthogonality is not sufficient by itself to determine all the elements of A. Corresponding to the arbitrary angle $\alpha$ in the bivariate case, with three variables $(x_1, x_2, x_3)$ there will be two such angles; with four variables, three angles; and so on.

Finally, let us consider the expected value of the transformed vector y and the variance-covariance matrices of the x's and the y's. We have

$$\mathcal{E}y = A\mathcal{E}(x-\mu) = A0 = 0$$

where 0 is a two-component vector of zeros. Thus the variance-covariance matrix of the y's is given by $\mathcal{E}(yy')$, a $2 \times 2$ matrix. Now

$$\mathcal{E}(x-\mu)(x-\mu)' = \mathcal{E}\begin{bmatrix} (x_1-\mu_1)^2 & (x_1-\mu_1)(x_2-\mu_2) \\ (x_2-\mu_2)(x_1-\mu_1) & (x_2-\mu_2)^2 \end{bmatrix}$$

$$= \begin{bmatrix} \mathcal{E}(x_1-\mu_1)^2 & \mathcal{E}(x_1-\mu_1)(x_2-\mu_2) \\ \mathcal{E}(x_1-\mu_1)(x_2-\mu_2) & \mathcal{E}(x_2-\mu_2)^2 \end{bmatrix}$$

$$\equiv \begin{bmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{21} & \sigma_{22} \end{bmatrix} \equiv \Sigma$$

It is occasionally useful to write $\sigma_{12} / (\sigma_{11}\,\sigma_{22})^{1/2} = p$ where p is, by definition, the correlation coefficient between $x_1$ and $x_2$.

Hence

$$\begin{aligned}\mathcal{E}(yy') &= \mathcal{E}\{A(x-\mu)(A(x-\mu))'\} \\ &= \mathcal{E}\{A(x-\mu)(x-\mu)'A'\} \\ &= A\{\mathcal{E}(x-\mu)(x-\mu)'\}A' = A\Sigma A'\end{aligned}$$

The results of the foregoing may be summed up as follows. A rigid rotation of two rectangular coordinate axes, after suitable translation to a new center of coordinates (if desired), is equivalent to a linear transformation of the original coordinate values $(x_1, x_2)$. If the 2 x 2 matrix of coefficients applied to the vector x is A then $AA' = I$, and this implies that the four elements of A are related. The variance-covariance matrix of the transformed variates $(y_1, Y_2)$ is given by $A\Sigma A'$ where $\Sigma$ is the variance-covariance matrix of the x's.

These results are perfectly general and will extend very easily to values of p in excess of 2. In fact the remaining material in this part relies heavily on the concept of a translation of the origin and a rigid rotation of rectangular coordinate axes.

EXAMPLE (III:1)

For the purposes of illustration, we shall consider a bivariate example, which for generality, has been taken as square, asymmetric

Consider the matrix:

$$A = \begin{bmatrix} 5 & 4 \\ 1 & 2 \end{bmatrix}$$

The characteristic equation of this matrix is.-

$$\begin{bmatrix} 5-\lambda & 4 \\ 1 & 2-\lambda \end{bmatrix},$$

which expands to,

$$\lambda^2 - 7\lambda + 6 = 0.$$

The solution of this equation yields the eigenvalues:

$$\lambda_1 = 6, \lambda_2 = 1.$$

To each of these solutions, there will be a corresponding eigenvector.

For $\lambda_1 = 6$, we have that:

$$-x_1 + 4x_2 = 0$$

and,

$$x_1 - 4x_2 = 0$$

As we know, both of these equations must be linearly dependent. From each of these, one obtains the proportionality relationship:

$x_1 : x_2 = 4:1$, hence, the eigenvector,

$$\beta_1 = \begin{matrix} 4 \\ 1 \end{matrix}.$$

Usually, one will want to normalize this, which in this case is,

$$\theta_1 = \frac{1}{\sqrt{17}} \begin{Bmatrix} 4 \\ 1 \end{Bmatrix}.$$

For the eigenvalue $\lambda_2 = 1$.

$$4x_1 + 4x_2 = 0$$
$$x_1 + x_2 = 0.$$

This yields the eigenvector:

$$\beta_2 = \begin{Bmatrix} 1 \\ -1 \end{Bmatrix},$$

which normalizes to,

$$\theta_2 = \frac{1}{\sqrt{2}} \begin{Bmatrix} 1 \\ -1 \end{Bmatrix}$$

---

These steps are basically the same as those needed in the first half of the computations required in <u>factor analysis</u> by the so-called "principal component model".

# CHAPTER IV

## FACTOR ANALYSIS

In this chapter, we shall discuss some of the philosophy and methods of factor analysis. This branch of multivariate work was introduced, more or less in bits and pieces, by the psychologists, who required some sort of procedure in order to reduce their highly multivariate data to manageable proportions. It is important to bear this in mind, for it should be adequately realised, that the basic model is not one devised by professional mathematicians, but one put together by social scientists.
Let us examine what the psychologists claim to do with the factor-analytic model as applied to their kind of data.
The psychologist devises a series of tests which he applies to his "patients", study individuals, etc., and on the basis of the reactions of his subjects to the tests, he hopes to be able to disclose something about the "hidden vectors of the mind".
We have here the two catch-terms "devising of tests" and "hidden vectors".

If the geologist is assured that his data and his problem fit in with this model, then go ahead! Otherwise, I stress caution in the application of factor analysis to data of any kind - so-called <u>shotgun application</u>. My experience and considered opinion suggest, that factor analysis should only be used, and then cautiously, when the basic premises underlying the specialized psychologic model are clearly understood.

Factor Analysis has been described as 'a statistical technique for reducing a large number of correlated variables to terms of a small number of uncorrelated variables. The correlated variables consist usually of measurements for observable traits; the uncorrelated variables (called "factors") are abstract hypothetical

components.

This definition by two psychologists may sound like a principal components analysis in which p correlated variates are transformed into p uncorrelated (orthogonal) variates in descending order of variability, only the first k being 'significant' in the summarization. However, although this may not have been too clear in the past, Factor Analysis is now distinguished from principal components analysis by two characteristics:

(i) Each of the p original variates is supposed analyzable into $m < p$ mutually uncorrelated 'common factors' with an uncorrelated residual ('unique') component which is not correlated with any of the remaining $p - 1$ variates;

(ii) The m orthogonal axes of 'common factors' may be rotated to new orthogonal or oblique axes to conform with theoretical ideas underlying the formulation of the model.

The effect of the first of these characteristics is that only a portion of the variances or the unities in the diagonal of the estimated p x p variance-covariance or correlation matrix, respectively, is regarded as due to the $m < p$ transformed variates. When $m = p$ the residual component vanishes, the whole of the variances (or unities) is accounted for, and we are back at the p-variate orthogonal transformation. If only k of these principal axes are used (because of the 'sphericity' of the remaining components) the result is similar (though not identical) to a Factor Analysis with m set equal to k. The two sets of results have, however, been arrived at with different models in mind.

## THE MODEL

The mathematical form of the model that Factor Analysis assumes is:

$$\underset{(p\times 1)}{\mathbf{x}} = \underset{(p\times 1)}{\boldsymbol{\mu}} + \underset{(p\times m)}{\boldsymbol{\Gamma}} \underset{(m\times 1)}{\mathbf{f}} + \underset{(p\times 1)}{\mathbf{u}} \qquad \text{(IV:1)}$$

where x is $N(\mu, \Gamma\Gamma' + \Delta)$, f is $N(0, I)$ u is $N(0, \Delta)$ and is independent of f, and $\Delta$ is a p x p diagonal matrix with nonnegative elements. The matrix $\Gamma\Gamma'$ is required to be of rank $m<p$, that is to say there are $p - m$ linear relations connecting the p rows of this symmetrical matrix.

We note that

$$\begin{aligned}\Sigma \equiv \mathcal{E}\{(x-\mu)(x-\mu)'\} &= \mathcal{E}\{(\Gamma f+u)(\Gamma f+u)'\} \\ &= \mathcal{E}\{\Gamma f(\Gamma f)'+\Gamma f u'+u(\Gamma f)'+uu'\} \\ &= \mathcal{E}(\Gamma ff'\Gamma')+\mathcal{E}(uu') \quad \text{since } \mathcal{E}(u) = 0 = \mathcal{E}(f) \text{ and u and f are independent with } \mathcal{E}(ff') = I \\ &= \Gamma\mathcal{E}(ff')\Gamma'+\Delta\end{aligned}$$

i.e.,

$$\Sigma = \Gamma\Gamma'+\Delta \qquad \text{(IV:2)}$$

Comparing with the PCA model,

$$x = \mu + A^{-1} \qquad y = \mu + A'y$$

$$(p \times 1)\ (p \times 1)\ (p \times p)\ (p \times 1)$$

since $AA' = I$. Further, on premultiplying by A', and post-multiplying by A, the relation

$$A\Sigma A' = \Lambda$$

we obtain

$$\Sigma = A'\Lambda A = (A'\Lambda^{1/2})(\Lambda^{1/2}A) = JJ' \quad \text{(say)} \qquad \text{(IV:3)}$$

where $J' = \Lambda^{1/2}A$.

This comparison shows that the m components of f replace the p components of y and that there is a 'residual' u in the factor analytic model which finds no place in principal components. This vector variate u 'extracts' $\Delta$ from the diagonal elements of $\Sigma$.

Clearly the usefulness of Factor Analysis lies in the possibility of representing x (with p components) in terms of the relatively few components of f. The smaller m is the happier we shall be. In fact, in the earliest psychological writings at the

beginning of this century m was taken as unity.

The technique of Factor Analysis thus consists of estimating from a p-variate sample of N observations x: (i) $\mu$ (which estimate we will hereafter assume to be $\bar{x}$, the observational mean vector), (ii) $\Gamma$, the (p x m) matrix of coefficients known as 'factor loadings' (or 'saturations'), and (iii) $\Delta$, the matrix of variances of the 'unique' variate $\mu$. Note particularly that f is not generally 'estimated' since it is the m-variate Normal sample value corresponding to an observed p-variate x and is thus of little consequence when the Normal parameters $\mu$, $\Gamma$, $\Delta$ are known (or estimated). This shows that it is the structure of the model that is regarded as of importance. It is thought that the observed x are samples from a Normal universe specified by the particular parameters mentioned.

The model we have described above has a peculiar indeterminacy. Suppose we simultaneously transform f and the rows of $\Gamma$ by the two (m x m) matrices of coefficients A and T, respectively, so that f becomes f* and $\Gamma$ becomes G,

i.e., $\mathbf{f}^* = \mathbf{Af}$ and $\mathbf{G} = \Gamma\mathbf{T}$ (IV:4)

[Note that we post-multiply a matrix to transform its rows.] Then

$$\begin{aligned}
\mathscr{E}\{(\mathbf{x}-\mu)(\mathbf{x}-\mu)'\} &= \mathscr{E}\{(\mathbf{Gf}^*+\mathbf{u})(\mathbf{Gf}^*+\mathbf{u})'\} \\
&= \mathbf{G}\mathscr{E}(\mathbf{f}^*\mathbf{f}^{*\prime})\,\mathbf{G}'+\Delta \quad \text{the other terms being zero} \\
&= \mathbf{G}\mathscr{E}\{\mathbf{Af}(\mathbf{Af})'\}\,\mathbf{G}'+\Delta \\
&= \mathbf{GAA}'\,\mathbf{G}'+\Delta \quad \text{since } \mathscr{E}(\mathbf{ff}') = \mathbf{I} \\
&= (\Gamma\mathbf{T})\,\mathbf{AA}'(\Gamma\mathbf{T})'+\Delta \\
&= \Gamma\mathbf{TA}(\mathbf{TA})'\,\Gamma'+\Delta \\
&= \Gamma\Gamma'+\Delta \quad \textit{provided} \quad \mathbf{TA}(\mathbf{TA})' = \mathbf{I}
\end{aligned} \tag{IV:5}$$

which means that TA must be orthogonal. If, therefore, the two matrices of the simultaneous transformations of $\Gamma$ and f are chosen so that their product TA is orthogonal then the model is left unchanged.

It is this feature of the Factor Analysis model that permits the investigator to rotate the so-called common-factor axes (the columns of $\Gamma$) to any convenient

positions. In other words, having estimated a set of coefficients $\Gamma$, the Factor Analyst may arbitrarily transform them by post-multiplication of $\Gamma$ by any (m x m) matrix T. Of course the effect of this transformation is that the original factors f become a set f* which is $N(0, \psi)$, where $\psi$ is a variance-covariance matrix uniquely determined by the fact that TA is orthogonal and T is known. These new 'factors' are thus correlated with one another (but not with u) unless A is orthogonal (i.e., $\mathcal{E}(f*f*') = AA' = I$). When this latter requirement is satisfied T is also orthogonal (because TA must be). This means that if we are loath to make a transformation to correlated factors by an oblique rotation of the common-factor axes we may nevertheless use an othogonal transformation (rotation) on $\Gamma$.

Now, we will avoid this indeterminacy of model (IV:1) by requiring each successive component variate in f to absorb as much as possible of the variability in x remaining after that of u have been 'extracted'. Both the procedures we now describe do this - though is slightly different ways.

A few biologists have published work involving the use of Factor Analysis. Five of these [Kraus and Choi (1958), Teissier (1955), Goodall (1954), Bailey (1956), and Teissier (1956)] used standardized x-values (so that $\Sigma$ was the correlation matrix) and ignored the diagonal elements of $\Delta$. Geological applications are rather more numerous than biological. In the two last-mentioned articles the authors used orthogonal rotations of the component (factor?) axes (namely by means of an orthogonal T) to clarify the biological significance of their results.

The expression 'centroid method' refers to the Thurstone (1947) approximation to Principal Factor Analysis. This approximate procedure is now really only justified when electronic computers are not available.

Principal Factor Analysis

Our purpose is now to estimate $\Gamma$ and $\Delta$ from a sample of N p-variate

observations. In view of the Normality assumption we know that this sample can be replaced (without loss of 'information') by the statistics $\hat{\mu}$ and $\hat{\Sigma}$ calculated therefrom. One difficulty is that we then have to estimate both $\Gamma$ and $\Delta$ from the relation

$$\Sigma = \Gamma\Gamma' + \Delta$$

Let us suppose, however, that we can make a reasonable 'guess' at the p diagonal elements of $\Delta$. If we write these as $\hat{\Delta}$ we are then left with the estimation of $\Gamma$ as

$$\hat{\Gamma} \quad = \{g_1 \quad g_2 \quad \cdots \quad g_m\} \quad \text{(say)}$$

$(p \times m)$

from the relation

$$\hat{\Sigma} - \hat{\Delta} = \hat{\Gamma}\hat{\Gamma}' \qquad \text{(IV:6)}$$

This relation reminds us of (IV:3) and suggests that we should estimate $\Gamma$ in the same way as we obtained J in (IV:3), namely by principal components analysis. In this case, however, we will be working with the matrix $\hat{\Sigma}-\hat{\Delta}$. Each successive eigenvalue of this matrix will 'absorb' as much as possible of the balance of the aggregate variance, namely of the trace * of $\hat{\Sigma}-\hat{\Delta}$. It should be remembered that any matrix of the form BB' is symmetric and all its eigenvalues are nonnegative, i.e., it is positive semi-definite. If, then, there are negative values among the (largest) m eigenvalues of $\hat{\Sigma}-\hat{\Delta}$ it shows that either (i) $\hat{\Delta}$ is an inappropriate estimate or (ii) $\hat{\Sigma}-\hat{\Delta}$ cannot be written as $\hat{\Gamma}\hat{\Gamma}'$.

We will accordingly write

$$|(\hat{\Sigma} - \hat{\Delta}) - \kappa \mathbf{I}| = 0 \qquad \text{(IV:7)}$$

and proceed to calculate the m largest roots $\kappa_1, \kappa_2, \ldots \kappa_m$, the remainder being assumed to be zero since $\Gamma$ has only m columns. Using the ith of these eigenvalues we may obtain the first $\overline{p-1}$ elements of $g_i$ by solving the first $\overline{p-1}$ of the equations

$$g_i'(\hat{\Sigma} - \hat{\Delta}) = \kappa_i g_i' \qquad (i = 1, 2, \ldots m) \qquad \text{(IV:8)}$$

It is easily confirmed that if we obtained the pth element of $g_i$ by requiring the sum of the squares of the elements of $g_i$ to be unity, we would then have to multiply each of its elements by $\kappa_i^{1/2}$ in order for $\Gamma\Gamma'$ to be equal to $\Sigma-\Delta$. We may telescope these steps by requiring

$$g_i'g_i = \kappa_i \qquad (i = 1, 2, \ldots m) \tag{IV:9}$$

This equation (making the sum of the squares of the p components of $g_i$ equal to a non-negative number $\kappa_i$) completes the estimation of $\Gamma$ based on the guessed value of $\Delta$.

We remark that a given set of data may produce what appear to be perfectly reasonable mathematical solutions of the equations (IV:7), (IV:8) and (IV:9) for all assumed values of m, e.g., for m = 1,2, ... p - 1. Only the statistical test just mentioned can distinguish between these solutions and indicate what is the smallest value of m which will 'fit' the data given the hypothesis of Normality.

A special case of the foregoing analysis shows why some authors prefer to discard Factor Analysis, with all its complications, in favor of principal components. Let us suppose that all the elements of $\Delta$ are the same and that we know $\delta^2$ where

$$\Delta = \delta^2 I$$

We may picture $\delta^2$ as relatively small since otherwise the m 'common factors' have 'explained' too small a part of x for the analysis to be helpful.

In this case equation (IV:7) becomes

$$|(\hat{\Sigma}-\delta^2 I)-\kappa I| = 0$$

i.e.,

$$|\hat{\Sigma}-(\kappa+\delta^2)I| = 0$$

But

$$|\hat{\Sigma}-\lambda I| = 0$$

showing that

$$\lambda_i = \kappa_i+\delta^2 \qquad (i = 1, 2, \ldots p) \tag{IV:10}$$

Now the principal axes are given by

$$\mathbf{a}_i'\hat{\Sigma} = \lambda_i \mathbf{a}_i' \qquad \text{(IV:11)}$$

together with

$$\mathbf{a}_i'\mathbf{a}_i = 1 \qquad (i = 1, 2, \ldots p) \qquad \text{(IV:12)}$$

On the other hand the components of $\Gamma$ are obtained in this case from

$$\mathbf{g}_i'(\hat{\Sigma} - \delta^2 \mathbf{I}) = \kappa_i \mathbf{g}_i'$$

i.e., from

$$\mathbf{g}_i'(\hat{\Sigma} - \delta^2 \mathbf{I}) = (\lambda_i - \delta^2)\mathbf{g}_i' \quad \text{by (8)}$$

i.e., from

$$\mathbf{g}_i'\hat{\Sigma} = \lambda_i \mathbf{g}_i' \qquad \text{(IV:13)}$$

together with

$$\mathbf{g}_i'\mathbf{g}_i = \kappa_i = \lambda_i - \delta^2 \qquad (i = 1, 2, \ldots m) \qquad \text{(IV:14)}$$

Since equation (IV:11) is the same as equation (IV:13) except for a possible constant multiplier, the only differences between the $\mathbf{a}_i'$ and $\mathbf{g}_i'$ lie (i) in their different (scaling) normalizations (IV:12) and (IV:14), respectively, and (ii) in the fact that we ignore the last p - m g-vectors. In this particular case, therefore, principal components and Principal Factor Analysis are likely to give very similar results if m is only a little less than p.

A peculiar feature of the model is that, so long as $m < p$ and $\Delta \neq 0$, the p-dimensional space of x (namely, the p-dimensional space of the principal axes) is not the same space as that of the m factor-axes. In fact the latter is not even 'embedded' in the former. This has the consequence that although we can envisage x as a point in a flattened rugby football of points in p dimensions that m f-vectors cannot be represented in the football at all. This is very different from the situation where we left the points in situ and rotated the p axes of reference.

Text of adequacy of first model

Having estimated $\Gamma$ and $\Delta$ our problem is to decide from these estimates whether, in the p-variate Normal universe from which we are sampling,

$$\underset{(p\times p)}{\Sigma} = \underset{(p\times m)}{\Gamma}\ \underset{(m\times p)}{\Gamma'} + \underset{(p\times p)}{\Delta}$$

This test may be referred to as the test for the correctness of m.

Now the logarithm of the likelihood ratio, namely the ratio of the likelihood of the sample on the basis of the above hypothesis to the likelihood based on an arbitrary $\Sigma$, can be shown to be proportional to

$$N \ln \{|\hat{\Gamma}\hat{\Gamma}' + \hat{\Delta}|/|\hat{\Sigma}|\} \qquad \text{(IV:15)}$$

If the method of estimation used is efficient - which, in general, means that Canonical Factor Analysis (see below) has been used - this criterion is approximately distributed as $\chi^2$ with $(p - m)(p - m - 1)/2$ degrees of freedom. We may use this test as an approximation even when Principal Factor Analysis is the method of estimation used and also when $\Sigma_R$, instead of $\Sigma$, has been operated on.

Numerical illustration

As illustrations of the technique of Principal Factor Analysis we will apply it to the 8 x 8 variance and correlation matrices obtained from the Blackith-Roberts grasshopper data. The matrices analyzed were designated matrix (c) (with elements calculated to six decimals) and matrix (d) (with five decimals).

The first step is to choose a reasonably low value of m which we hope will lead to an adequate fit of the data. A rule-of-thumb used in psychometric texts is to choose m equal to the number of groupings of the p-variates which are revealed by the correlation matrix. References show that the variates numbered 2, 3 and 4 are fairly highly correlated and might thus constitute a group. On the other hand the variates 5, 6, 7, 8 and 9 are intercorrelated at a noticeably lower level and might be collected together in a second group. We may thus attempt a Factor Analysis with two 'common factors', i.e., with m = 2.

The next step is to make initial 'guesses' about the elements of $\Delta$. A feature of many iterative procedures is that no matter what starting values are used the final result is the same. Unfortunately this is not true with Principal Factor Analysis. Using an 11 x 11 matrix of observed correlations between assessments of primary emotions, Wrigley (1959) has shown that substantially different estimates of $\hat{\Gamma}$ are produced by the use of (i) zeros, (ii) unities, and (iii) the reciprocals of the corresponding element of the inverse of $\hat{\Sigma}_R$, in the principal diagonal of $\hat{\Delta}$. Incidentally his experiment indicated that when m was 4 or more about 100 iterations were required to produce successive values of $\hat{\Gamma}\hat{\Gamma}'$ that 'varied only slightly'.

Another disturbing feature of Wrigley's results was the frequency with which the 'final' estimate of $\Delta$ contained one or more negative elements. In fact in the two sets of solutions for m = 1, 2, 3, ... 10 with initial estimates (i) and (iii), respectively, only those for m = 3 avoided this failure of model (IV:1) which requires all elements of $\Delta$ to be nonnegative. While in the Wrigley example m = 3 was thus the only possible solution the use of the statistical criterion could still result in a judgement of 'bad fit'. This would be an example of the total failure of the factor analytic model.

In the numerical example we have chosen to use number (iii) of Wrigley alternative for the initial estimate for $\hat{\Delta}$. This is because it has a rather tenuous validity from a theoretical viewpoint (Harman, Chapter 5, 1960). That is to say, if we are operating on $\hat{\Sigma}_R$ the first estimate of $\Delta$ is

$$\hat{\Delta} = (\text{diag}\,\hat{\Sigma}_R^{-1})^{-1}$$

while if we are using $\hat{\Sigma}$ the corresponding estimate is

$$\hat{\Delta} = \{\text{diag}\,\hat{\Sigma}^{-1}\}^{-1}$$

Now with m = 2 criterion (IV:13) is approximately distributed as $\chi^2$ with 15 degrees of freedom. The 5% critical value of this distribution is 24.996. Using Principal Factor Analysis the test criterion (with N = 368) was 33.829 for matrix (c)

and 36.667 for matrix (d). Thus two 'common factors' are insufficient to fit the data. With m = 3 the $\chi^2$ for the variance-covariance matrix (c) is 17.478 in comparison with a 5% critical value of 18.307 (10 degrees of freedom). But it is only when we reach m = 4 that the calculated vale of (IV:13) for matrix (d) is below the corresponding critical value of $\chi^2$ (namely, 9.951 and 12.592, respectively). This awkward feature of Principal Factor Analysis, namely that two different factor models apply to one and the same set of data scaled in two different ways, is avoided by the technique to be developed below. We will therefore abstain from reproducing and 'interpreting' the vectors of 'loadings' in the two different $\hat{\Gamma}$. The table below summarizes the 'fit of the original matrices by comparing the diagonal (common factor) elements of $\hat{\Gamma}\hat{\Gamma}'$ with those of the original matrix.

Original variances and portion accounted for by $m$ common factors in a Principal Factor Analysis

| Variate no. | Matrix (c) Original | $m = 3$ | Matrix (d) Original | $m = 4$ |
|---|---|---|---|---|
| 2 | 0·0138 | 0·0123 | 1 | 0·851 |
| 3 | 0·0150 | 0·0107 | 1 | 0·751 |
| 4 | 0·2545 | 0·1887 | 1 | 0·703 |
| 5 | 0·0198 | 0·0084 | 1 | 0·468 |
| 6 | 0·0097 | 0·0029 | 1 | 0·513 |
| 7 | 0·0197 | 0·0102 | 1 | 0·482 |
| 8 | 0·0015 | 0·0003 | 1 | 0·298 |
| 9 | 0·4155 | 0·1777 | 1 | 0·472 |
| All | 0·7495 | 0·4112 | 8 | 4·538 |

It may be mentioned that, with the variance-covariance matrix, it required 18 iterations to produce a $\hat{\Delta}$, the elements of which did not differ from the preceding $\hat{\Delta}$ by as much as 5 units in the fifth decimal place. On the other hand with matrix (d) (m = 4) after 41 iterations there was still one element in the principal diagonal of $\hat{\Delta}$ that differed by as much as 2 units in the third decimal place from the corresponding element of the estimate of $\hat{\Delta}$ immediately preceding. These results, essentially confirming those of Wrigley (1959), lead us to view with some suspicion the published results of Factor Analyses obtained after small numbers of iterations.

A feature of the foregoing numerical procedures was the choice of a small m-value which was only discarded in favor of a larger value because the resulting

model did not fit the observations. This means that we have been seeking the model representation with the smallest possible m. In turn this implies that the elements of $\Delta$, the variances of the u-variates, are to be made as large as possible without overly disturbing the fit.

The end result is that the m-column matrix $\hat{\Gamma}$ will necessarily 'discard' into the u-variates a (much) larger portion of the aggregate variance (as measured by the sum of the diagonal elements of $\hat{\Sigma}$) than the first m columns of the corresponding principal components matrix J.

Furthermore, the number, $\mathcal{K}$, of significantly different eigenvalues (i.e., nonisoclinic variation which lends itself to interpretation) is likely to be larger than the smallest satisfactory value of m. Thus, for example, the m = 3 'common factors' of the Principal Factor Analysis of matrix (c) accounted for 55% of the trace of the original 8 x 8 matrix. An equal number of eigenvalues of the matrix accounted for 93%, while all eight eigenvalues were deemed to be significantly different. The statistician cannot help feeling dissatisfied with a technique that characterizes as 'unique' to that animal or plant group as much as 45% of its aggregate variability.

## Canonical Factor Analaysis

The validity of Principal Factor Analysis as a method of estimating $\Gamma$ and $\Delta$ may be seriously questioned on the ground that it produces different results when x is scaled. Reverting to the model relation we note that division of each component of x by its standard deviation (real or estimated) $\sigma_{ii}^{1/2}$ is equivalent to premultiplication of each side of (IV:1) by $(\text{diag}\Sigma)^{-1/2}$. But

$$(\text{diag}\,\Sigma)^{-1/2}(\mu+\Gamma f+u) = (\text{diag}\,\Sigma)^{-1/2}\mu+\{(\text{diag}\,\Sigma)^{-1/2}\Gamma\}f+(\text{diag}\,\Sigma)^{-1/2}u$$

The first of the three terms of this result shows that each component of $\mu$ is being standardized (as we should expect), while the last term shows that the transformed p-variate u retains a mean of 0 but has a variance matrix $\Delta_1$, say, given by

$\Delta_1 = \Sigma^{-1/2}\Delta\Sigma^{-1/2}$. Finally, if we are to retain the notion of unit Normal variates for the p components of f, the rows of $\Gamma$ have been scaled by division by $\sigma_{11}^{1/2}$, $\sigma_{22}^{1/2}$, ... $\sigma_{pp}^{1/2}$, respectively. This relationship is far from being satisfied by the Principal Factor solutions (i.e., estimates of $\Gamma$) of the variance-covariance and correlation matrix, respectively.

We will accordingly outline a preferred technique in which the foregoing scaling relationships hold good between the two solutions. This invariant procedure is obtained as the maximum-likelihood solution of the problem of estimating $\Gamma$ and $\Delta$.

As we may surmise, the noninvariance to scaling of Principal Factor Analysis is because it utilizes $\Sigma-\Delta$ instead of $\Delta^{-1/2}(\Sigma-\Delta)\Delta^{-1/2}$. The 'residual' or 'within' variance-covariance matrix is now $\Delta$ and the 'between' variability is represented by $\Sigma-\Delta$. We must thus replace equation (IV:7) by

$$|\hat{\Delta}^{-1/2}(\hat{\Sigma}-\hat{\Delta})\hat{\Delta}^{-1/2}-\nu\mathrm{I}| = 0$$

and $\hat{\Gamma}$ will be determined from

$$g_i'\hat{\Delta}^{-1/2}(\hat{\Sigma}-\hat{\Delta})\hat{\Delta}^{-1/2} = \nu_i g_i' \qquad (i = 1, 2, \ldots m)$$

together with

$$g_i'\hat{\Delta}^{-1}g_i = \nu_i$$

This solution is essentially the same as that provided by Lawley in his second paper on this subject. It has been the basis of all published numerical applications since then. In order to simplify the procedure for desk calculators Lawley proposed an iterative solution which avoided the need to solve the above determinantal equation.

Howe (1955) proposed an alternative procedure for use with desk calculators but by then Rao (1955) had arranged for a direct computer solution of the foregoing determinantal and vectorial equations. This program was used by Bechtoldt (1961) on a 17 x 17 correlation matrix with m = 6. This author states that he required only five

iterations to produce successive estimates of $\Delta$ which did not differ in any element by more than $\pm$ 0.01.

But the direct solution of the foregoing equations has the serious disadvantage that if any element of $\hat{\Delta}$ is nearly zero convergence of the iterative process is likely to be slow or even to fail. We propose, therefore, to rewrite the equations in a different form - although the final estimate of $\Gamma$ will be unchanged.

The above determinantal equation is equivalent to

$$|(\hat{\Sigma}-\hat{\Delta})-\nu\hat{\Delta}| = 0$$

i.e.,

$$|(\hat{\Sigma}-\hat{\Delta})-\nu(\hat{\Sigma}-\overline{\hat{\Sigma}-\hat{\Delta}})| = 0$$

i.e.,

$$|(1+\nu)(\hat{\Sigma}-\hat{\Delta})-\nu\hat{\Sigma}| = 0$$

or

$$|(\hat{\Sigma}-\hat{\Delta})-\theta\hat{\Sigma}| = 0 \qquad \text{(IV:16)}$$

where $\theta = \nu\ (1+\nu)$. This is the original from of Lawley's maximum-likelihood solution.

We now use the m largest roots of (IV:16) to obtain the columns $g_i$ of $\Gamma$, namely

$$g_i'(\hat{\Sigma}-\hat{\Delta}) = \theta_i g_i'\hat{\Sigma} \qquad (i = 1, 2, \ldots m) \qquad \text{(IV:17)}$$

together with the normalization

$$g_i'\hat{\Sigma}^{-1}g_i = \theta_i \qquad \text{(IV:18)}$$

Having completed this solution based on a preliminary $\hat{\Delta}$ we obtain a second, improved estimate $\hat{\Delta}$ from the relation

$$\hat{\hat{\Delta}} = \text{diag}\,(\hat{\Sigma}-\hat{\Gamma}\hat{\Gamma}') \qquad \text{(IV:19)}$$

and insert this in place of $\hat{\Delta}$ in (IV:16). Successive iterations are carried out until two estimates of $\hat{\Delta}$ agree to the desired accuracy. The final $\hat{\Delta}$ and the $\hat{\Gamma}$ that results from its use in (IV:16), (IV:17) and (IV:18) - assuming that no value of $\theta_i$ is negative - constitute the Canonical Factor Analysis.

Testing the adequacy of model (IV:1)

When Canonical Factor Analysis has been used to estimate $\Gamma$ and $\Delta$ the factor N in the test criterion can be refined and the second term simplified. The result is

$$\left\{N-m-\frac{2(p-m)+7-2/(p-m)}{6}\right\}\sum_{j=m+1}^{p}\ln(1-\theta_j) \qquad \text{(IV:20)}$$

which indicates that it is measuring the deviations from the hypothetical universal zeros of the smallest p-m sample roots of

$$|\hat{\Gamma}\hat{\Gamma}'-\theta\hat{\Sigma}| = 0$$

Numerical illustration

The foregoing techniques were applied to the 8 x 8 Blackith-Roberts matrix. As with Principal Factor Analysis, d of the Appendix was chosen as 4 so that the elements of the final $\hat{\Delta}$ can be assumed correct to four decimal places.

Trial with m = 2 resulted in a $\chi^2$-value significant at the 1% level. However, with m = 3 the criterion (IV:20) proved nonsignificant ($\chi^2$ = 14.484 with 10 degrees of freedom). The maximum-likelihood (i.e., in a certain sense, the 'best') factor structure of matrix (c) or (d), indifferently, thus agrees with that derived from a Principal Factor Analysis of matrix (c). Very few published Factor Analyses have operated on variance-covariance matrices.

The matrix $\Gamma$ produced after 23 iterations was

$$\hat{\Gamma} = \begin{bmatrix} 0{\cdot}1086 & -0{\cdot}0173 & 0{\cdot}0021 \\ 0{\cdot}1011 & -0{\cdot}0202 & -0{\cdot}0033 \\ 0{\cdot}3846 & 0{\cdot}1801 & 0{\cdot}0032 \\ 0{\cdot}0836 & 0{\cdot}0173 & -0{\cdot}0402 \\ 0{\cdot}0504 & -0{\cdot}0195 & 0{\cdot}0048 \\ 0{\cdot}0777 & 0{\cdot}0231 & 0{\cdot}0494 \\ 0{\cdot}0174 & 0{\cdot}0044 & -0{\cdot}0104 \\ 0{\cdot}3371 & 0{\cdot}2799 & 0{\cdot}0104 \end{bmatrix}$$

while the (diagonal) elements of $\hat{\Delta}$ were 0.0017, 0.0043, 0.0741, 0.0109, 0.0068, 0.0107, 0.0010, and 0.2234, respectively. Since $tr\hat{\Sigma} = 0.749382$ and $tr\hat{\Delta} = 0.33287$ (the last figure being spurious) we see that the common factors 'absorb' 55.6% of the aggregate variance. In our principal components analysis of the first three eigenvalues absorbed 95.4% of this variance. It should be clear by now that principal components and Factor Analysis are totally different techniques. It is indeed unfortunate that they have become confused in the literature.

As to the interpretation of the 'loadings' in $\hat{\Gamma}$ we can only draw the biologist's attention to (i) the 'general factor' and a second (bi-polar) factor in which the hind femoral and the elytron lengths play a dominant role, (ii) the antithetical position taken by the aggregate of the width of the grasshopper's head, its pronotal width, and its prozonal length in relation to the other five measured parts which are weighted differently, and (iii) the importance of the metazonal length and the hind femoral width, offsetting one another, in the third factor. Blackith (1960) has made a Centroid Factor Analysis of all ten (instead of eight) variates and the reader is referred to that paper for further details.

## Rotation

The objective and techniques associated with the 'rotation' of the m estimated common factor axes into supposedly more meaningful positions occupy a large part of the existing texts on Factor Analysis. Thus one chapter of Burt (1940), three chapters of Thurstone (1947), three of Thomson (1951), six (shorter) chapters of Cattell (1952), two of Adcock (1954), two of Fruchter (1954), four chapters of Harman (1960), and one chapter of Lawley and Maxwell (1963) are devoted specifically to the computational procedures of rotation while other chapters of these books usually contain discussions pro and con.

Two different types of argument are used to justify the rotation of a set of common factor axes found by some standard computational procedure - usually the Centroid approximation to Principal Factor Analysis. The first of these stems from the psychologist's feeling that he knows what 'factors' are common to certain of the p variates observed (which are usually 'tests' of one kind or another in his case) and which variates do not call these or other 'factors' into play. He can therefore group the variates together in a significant manner and can rotate the factor axes so that the loadings (g-coefficients in our notation) attached to factors that should not appear in the specified group of tests are (approximately) zero.

This argument lost its appeal when it appeared that psychologists had different views on the 'factors' that entered into the various tests used. Furthermore, the distribution of a linear compound of a number of random variables tends towards the Normal as the number of variates increases (unless one or more of the coefficients used dominate the rest).

However, indices constructed from Normally distributed variates may well have distributions that are far from Normal. An immediate example is provided by the sum of the squares of N Normally distributed variates measured from the sample mean. This sum is distributed proportionally to $\chi^2_{N-1}$ (the constant of proportionality being $\sigma^2$, the variance of the sampled universe), and this distribution only tends to Normality as N becomes large. Another example is the quotient of two Normally distributed variables.

But not all the procedures we have introduced are equally disturbed by a lack of Normality in the observations. The methods of Part A are all 'robust' in the sense that deviations from Normality of the distribution of the residual e are relatively unimportant (Chapter 10 of Scheffé, 1959). Though numerical examples are scarce in the

literature on this subject we would think that Bartlett's test of the equivalence of h variance-covariance matrices would be very sensitive to departures from Normality and that the same would be true of the tests of the sizes of k and m. In fact the whole concept of ellipsoids of variation and the rotation of axes to produce uncorrelated variates is so closely linked with the multivariate Normal distribution that these methods should really only be applied to data of this type. On the other hand the choice of the directions of the canonical axes has nothing to do with Normality, but the subsequent test of 'how many axes' has. This encourages us to use canonical analysis on quantal variates but leaves the test of its success rather rough and ready.

We have mentioned that some biologists have used the terminology of Factor Analysis in describing principal components analyses of multivariate data. Others, understandably, have used simpler techniques to achieve a similar objective. In addition there is a group of biological articles in which factor analytic methods have been used to discriminate between species.

(i) Principal components analysis is intended to achieve a parsimonious summarization of a random sample from a single universe of multivariate Normal measurements;

(ii) Canonical analysis is a procedure of discriminating as clearly as possible between two or more multivariate Normal universes with the same variance-covariance matrix; and

(iii) Factor Analysis is an attempt to elicit the underlying Normal multivariate structure of a universe that can be sampled with respect to many correlated variates.

While (i) may tend to merge into (iii) when we ignore the smallest p-k eigenvalues of its analysis, conceptually the bases and the computational procedures of the two techniques are (or should be) quite different.

Now suppose we have measured p variates on $N_j$ specimens of a given species of bee ($j = 1, 2, \ldots h$); $\sum_{j=1}^{h} N_j = N$. Some of these variates may be qualitative (e.g., the color of the abdomen), others discrete (e.g., the number of antennal segments), while some will be measured lengths or weights. Possibly some of these measurements will have been transformed by taking their logarithms or their square roots, etc., in order to make them conform more closely to Normal form. The most cogent procedure to discriminate between these h species of bees and to discover their underlying affinities is canonical analysis.

Following Bartlett (1948), let us examine the determinantal equation of such an analysis on the supposition that $N \to \infty$ and that p is larger than h-1. This equation can be written as

$$|\Xi - \nu\Sigma| = 0 \qquad \text{(IV:21)}$$

and has h-1 nonzero (positive) roots. Here $\Xi$ is the 'between species', and $\Sigma$ the 'within species', variance-covariance matrix. Now, if a symmetrical p x p determinant is of rank h-1 < p its order can be reduced from p to h-1 by the same linear operations on its rows and columns. We may therefore rewrite the foregoing determinantal equation as

$$|\Xi^* - \nu\Sigma^*| = 0 \qquad \text{(IV:22)}$$

where both matrices are (h-1) x (h-1) and symmetric.

We now suppose that the off-diagonal terms of $\Sigma^*$ are zero (a very severe restriction) and thus replace our equation by

$$|\Xi^* - \nu \operatorname{diag}\Sigma^*| = 0 \qquad \text{(IV:23)}$$

In order to reduce the matrix $\Xi^*$ to the form of a correlation matrix we must pre-multiply and post-multiply by $(\operatorname{diag}\Xi^*)^{-1/2}$. Doing this to the whole matrix expression results in

$$|R_P - \nu(\operatorname{diag}\Xi^*)^{-1/2} \operatorname{diag}\Sigma^* (\operatorname{diag}\Xi^*)^{-1/2}| = 0 \qquad \text{(IV:24)}$$

where

$$\mathbf{R}_P = (\text{diag}\,\Xi^*)^{-1/2}\,\Xi^*\,(\text{diag}\,\Xi^*)^{-1/2}$$

is the (h-1) x (h-1) correlation matrix between h species with the 'average species' eliminated, and the term which $\nu$ multiplies is a diagonal matrix, the typical element of which is the jth diagonal element of $\Sigma^*$ divided by the jth diagonal element of $\underline{\underline{\Sigma}}^*$. Now, if these diagonal elements can be replaced by unities (which is another strong restriction) the determinantal equation reduces to

$$|\mathbf{R}_P - \nu\mathbf{I}| = 0 \qquad \text{(IV:25)}$$

and the canonical analysis is then the equivalent of a principal components analysis of a matrix correlation coefficients between individual species of bees.

The foregoing analysis indicates that if (IV:1) the measured variates are uncorrelated (which could, of course, be achieved by a preliminary principal components analysis of $\Sigma$), and (IV:2) the corresponding diagonal elements of $\underline{\underline{\Sigma}}$ and $\Sigma$ are equal, then a canonical analysis may be replaced by a principal components analysis of correlation coefficients between the different species. We might attempt to estimate $R_p$ by calculating the h(h-1)/2 correlation coefficients between each pair of species of bees, any species being represented by p specified measurements made on one of its members.

We have thus seen that the replacement of a canonical analysis by a principal components analysis of a matrix of 'between species' correlation coefficients is plagued by unjustified assumptions. However, it is logically more satisfactory than the concept of drawing a random sample of p h-tuple measurements from an h-variate Normal universe, when p is actually a carefullychosen set of characteristics and the h variates are the species that Chance has made conveniently available to the biologist! The arbitrary nature of the subsequent Factor Analysis is emphasized when we try to apply maximum-likelihood and to determine the 'proper' value of m.

Matsakis (1957b), Sokal (1958b), Morishima and Oka (1960), and Rohlf and Sokal (1962), employ the dubious techniques based on between-species correlation coefficients. In all cases of this type we would recommend the replacement of the measurements on a single animal per species by measurements on at least two animals and the subsequent canonical analysis of the data. If certain of the characters do not vary at all 'within' a species they can be used as distinguishing marks supplementary to the species name.

The attempt to classify animals and plants by means of indices of 'between species' likeness has not been limited to Factor Analysis. Although some of these indices involve correlation coefficients or other statistical measures of similarity none of them regard the measured animals, plants or bacteria as samples from specific universes - as they surely must be. Thus none of these techniques, cited by Sneath and Sokal (1962), can provide any measure of the correctness of their classifications nor can convincing reasons be given for preferring any one method over the others.

There is, however, a case where correlations between observed species have statistical validity. Goodall (1954) has made a principal components analysis of a sample of 32 equal rectangular areas (160 x 80 m) of virgin mallee scrub vegetation. The sample areas were measured as to their percentage cover by each of 14 ( = p) species of plants, each percentage reading being transformed by the arcsine-root transformation. The first five principal axes of the resulting 14 x 14 correlation matrix were found to be 'significant' by Bartlett's criterion. Similarly Dagnelie's (1960) paper, cited in the exhibit, utilized coded qualitative variates in calculating the correlation coefficients between plant species observed in greater or lesser abundance in a sample of 80 beech groves. Another illustration is found in Reyment (1963). These papers provide examples where correlation 'between species' are valid parameters of a single p-variate Normal universe of 'areas'.

In fact, with this type of material it is hard to see the justification for calculating correlation coefficients between every pair of p areas based on a 'sample' of 20 'chosen' species (Williams and Lambert, 1961).

To this point, we have been concerned with applications of methods of factor analysis that have been available in the literature for several years. I shall now take up a newer method, in which several differences from the usually occurring factor model occur. In connection herewith, we shall again review some of the characteristics of principal component analysis, in order to be quite clear concerning the differences in the two types of models.

We shall consider an example involving correlations between species numbers.

The correlation coefficients between pairs of variables (species), ordered in a correlation matirx, are subjected to the mathematical procedure known as a transformation (equation (IV:1)). The new matrix D in equation (IV:1) has all off-diagonal elements equal to zero and its diagonal elements add up to the same total as the diagonal elements of the original correlation matrix. This diagonalization process is of wide application in applied mathematics. It was used by Jacobi over a hundred years ago in his study of the orbits of the planets and his solution is the one used in most electronic computer programmers today. Many statistical procedures make use of the process, including the statistical treatment of certain taxonomical problems and a problem in population dynamics.

The element $d_1$ of matrix D is larger than any of the succeeding elements. We shall here refer to it as an eigenvalue; other terms in use for it are 'latent root' and 'characteristic root'. Each eigenvalue is associated with two eigenvectors (for an asymmetric matrix there are two different eigenvectors, but in this paper we shall only be concerned with symmetric matrices, where the eigenvectors do not differ); other terms are 'latent vector' and 'characteristic vector'. The relationship is indicated in equation (IV:2).

It is easy to see that the elements of D represent another distribution of the total variance of the original matrix (the sum of its diagonal elements—this is called the *spur* or the *trace* of a matrix; thus, spur A = spur D in equation (2)).

In the taxonomical analysis of, for example, ostracods, the first eigenvalue is usually many times greater than the second eigenvalue and we say that most of the variation is concentrated to the *first principal component*. This is given by the elements of the first eigenvector, which are employed as the coefficients in an equation. Thus, if the elements of the eigenvector are ($b_1$, $b_2$, $b_3$), and the dimensions measured, $x_1$, $x_2$, $x_3$, the first principal component may be written out as

$$U_1 = b_1x_1 + b_2x_2 + b_3x_3. \qquad \textbf{(IV:26)}$$

The variance of $U_1$ is given by $d_1$, the first element of D.

In our palaeoecological problem the $x_i$ are species and the $b_i$ denote the importance of species $x_i$ in a particular principal component. We should therefore be able to recognize associations of species, yielded by the principal components, the relative significance of each species being indicated by the corresponding coefficient of the pertinent eigenvector (in statistical parlance this coefficient is sometimes called a 'weight' or 'weighting'; the corresponding psychometric term of factor analysis is 'loading').

It will usually be found that the first few eigenvalues account for almost all of the variance. The question is, then, whether any importance is attached to the remaining eigenvalues. This may be tested by seeing if the remaining eigenvalues are indistinguishable from each other (isotropic residue), which also implies that they are unimportant with respect to magnitude. Obviously, all eigenvalues must, in a sense, be significant (*cf.* Kendall 1957), unless we have a matrix, the rank of which is less than its order, which is a remote possibility in the kinds of biological problems we consider here. What the aforementioned test helps us to do is to decide the number of important eigenvalues; that is, in the ecological problem under review, the ones that have been most important in determining the frequencies in the various species associations, indicated by the elements of the corresponding eigenvalues. One would like to be able to test these coefficients for significance, but, as far as the author is aware, no such tests appear to be available. Consequently, there must be an element of subjectivity attached to decisions regarding the importance of any vector element.

Up to now the discussion has been in terms of principal component analysis. The ecological interpretation of the results of factor analysis is made in much the same way. However, the structure of factor analysis differs somewhat from principal component analysis, as indicated by equation (3), for it includes a term for random variation. The 'communalities' of factor analysis, discussed further on, are derived from the random matrix, as is indicated by equation (4). The method of factor analysis employed in the present paper would appear to be an improvement over previous methods, as it provides a more realistic assumption concerning the matrix of residual variances (equation (5)).

## STATEMENT OF THE PROBLEM

We have, say, $k$ samples from boreholes in which $p$ different species occur (not necessarily all $p$ species occur in each of the $k$ samples). For each sample a certain arbitrary total number of individuals, $N$, are picked (randomly), this $N$ being the same for all samples.

A matrix T of 'scores' (relative frequencies) results, each row of which consists of the score for each of the $p$ species in the $j$-th sample ($j = 1, \ldots, k$).

$$\mathbf{T} = \begin{bmatrix} v_{11}s_1 & . & v_{12}s_2 & . & . & . & . & v_{1p}s_{1p} \\ . & . & . & . & . & . & . & . \\ v_{p1}s_1 & . & . & . & . & . & . & v_{pp}s_p \end{bmatrix} \qquad \text{(IV:27)}$$

here $v_{ik}$ represents the score for species $s_k$ ($v_{ik} \geqslant 0$).

It is desired to use these observations on the frequencies of the species to ascertain the extent of relationship in their occurrence. That is, one would like to see if species that react in the same way to all factors in their environment can be marshalled into one group, and those that are adversely affected by some environmental facet can be made to show up in another group or groups.

For any symmetric ($k \times k$) matrix A, there exists an orthogonal matrix B such that

$$\mathbf{B'AB} = \mathbf{D} = \begin{bmatrix} d_1 & 0 & . & . & . & . & 0 \\ . & . & . & . & . & . & . \\ 0 & . & . & . & . & . & d_k \end{bmatrix} (d_1 > d_2 \ldots > d_k) \qquad \text{(IV:28)}$$

If the matrix A is positive definite, then the $d_i > 0$. (In other words a positive definite symmetric matrix has only positive eigenvalues.)

Corresponding to each eigenvalue there will be an eigenvector, c, such that if $d$ is any eigenvalue of A, c is a non-zero vector satisfying the equation

$$(\mathbf{A} - d\mathbf{I}) = 0 \qquad \text{(IV:29)}$$

These eigenvalues and eigenvectors provide the basis for the ecological analysis. Primarily we are interested in the components of the eigenvectors, which we may regard as *ecospectra*; i.e. the observational vectors are broken up into functionally connected spectra of ecological reaction. The same reasoning is applicable to the matrix of factor loadings of *factor analysis*, treated as method 2.

*Computational procedure for method* 1

1. Note the frequencies of each of the $p$ species in $k$ samples for a total of $N$ individuals.

2. Arrange the data for each sample in rows the one above the other; the order of the rows with respect to each other should be random (important for palaeoecological data in order to eliminate the possibility of serial correlation).

*Multivariate analysis*

| | $s_1$ | . | . | . | . | . | . | . | $s_p$ |
|---|---|---|---|---|---|---|---|---|---|
| Sample 1 | $v_{11}$ | . | . | . | . | . | . | . | $v_{1p}$ |
| . | . | . | . | . | . | . | . | . | . |
| Sample $k$ | $v_{p1}$ | . | . | . | . | . | . | . | $v_{pp}$ |

The $v_{ij}$ represents the score for the $j$-th species in the $i$-th sample. Because of the constant $\Sigma v_{ij} = N$, these scores are proportions which may be directly compared between samples. Logically, the greater one takes the value of $N$, the more likely are the components of the observational vectors to approach stability.

3. Compute all combinations of correlation coefficients between frequencies of species. It may be more suitable to express the data as decimal proportions at this stage.

4. The correlation coefficients are arranged into a matrix of correlations and this correlation matrix is reduced to its canonical form.

Owing to the closed-form nature of this table, there will be a small spurious correlation. Thus, a correlation of zero, will appear as -0.06. This is clearly unimportant in the present connection.

This is best done on an electronic computer as the arithmetical labor involved for more than three or four variables is prohibitive.

*Interpretation*

Each eigenvalue gives a quantitative indication of the relative importance of the species combination represented by the components of the associated eigenvector. Particularly in large matrices, it may be found that only the first few eigenvalues will be meaningful; the remaining eigenvalues can be tested to see whether they differ significantly from an isotropic residue (Lawley 1956). The eigenvector may be written out in equation form as

$$z = a_1b_1 + a_2b_2 + \ldots\ldots + a_pb_p$$

The usual computer solution of an eigenvalue problem provides normalized vector components so that an $a_i$ may receive any value between 0 and 1, depending on the importance of the contribution of the variable. Hence, the variables (= species) of the above equation that have large coefficients are said to be causing most of the variation represented in the eigenvalue. The reasoning is further elucidated in the worked example.

*Example using method* 1

The example considered here is palaeontological. It will be apparent, however, that the same reasoning may also be applied to data concerning relative abundances of living associations (*cf.* Greig-Smith 1957, p. 147).

The first 600 randomly selected individuals of twenty-eight borehole samples from western Nigeria, containing seventeen species of Paleocene (Cenozoic) ostracods were used to construct the correlation matrix given in Table 1. (Significant correlations printed in bold face type; these are 'total' correlations not taking the influences of the other $k$-2 variables into account.)

Table 1. *Correlation matrix based on collections of seventeen species of fossil ostracods*

| | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1·0000 | | | | | | | | | | | | | | | | |
| −0·3176 | 1·0000 | | | | | | | | | | | | | | | |
| 0·1363 | −0·1902 | 1·0000 | | | | | | | | | | | | | | |
| 0·1804 | −0·1775 | −0·3495 | 1·0000 | | | | | | | | | | | | | |
| −0·2547 | 0·1606 | −0·1285 | −0·1838 | 1·0000 | | | | | | | | | | | | |
| −0·0267 | −0·2299 | −0·1599 | 0·2067 | −0·0948 | 1·0000 | | | | | | | | | | | |
| −0·0038 | −0·1499 | −0·2642 | 0·0481 | −0·1279 | **0·4711** | 1·0000 | | | | | | | | | | |
| −0·1540 | −0·1756 | −0·2558 | −0·2026 | −0·1547 | 0·0639 | 0·0332 | 1·0000 | | | | | | | | | |
| −0·3659 | −0·1807 | −0·1862 | −0·0920 | −0·2624 | 0·0435 | 0·0945 | 0·0808 | 1·0000 | | | | | | | | |
| −0·0509 | −0·0028 | 0·2996 | −0·3229 | 0·2043 | −0·0899 | 0·3722 | 0·2368 | −0·0499 | 1·0000 | | | | | | | |
| −0·1097 | −0·2288 | −0·2509 | −0·2021 | **0·3699** | −0·1206 | 0·0400 | −0·1809 | −0·0083 | 0·0675 | 1·0000 | | | | | | |
| −0·2104 | −0·1734 | −0·1110 | 0·0320 | **0·5883** | −0·1685 | −0·2358 | −0·2040 | −0·1603 | −0·1889 | **0·7303** | 1·0000 | | | | | |
| −0·3085 | −0·3166 | 0·3419 | 0·0981 | −0·2349 | 0·0722 | −0·2635 | 0·0903 | 0·2735 | −0·1878 | −0·2679 | 0·0075 | 1·0000 | | | | |
| −0·0213 | −0·2386 | 0·2331 | −0·0183 | −0·1927 | 0·0792 | −0·1421 | −0·1107 | −0·0871 | −0·1408 | −0·1212 | −0·1286 | 0·2957 | 1·0000 | | | |
| 0·0130 | −0·2420 | −0·1383 | −0·0855 | −0·1567 | 0·3008 | 0·0942 | −0·0079 | 0·3093 | −0·0515 | −0·0136 | −0·2747 | 0·0277 | −0·0811 | 1·0000 | | |
| 0·1513 | −0·1549 | −0·5959 | 0·4269 | −0·3038 | 0·2614 | 0·2442 | −0·3562 | 0·3636 | −0·2613 | 0·0118 | −0·2447 | −0·1761 | −0·0099 | 0·4261 | 1·0000 | |
| 0·1389 | −0·1714 | −0·4500 | 0·1173 | 0·2005 | 0·2841 | 0·3086 | −0·4612 | 0·0315 | −0·1857 | 0·4265 | 0·3249 | −0·3127 | −0·1052 | −0·0165 | 0·4563 | 1·0000 |

The results of the principal component analysis are shown in Table 2.

We note that all eigenvalues are positive, which is a result of the correlation matrix being symmetric and positive definite. A very interesting aspect of the analysis is that the first component only accounts for 18% of the total variation. Compared with applications of principal component analysis in taxonomic work this is low indeed (*cf.* Reyment 1963). The first three eigenvalues add up to 46·57% of the total variation and the first four to 57%. The slight differences between successive eigenvalues indicate the great computational difficulties that would have been experienced had the work been done manually. The differences between successive eigenvalues are smaller between the first and last few eigenvalues than between the intermediate ones. We note a further feature of the results, notably, that the elements of the first eigenvector are not all positive, which is a situation differing from that usually found in principal component analysis of biological materials (*cf.* Reyment 1963).

One interpretation that suggests itself is that the several eigenvalues of roughly the same magnitude may indicate the existence of more than one ecological factor and that all are of roughly the same importance.

Using Lawley's (1956) test of significance it was found that the first five eigenvalues are significantly different from each other, while the remaining twelve may be regarded as being isotropic.

Approximate 95% confidence intervals for the eigenvalues were found as follows:

The procedure is to choose $l$ and $u$ so that

$$\Pr[nl<\chi^2_n] = \sqrt{(1-\epsilon)};\ \Pr[\chi^2_n<nu] = \sqrt{(1-\epsilon)}.$$

$$\text{Then } \Pr\left[\frac{\lambda_m(\mathrm{S})}{u} < \text{all } \lambda\,(\Sigma) < \frac{\lambda_M(\mathrm{S})}{l}\right] \geqq 1-\epsilon. \qquad \text{(IV:30)}$$

($\lambda$ = the eigenvalues and $n = N-1$; $\lambda_m$ = smallest eigenvalue and $\lambda_M$ = largest eigenvalue)

These intervals are $0{\cdot}07 < \text{all } \lambda_i < 35{\cdot}08$, with a probability of 0·96.

A complication encountered when interpreting the meaning of the eigenvectors is that the standard deviations of the variables are frequently very different, which is an outcome of the nature of the data. In the present problem it might have been advisable to perform some sort of scaling operation on the input material. The vector of standard deviations is:

$\mathbf{s}'$ = (0·1384, 0·1593, 0·0609, 0·0475, 0·0247, 0·0123, 0·0425, 0·0714, 0·0829, 0·0213, 0·0514, 0·0850, 0·0733, 0·0808, 0·0431, 0·0421, 0·0101)

Dividing each component of the eigenvectors by the appropriate standard deviation, as is often done in principal component analysis based on a correlation matrix, did not lead to intelligible results.

A rough interpretation of the eigenvectors corresponding to the first five eigenvalues is that the components of the first vector are mainly suggestive of some environmental factor which influences species 16 and 17 in one direction and species 3 in another. The second eigenvector is mainly concerned with species 5, 11 and 12 and suggests the exist-

Table 2. *The principal component analysis*

| Eigenvalue | Percentage of total variation and diff. $\lambda_i - \lambda_{i+1}$ | | Eigenvectors |
|---|---|---|---|
| 3·1052 | 18·27 | | (0·0832, −0·0973, −0·4252, 0·2438, 0·0065, 0·2300, 0·2360, −0·2753, 0·1080, −0·1759, 0·1954, 0·0852, −0.1990, −0·1003, 0·1690, 0·4427, 0·4435) |
| 2·8655 | 16·86 | 1·41 | (−0·1007, 0·1494, −0·0737, −0·1465, 0·4621, −0·2244, −0·1195, −0·1376, −0·2157, 0·0651, 0·3962, 0·4695, −0·2127, −0·1497, −0·2506, −0·2386, 0·1672) |
| 1·9441 | 11·44 | 5·42 | (−0·0323, 0·1795, −0·0531, −0·2826, 0·0705, 0·1092, 0·4530, 0·2376, 0·0294, 0·5075, −0·0079, −0·2717, −0·3912, −0·3246, 0·1172, −0·0308, −0·0144) |
| 1·7764 | 10·45 | 0·99 | (0·4703, 0·3122, −0·0128, 0·2283, −0·0938, −0·1424, −0·0681, −0·2084, −0·4765, −0·0761, −0·2868, −0·2290, −0·3481, −0·0224, −0·2487, 0·0270, −0·0126) |
| 1·6070 | 9·45 | 1·00 | (−0·4411, 0·5877, −0·3302, 0·0077, 0·0149, −0·1952, −0·2028, −0·0827, 0·2877, −0·2192, −0·1959, −0·1202, 0·0128, −0·2083, 0·0313, 0·1065, −0·1587) |
| 1·1766 | 6·92 | 2·53 | (−0·3485, 0·1197, −0·1118, 0·3392, 0·1588, 0·4946, 0·3246, 0·1046, −0·2307, 0·0051, −0·2213, 0·0641, 0·1974, 0·1327, −0·3858, −0·1939, 0·0337) |
| 1·0167 | 5·98 | 0·94 | (−0·1522, 0·2167, 0·1220, −0·4273, 0·0597, 0·0898, 0·0825, −0·4874, −0·0272, 0·1011, −0·0061, −0·1876, −0·0137, 0·6258, 0·0580, 0·0531, 0·1796) |
| 0·8610 | 5·06 | 0·92 | (0·0465, 0·0717, 0·0209, −0·1252, 0·3317, 0·4378, −0·2595, 0·0706, −0·3452, −0·2520, −0·1079, 0·0190, −0·0469, −0·0974, 0·6129, −0·0976, −0·1053) |
| 0·6940 | 4·08 | 0·98 | (0·0110, −0·0613, 0·2005, 0·3263, 0·3224, −0·1049, 0·0111 −0·5049, 0·0432, 0·4600, −0·2023, 0·0037, 0·3202, −0·2689, 0·1878, 0·0864, −0·0869) |
| 0·6050 | 3·56 | 0·52 | (−0·1580, −0·1108, −0·3080, 0·3368, 0·1484, −0·2197, −0·0507, 0·1772, −0·1719, 0·2609, 0·1767, −0·0202, −0·2246, 0·4767, 0·1991, 0·1872, −0·4107) |
| 0·4302 | 2·53 | 1·03 | (0·1081, −0·1690, 0·0023, −0·0274, 0·5267, 0·0105, −0·3173, 0·2839, 0·3200, 0·0632, −0·3887, −0·1422, −0·1655, 0·1482, −0·2202, 0·1739, 0·3072) |
| 0·3077 | 1·81 | 0·72 | (−0·2783, 0·2141, 0·5282, 0·1785, −0·2111, 0·2332, −0·3230, 0·1314, −0·1002, 0·1839, 0·2030, 0·0830, −0·1806, −0·0689, −0·0704, 0·4424, 0·1234) |
| 0·2286 | 1·34 | 0·47 | (0·2314, 0·0514, 0·0591, 0·1078, 0·0234, 0·4292, −0·0955, −0·2610, 0·4973, −0·0047, 0·1418, 0·1084, −0·3362, 0·0570, −0·1597, −0·1986, −0·4542) |
| 0·2025 | 1·19 | 0·15 | (0·2141, 0·1178, −0·4328, −0·0664, −0·1403, 0·2547, −0·4470, 0·0370, −0·0612, 0·4007, 0·2759, −0·2355, 0·3607, −0·0298, −0·1124, −0·0899, 0·1232) |
| 0·1194 | 0·70 | 0·49 | (−0·1768, −0·0193, 0·0062, 0·3237, −0·2712, −0·0804, −0·1662, −0·0321, 0·1293, 0·1536, −0·1420, 0·0268, −0·2809, 0·1001, 0·3031, −0·5810, 0·4188) |
| 0·0403 | 0·24 | 0·46 | (−0·1869, −0·1720, 0·1738, 0·1889, 0·2332, −0·0474, −0·0011, −0·0830, −0·0181, −0·2729, 0·4286, −0·7038, −0·0342, −0·1215, −0·0848, −0·1680, 0·0301) |
| 0·0198 | 0·11 | 0·13 | (0·3735, 0·5383, 0·1971, 0·2544, 0·1939, −0·1027, 0·2350, 0·2889, 0·2147, −0·0601, 0·2427, 0·0343, 0·2514, 0·2109, 0·1967, −0·0496, 0·1472) |

ence of some environmental factor that influences all of these in the same way. The third eigenvector indicates the existence of an environmental factor that influences species 7 and 10 in one way and species 13 and 14 oppositely. The fourth vector suggests a factor that influenced species 1 and 2 in one direction and species 9 and 13 in another. The fifth component would appear to indicate an influence on species 1 and 5 in one direction and on species 2 in another.

Now, in principal component analysis it is assumed that the variation observed is mainly due to factors having characteristic effects on each of the variates. It is therefore open to question whether the foregoing analytic model is really the best way of attacking the problem. The related technique of factor analysis takes the possibility of extraneous variation into account. The $r_{ii} = 1$ terms of the correlation matrix are replaced by so-called 'communalities' that represent the proportion of the variation of each variate which is due to factors operating on some or all of the other variables, the rest being assumed due to chance. It should be pointed out that the communality concept of factor analysis is not accepted by all quantitative biologists as being valid. Also, in principal component analysis one is primarily interested in studying the *shape* of the scatter of the observations.

The method of *factor analysis* here employed is that proposed by Jöreskog (1962). As it involves certain modifications it is given detailed discussion.

The fundamental postulate of factor analysis is

$$\mathbf{S} = \alpha\beta + \epsilon \qquad \text{(IV:31)}$$

where $\alpha = (\alpha_{it})$, $\beta = (\beta_{tv})$ and $\epsilon = (\epsilon_{iv})$. Here $\alpha_{it}$ and $\beta_{tv}$ are non-random quantities, while matrix $\epsilon$ is a set of random variables for which $E(\epsilon) = 0$ and $E(\epsilon\epsilon') = n\Delta$, where $\Delta$ is a positive diagonal matrix, the diagonal elements of which are the residual variances. We note that the $\alpha_{it}$ are termed factor loadings (= coefficients) of $k$ common factors and the $\beta_{tv}$ are the factor values of the individuals.

Factor analysis assumes the covariance matrix to be made up of two matrices:

$$\Sigma = \alpha\alpha' + \Delta \qquad \text{(IV:32)}$$

The diagonal elements of $\Sigma - \Delta$ are called *communalities*.

Under the common assumption of factor analysis that $\Delta = \sigma^2\mathbf{I}$, where $\sigma^2$ is a positive constant and $\mathbf{I}$ is the unit matrix, a test for significant eigenvalues may give a large number of common factors, since the $p-k$ eigenvalues of the population correlation matrix (which is what is used in most analyses in place of the covariance matrix) are not likely to be nearly isotropic unless $k$ is large. In order to find the least number of common factors Jöreskog (1962) has proposed another assumption for $\Delta$, namely, that

$$\Delta = \theta(\text{diag}\ \Sigma^{-1})^{-1} \qquad \text{(IV:33)}$$

where $\theta$ is a positive scalar and $\Sigma$ is non-singular.

*Computational procedure for method* 2

We shall write C for the correlation matrix. Beginning with the sample correlation matrix the following steps are required:

1. Compute $D = \text{diag.}^{-1/2}$

2. $C^* = D\ 1/2\ CD\ 1/2$

3. $C^* - l^* I$ , where the $l^*_i$ are the eigenvalues of $C^*$. At the same time the corresponding eigenvectors are found, $Z_i^{*\prime} = (Z^*_{1i}, Z^*_{2i}, \ldots, Z_{pi})'$.

4. The diagonal elements of $L^*$ are tested for significance. $Lk^*$ is the diagonal matrix of the first k eigenvalues of $L^*$.

5. Then one computes

$$A^* = Z_k^*(L_k^* - t_k I)^{\frac{1}{2}} \qquad \text{(IV:34)}$$

where t is defined as

$$t_k = \frac{1}{p-k} \sum_{i=k+1}^{p} l_i^*$$

and A, the estimate of $\alpha$, is $A = D^{-1/2} A^*$.

6. The diagonal matrix of estimated residual variances R is

$$R = t\ D^{-1} \qquad \text{(IV:35)}$$

7. In order to find a lower confidence limit for k one first tests the hypothesis $k = 1$ against $k > 1$; if this hypothesis is rejected one test $k = 2$ against $k > 2$, and so on. The number of factors for which the test does not show significance is determined. If the level of significance is 5% in each test, this method leads to a lower confidence limit for the number of common factors. The test criterion is (after Lawley 1956 and Whittle 1952)

$$c_k = -(n-1)[\log (l^*_{k+1} \cdots l^*_p) - (p-k) \log t] \qquad \text{(IV:36)}$$

where t is defined as in a foregoing equation and the $l^*_i$ are the eigenvalues of $C^*$. If the hypothesis is true and if n is large, $c_k$ is approximately distributed as $\chi^2$ with degrees of freedom

$$d_k = \tfrac{1}{2}(p-k+2)(p-k-1) \qquad \text{(IV:37)}$$

We use the same correlation matrix as in the previous example, that is, the data of Table 1. We observe that the number of observational vectors, twenty-eight, is relatively small for an analysis in which five factors appear to be of some consequence. Furthermore, the observations deviate widely from normality, which is not a desirable situation in a factor analysis. Component analysis does not necessarily require normally distributed variables but the tests of significance for it as well as for factor analysis do. The results of the computations appear in order below.

1.
$$\mathbf{D} = \begin{bmatrix} 9{\cdot}29 & & & & & & & & & & & & & & & & & \\ & 15{\cdot}94 & & & & & & & & & & & & & & & & \\ & & 4{\cdot}90 & & & & & & & & & & & & & & & \\ & & & 5{\cdot}90 & & & & & & & & & & & & & & \\ & & & & 5{\cdot}15 & & & & & & & & & & & & & \\ & & & & & 2{\cdot}55 & & & & & & & & & & & & \\ & & & & & & 4{\cdot}93 & & & & & & & & & & & \\ & & & & & & & 5{\cdot}67 & & & & & & & & & & \\ & & & & & & & & 4{\cdot}26 & & & & & & & & & \\ & & & & & & & & & 3{\cdot}81 & & & & & & & & \\ & & & & & & & & & & 8{\cdot}90 & & & & & & & \\ & & & & & & & & & & & 12{\cdot}97 & & & & & & \\ & & & & & & & & & & & & 5{\cdot}60 & & & & & \\ & & & & & & & & & & & & & 3{\cdot}75 & & & & \\ & & & & & & & & & & & & & & 3{\cdot}95 & & & \\ & & & & & & & & & & & & & & & 4{\cdot}77 & & \\ & & & & & & & & & & & & & & & & 4{\cdot}25 \end{bmatrix}$$

2. Computing $\mathbf{C}^* = \mathbf{D}^{\frac{1}{2}}\mathbf{C}\mathbf{D}^{\frac{1}{2}}$ and extracting the eigenvalues gives the matrix

$$\mathbf{L}^* = \begin{bmatrix} 23{\cdot}76 & & & & & & & & & & & & & & & & & \\ & 20{\cdot}35 & & & & & & & & & & & & & & & & \\ & & 14{\cdot}62 & & & & & & & & & & & & & & & \\ & & & 11{\cdot}17 & & & & & & & & & & & & & & \\ & & & & 9{\cdot}66 & & & & & & & & & & & & & \\ & & & & & 5{\cdot}79 & & & & & & & & & & & & \\ & & & & & & 4{\cdot}80 & & & & & & & & & & & \\ & & & & & & & 3{\cdot}60 & & & & & & & & & & \\ & & & & & & & & 3{\cdot}17 & & & & & & & & & \\ & & & & & & & & & 2{\cdot}73 & & & & & & & & \\ & & & & & & & & & & 2{\cdot}32 & & & & & & & \\ & & & & & & & & & & & 1{\cdot}57 & & & & & & \\ & & & & & & & & & & & & 1{\cdot}12 & & & & & \\ & & & & & & & & & & & & & 0{\cdot}90 & & & & \\ & & & & & & & & & & & & & & 0{\cdot}57 & & & \\ & & & & & & & & & & & & & & & 0{\cdot}32 & & \\ & & & & & & & & & & & & & & & & 0{\cdot}13 \end{bmatrix}$$

Up to this point factor analysis does not differ essentially from principal component analysis. Compared with the results of the principal component analysis it is interesting to note that the eigenvalues are percentually of the same order of magnitude. More essential, however, is the fact that the $p-k$ smallest eigenvalues of $\mathbf{C}^*$ tend to be more like each other than do the $p-k$ smallest eigenvalues of $\mathbf{C}$.

The first eigenvalue represents 22·29% of the total variation, the second 19·09% and the third 13·72%. The first three eigenvalues add up to 55·1% of the total variation, and the first four to 65·6%; these figures indicate a somewhat greater concentration of variation to the first eigenvalues than was found for the correlation matrix alone.

3. The results of the tests of significance for the number of common factors are given in the following table. (If $d_k > 100$, $c_k$ is a $\lambda$-value, otherwise $c_k$ is a $\chi^2$-value.)

| $k$ | $t$ | $d_k$ | $c_k$ |
|---|---|---|---|
| 1 | 5·18 | 135 | 5·35*** |
| 2 | 4·16 | 119 | 4·17*** |
| 3 | 3·42 | 104 | 3·27*** |
| 4 | 2·82 | 90 | 126·47** |
| 5 | 2·25 | 77 | 95·67 |
| 6 | 1·93 | 65 | 79·26 |
| 7 | 1·64 | 54 | 64·69 |
| 8 | 1·43 | 44 | 54·30 |
| 9 | 1·21 | 35 | 44·04 |
| 10 | 0·99 | 27 | 33·84 |

The approximate 95% confidence interval for the eigenvalues is, using the same formula as before, $0{\cdot}08 < \text{all } \lambda_i < 45{\cdot}73$, with a probability of 0·96.

Hence the smallest number of factors that may be considered to fit the model is $k = 5$, which agrees with the principal component analysis. However, since the sample size is small and the data not truly multivariate normal, it might be wise to consider a further five factors.

4. To find the unrotated factor loadings one computes

$$\mathbf{A} = \mathbf{D}^{\frac{1}{2}}\mathbf{A}^* \qquad \text{(IV:38)}$$

where $\mathbf{A}^* = \mathbf{Z}_k^*(\mathbf{L}_k^* - t_k\mathbf{I})^{\frac{1}{2}}$. It is common in psychometry to 'rotate' these vectors by some graphical device or other in order to present the data in 'standard' terms; rotation is not considered useful in most biological situations by the present writer. For $k = 10$, the matrix of factor loadings is

$$\mathbf{A} = \begin{bmatrix}
-0{\cdot}15 & -0{\cdot}47 & 0{\cdot}30 & -0{\cdot}72 & -0{\cdot}03 & 0{\cdot}09 & 0{\cdot}09 & 0{\cdot}01 & -0{\cdot}04 & -0{\cdot}06 \\
-0{\cdot}25 & 0{\cdot}90 & 0{\cdot}21 & -0{\cdot}02 & -0{\cdot}06 & 0{\cdot}03 & 0{\cdot}00 & -0{\cdot}03 & 0{\cdot}00 & -0{\cdot}01 \\
-0{\cdot}21 & -0{\cdot}12 & -0{\cdot}69 & -0{\cdot}36 & 0{\cdot}02 & 0{\cdot}13 & -0{\cdot}18 & 0{\cdot}03 & -0{\cdot}13 & -0{\cdot}05 \\
-0{\cdot}02 & -0{\cdot}30 & 0{\cdot}38 & 0{\cdot}07 & -0{\cdot}54 & -0{\cdot}48 & 0{\cdot}09 & 0{\cdot}00 & -0{\cdot}09 & 0{\cdot}20 \\
0{\cdot}59 & 0{\cdot}40 & -0{\cdot}06 & -0{\cdot}10 & 0{\cdot}10 & -0{\cdot}14 & -0{\cdot}11 & 0{\cdot}44 & 0{\cdot}02 & 0{\cdot}05 \\
-0{\cdot}10 & -0{\cdot}31 & 0{\cdot}21 & 0{\cdot}28 & 0{\cdot}12 & -0{\cdot}29 & -0{\cdot}25 & 0{\cdot}20 & 0{\cdot}32 & -0{\cdot}24 \\
-0{\cdot}09 & -0{\cdot}21 & 0{\cdot}35 & 0{\cdot}19 & 0{\cdot}56 & -0{\cdot}37 & -0{\cdot}31 & -0{\cdot}11 & -0{\cdot}06 & -0{\cdot}08 \\
-0{\cdot}25 & -0{\cdot}09 & -0{\cdot}52 & 0{\cdot}06 & 0{\cdot}36 & -0{\cdot}32 & 0{\cdot}41 & -0{\cdot}10 & 0{\cdot}22 & -0{\cdot}03 \\
-0{\cdot}09 & -0{\cdot}18 & -0{\cdot}01 & 0{\cdot}69 & 0{\cdot}12 & 0{\cdot}26 & 0{\cdot}17 & -0{\cdot}07 & -0{\cdot}27 & -0{\cdot}08 \\
-0{\cdot}09 & 0{\cdot}08 & -0{\cdot}21 & -0{\cdot}14 & 0{\cdot}66 & -0{\cdot}16 & -0{\cdot}21 & 0{\cdot}07 & -0{\cdot}31 & 0{\cdot}27 \\
0{\cdot}85 & -0{\cdot}02 & 0{\cdot}08 & 0{\cdot}02 & 0{\cdot}27 & 0{\cdot}18 & 0{\cdot}05 & -0{\cdot}17 & 0{\cdot}02 & 0{\cdot}10 \\
0{\cdot}93 & 0{\cdot}09 & -0{\cdot}13 & -0{\cdot}02 & -0{\cdot}17 & -0{\cdot}07 & 0{\cdot}04 & 0{\cdot}00 & -0{\cdot}01 & -0{\cdot}04 \\
-0{\cdot}11 & -0{\cdot}26 & -0{\cdot}57 & 0{\cdot}38 & -0{\cdot}44 & 0{\cdot}01 & -0{\cdot}14 & 0{\cdot}05 & -0{\cdot}14 & -0{\cdot}10 \\
-0{\cdot}11 & -0{\cdot}24 & -0{\cdot}23 & 0{\cdot}01 & -0{\cdot}26 & 0{\cdot}20 & -0{\cdot}49 & -0{\cdot}20 & 0{\cdot}37 & 0{\cdot}25 \\
-0{\cdot}16 & -0{\cdot}32 & 0{\cdot}16 & 0{\cdot}33 & 0{\cdot}22 & 0{\cdot}35 & 0{\cdot}17 & 0{\cdot}46 & 0{\cdot}15 & 0{\cdot}09 \\
-0{\cdot}10 & -0{\cdot}36 & 0{\cdot}69 & 0{\cdot}34 & -0{\cdot}06 & 0{\cdot}16 & 0{\cdot}05 & 0{\cdot}00 & -0{\cdot}02 & 0{\cdot}12 \\
0{\cdot}47 & -0{\cdot}18 & 0{\cdot}58 & 0{\cdot}05 & 0{\cdot}07 & 0{\cdot}03 & -0{\cdot}23 & -0{\cdot}02 & 0{\cdot}01 & -0{\cdot}29 \\
\hline
2{\cdot}47 & 1{\cdot}93 & 2{\cdot}54 & 1{\cdot}70 & 1{\cdot}66 & 0{\cdot}95 & 0{\cdot}84 & 0{\cdot}56 & 0{\cdot}54 & 0{\cdot}40
\end{bmatrix}$$

The last row gives the contribution of each factor to the total test variance.

The results show clearly that these ten factors reproduce both the correlations and variances of the original matrix very well and it is quite possible that even a lesser number of factors would do this too. The communalities are found from finding diag C—R, where $R = fD^{-1}$. That is, (1—estimated residual variance). There are in order:

0·90, 0·94, 0·80, 0·84, 0·81, 0·62, 0·80, 0·83, 0·77, 0·74, 0·89,
0·93, 0·83, 0·74, 0·75, 0·80, 0·25.

The first factor indicates that species 5, 11, 12 and possibly 17 are affected in the same way by some environmental factor; the other elements of the vector do not differ significanlty from zero. The second factor suggests an environmental control that mainly influences species 2 but also to a lesser degree, in the same direction, species 5. Species 1, 15 and 16 are influenced in the opposite direction. The third factor seems to represent an environmental control that affects species 3, 8 and 13 in one direction and species 16 and 17 in another direction. The fourth vector would appear to indicate the influence of an environmental factor which rather strongly affects species 1 in one direction and about equally as strongly species 9 in the reverse direction; other affected species are 3, 13, 15 and 16. The fifth vector suggests relatively strong influences on species 4, 7 and 10 in opposite directions, and lesser influences on species 8 and 13. These five factors seem to be the most informative. The remaining five factors included in the matrix of factor loadings are remarkable in that none of them suggests a strong reaction of any environmental agent on any of the species.

Hence, it would seem that most of the variation in frequencies of the seventeen species may have been controlled by five environmental stimuli of some kind or other (for example, temperature, light, salinity, variation in chemical proportions of sea water, pH, redox).

If as a criterion of non-reactivity to environment we take small factor loadings it may be suggested that species 6, 14 and 15 are euryoic and this agrees extremely well with what is to be observed qualitatively in the material. Judging from occurrences in the borehole samples one gains the impression that species 1 and 9 are also euryoic. Both of these are strongly affected only by the fourth factor, which may indicate that this factor is an unimportant one with respect to actual distribution. None of the species appears to give the impression of being stenoöic.

It will be seen that the results of the factor analysis tend to differ somewhat in detail from the component analysis, although both give a strong impression of the operation of several approximately equally important environmental factors.

Inasmuch as the factor analysis model would seem to be more suitable for the kind of ecological data here treated (remembering we are not studying the shape of the distribution) the results obtained with its aid could be more descriptive of the actual situation.

**The program has been adjusted for the IBM 7040 and is running for most data. Several copies of the binary cards are presently available at K.U. It will not work for off-diagonal correlations of 1. and stops if negative eigenvalues are encountered.**

SUMMING UP:

It is important to have clearly in mind what it is you are attempting to find out about your data. I think in most cases, a principal component analysis, or an analysis by canonical variates, is the applicable procedure and not factor analysis. As I have tried to point out in the foregoing, it is imperative that the data be considered at length before the factor model is selected. Even then, the results should be tested for interpretation against the same data run in other multivariate procedures. The concept of "when in doubt, throw in a factor analysis" is most definitely not a sound one.

# CHAPTER V

## THE METHOD OF CANONICAL VARIATES

This important multivariate-statistical procedure was devised by H. Hotelling in 1936 as a part of his program for treating the complex data obtained by social scientists.

The procedure has not achieved recognition to the extent it deserves and in many respects other multivariate methods, and non-statistical methods (Numerical Taxonomy) have been, and are being, applied, although these are inferior in properties to the Hotelling method.

The first basic requirement for the application of canonical variate analysis (hereinafter referred to as CVA) is that the covariance matrices fulfil the requirement of equality, notably, that,

$$\Sigma_1 = \ldots\ldots = \Sigma_q, \qquad \text{(V:1)}$$

where there are q parent populations.
If this requirement has been met, our next concern is to examine h mean vectors, $\mu^{(i)}$, i = 1, h, which will, to the greatest degree possible, on the grounds of the variables analysed, bring out the relative closeness of any pairs of mean vectors and will also bring out such ordering as exists in these mean vectors.

At first sight the most promising approach is to transform the original $p$ axes of coordinates to the *common* principal component axes. For this purpose we would calculate the common estimate of the variance–covariance matrix, namely

$$\hat{\Sigma} = (N-h)^{-1}\left\{\sum_{i=1}^{h}(N_i-1)\hat{\Sigma}^{(i)}\right\} \qquad \text{(V:2)}$$

and proceed to find its eigenvalues and eigenvectors (principal axes). The $h$ transformed mean vectors could then be compared quite easily since they each consist of $p$ variate values which are mutually independent and which are ranked in descending order of variance size.

The foregoing approach forms the basis of Rao's concept of the generalized statistical distances between end points of the h, standardized mean vector in, say, k-space.

Each of the $\binom{h}{2}$ distances between these standardized means (where, for example, the ith component of any mean vector is standardized by dividing it by $\lambda_i^{1/2)}$ is calculated from the formula

$$D^2 = d_1^2+d_2^2+\ldots+d_p^2 \qquad \text{(V:3)}$$

where $d_i$ is the difference between the ith components of the specified pair of (estimated) mean vectors, and D is the required distance. The mutual relationships connecting the distances between any group of two, three, four, ... of the h samples then form the basis of a subdivision into 'group constellations' describing the affinities of the h samples.

Besides computing the $\binom{h}{2}$ distances Blackith (1960) recommends the calculation of the angles between any pair of vectors joining the end of a given mean vector to the ends of two other mean vectors in other groups. If $d_1$, $d_2$, ... $d_p$ are defined as above and if $\delta_1$, $\delta_2$, ... $\delta_p$ be the corresponding values measured from the first mean to a third sample mean, the angle between the two 'distance' vectors is $\theta$ given by

$$\cos\theta = (d_1\delta_1+d_2\delta_2+\ldots+d_p\delta_p)/D_1D_2 \qquad \text{(V:4)}$$

$D_1$ and $D_2$ being the two distances under consideration.

This technique has been used in a number of biological connections. In geology, we find geologists progressing by trying to do the same thing by means of the factor analytic model. For examples, see the Computer Applications in the Earth Sciences Colloquium at KU, December 15-16, 1966.

This technique suffers from the following defect. Suppose that p is fairly large and that h = 2. Clearly the whole of the comparison between the two universes

is contained in the one-dimensional comparison of two points on a straight line, these points representing the ends of the two (estimated) mean vectors. In other words, the ends of the mean vectors can be compared in a single dimension, a subspace of the original p-dimensional space.

This argument extends quite simply to the case where h = 3 and the comparison of the three p-dimensional mean vectors is made by plotting the projections of the ends of the three vectors on a plane (i.e., in two dimensions). This plotting is conveniently done by using a pair of axes at right angles to one another. In general, the comparison of a number of universes h<p should be made in a space of h - 1 dimensions rather than a space of p dimensions.

Another criticism of the 'distance' technique is that its rationalization depends heavily on all p variates being measured in the same units. Furthermore, experience suggests that the dimensionality of the $D^2$ comparisons with biometric data will generally be greater (even for h > p) than that of the method now to be described.

Canonical Axes

Consider the whole p-dimensional sample space of the h universes. Since the variance-covariance matrices of these universes are supposed equal we could represent the differences between the mean vectors of these h universes by a model of the form

$$\underset{(N\times p)}{\mathbf{X}} = \underset{(N\times h)}{\mathbf{Z}'}\ \underset{(h\times p)}{\mathbf{B}} + \underset{(N\times p)}{\mathbf{E}}\,. \qquad \text{(V:5)}$$

Note that h rows of B are needed to account for (i) the general mean of the whole sample of N, where $N = N_1 + N_2 + \ldots + N_h$, and (ii) the h - 1 differences between the means of the h different universes. Observe that the variance-covariance matrix of each of the p-variate sample 'errors' is $\Sigma\,\Omega$.

Out objective is to derive a transformation

$$\mathbf{y} = \mathbf{C}\mathbf{x} \qquad \text{(V:6)}$$

which will emphasize the differences between the means of the h universes (or, rather, of their sample estimates). For example, if h = 2 the first axis along which to measure the first component of y should be the line joining the ends of the two mean vectors of these universes, or a line parallel thereto. When h = 3 this first axis should pass as closely as possible through all three points representing the ends of the mean vectors of the three universes.

Now, suppose that the $\Omega$-model represented by (V:5) is replaced by a model in which the B matrix has degenerated into a single row of $\beta$'s representing the p means of all the N x's. We may write the estimated variance-covariance matrix of this model as $\hat{\Sigma}_{\omega}$. Our interest then resides in the variance-covariance matrix of the difference between the models $\Omega$ and $\omega$ for this measures the variability 'between' the h groups. It is estimated by

$$\{(N-1)\hat{\Sigma}_{\omega} - (N-h)\hat{\Sigma}_{\Omega}\}/(h-1) \qquad \text{(V:7)}$$

or, by

$$\{\hat{\mathrm{B}}'_{\Omega}\mathrm{ZZ}'\hat{\mathrm{B}}_{\Omega} - \hat{\mathrm{B}}'_{\omega}\mathrm{ZZ}'\hat{\mathrm{B}}_{\omega}\}/(h-1)$$

We will write this matrix as $\hat{\underline{\underline{\Box}}}$.

We note, in passing, that both $\Omega$ and $\omega$ could include other explanatory z-vectors. For example, we could compare the means of h groups of frogs after allowing for age differences by a linear or quadratic function.

Previously, we found a transformation

$$\mathbf{y} = \mathbf{A}\mathbf{x}$$

such that the first axis was inclined along the direction of the maximum variability among the N p-dimensional observations. Then a second axis, at right angles to the first, was inclined in the direction of the next greatest variability. And so on. The procedure we developed could have been arrived at by determing the components

$a_1'\Sigma a_1$ and (since an unrestricted maximization leads to infinite components for $a_1'$ ) we would have had to require the maximization to be subject to $a_1'a_1 = 1$. Then, one would conclude, that the components of $a_1'$ were obtainable from the p equations

$$a_1'\Sigma = \lambda_1 a_1'$$

where $\lambda_1$ is the largest root of

$$|\Sigma - \lambda I| = 0$$

This process could then have been repeated for $a_2'$ corresponding to the second largest root $\lambda_2$; and so on.

We may express the whole series of maximizations in matrix form by saying that we were required to maximize the variance-covariance matrix

$$A\Sigma A'$$

subject to

$$AA' = I$$

and that the elements of A were then determinable by the p sets of p equations

$$A\Sigma = \Lambda A$$

where the elements of the diagonal matrix $\Lambda$ are obtained as the p roots of the determinantal equation

$$|\Sigma - \lambda I| = 0$$

Although we can no longer think in terms of directed ellipsoids of variation and their major and minor axes our present problem is nevertheless quite similar. We wish our first transformed axis to be inclined in the direction of the greatest variability 'between' the h means; then our second axis, at right angles to the first, is to be inclined in the direction of next greatest variability and so on. And whereas in our first problem we operated with a (p x p) variance-covariance matrix $\Sigma$ we are now using a (p x p) variance-covariance matrix $\Xi$ .

Our transformation matrix C of (V:2) thus has a first row $c_1'$ which maximizes the 'between' variance of $y_1$ , namely

$$c_1'\Xi c_1 \qquad \text{(V:8)}$$

Since this maximization would lead to infinite components for $c_1'$ we must impose restrictions on the transformation matrix. We thus stipulate that the variates y shall be uncorrelated (i.e., have zero covariances) and each be of unit variance. But the variance-covariance matrix of the transformation (V:2) is $C\Sigma_\Omega C'$ and we are thus making the maximization of (V:8) subject to

$$c_1'\Sigma_\Omega c_1 = 1 \tag{V:9}$$

The foregoing requirements for the first row of C can be extended to the second, third,... rows so that we may finally state our problem in matrix form as follows: To maximize the 'between' variance-covariance matrix

$$C\Xi C' \tag{V:10}$$

subject to

$$C\Sigma_\Omega C' = I \tag{V:11}$$

Now we may write (V:11) in the form

$$(C\Sigma_\Omega^{1/2})(\Sigma_\Omega^{1/2}C') = (C\Sigma_\Omega^{1/2})(C\Sigma_\Omega^{1/2})' = I$$

and putting

$$C\Sigma_\Omega^{1/2} = F$$

this is the same as

$$FF' = I \tag{V:12}$$

With this notation (V:10) becomes

$$\begin{aligned}&(C\Sigma_\Omega^{1/2})\Sigma_\Omega^{-1/2}\Xi\Sigma_\Omega^{-1/2}(C\Sigma_\Omega^{1/2})'\\ &= F(\Sigma_\Omega^{-1/2}\Xi\Sigma_\Omega^{-1/2})F'\end{aligned} \tag{V:13}$$

Our procedure is then as follows. We first find the roots $\nu_1 > \nu_2 > \nu_3 > \ldots$ of the pth degree polynomial equation

$$|\Sigma_\Omega^{-1/2}\Xi\Sigma_\Omega^{-1/2} - \nu I| = 0 \tag{V:14}$$

and then determine the p components of $f_i'$ the ith row of F from the p equations

$$f_i'\Sigma_\Omega^{-1/2}\Xi\Sigma_\Omega^{-1/2} = \nu_i f_i' \tag{V:15}$$

supplemented by the relation

$$\mathbf{f}_i'\mathbf{f}_i = 1 \quad . \tag{V:16}$$

Finally, we find the p components of $\mathbf{c}_i'$ from the p relations

$$\mathbf{c}_i'\Sigma_\Omega^{1/2} = \mathbf{f}_i' \tag{V:17}$$

As we might suspect, the p roots of (V:14) are only distinguishable when $p \leqslant h - 1$. When $p > h - 1$ there are $(p - h - 1)$ zero roots and $h - 1$ distinguishable other roots.

EXAMPLE (V:1)

As an example of the calculations, we shall consider some bivariate data on frogs from Southern Sweden. The first step is to present the data in the form of the multivariate analysis of variance (often merely referred to as MANOVA). This is shown in the following table:

| | Degrees of Freedom | Length | Breadth |
|---|---|---|---|
| P matrix | 1 | 10·80 | 16·15 |
| | | 16·15 | 24·15 |
| W matrix $= 47\hat{\Sigma}_\Omega$ | 47 | 841·45 | 938·09 |
| | | 938·09 | 1099·65 |
| 'Total' matrix $= 48\hat{\Sigma}_\omega$ | 48 | 852·25 | 954·24 |
| | | 954·24 | 1123·80 |

The W matrix is obtained directly from $\hat{\Sigma}$ while, e.g., the sum of squares of cranial lengths 'between' the sexes is

$$\frac{(35 \times 22{\cdot}860)^2}{35} + \frac{(14 \times 21{\cdot}821)^2}{14} - \frac{(35 \times 22{\cdot}860 + 14 \times 21{\cdot}821)^2}{49}$$

As a matter of interest we may calculate the ratio of the determinants $|47\hat{\Sigma}_\Omega|$ to $|48\hat{\Sigma}_\omega|$, namely 45,288/47,185 = 0·9598.

For computations on a desk calculator the determinantal equation is best pre- and post-multiplied by $W^{1/2}$ giving

$$|P - \phi W| = 0$$

We have stated that when $p > h-1$ (as it is here) this equation in $\phi$ will only have $h-1$ non-zero roots. This is easily verified here ($h = 2$) when we insert the numerical values of P and W into the quadratic, for the term not involving $\phi$ (namely $10{\cdot}80 \times 24{\cdot}15 - 16{\cdot}15 \times 16{\cdot}15$) vanishes. Let us, however, proceed less directly in a manner that extends quite easily to larger values of $p$.

Let us write

$$\Delta(\phi) \equiv |\mathrm{P} - \phi \mathrm{W}| \equiv \phi^{p-1}(\alpha + \beta\phi)$$

where $\alpha$ and $\beta$ are determinable coefficients. Then the equation

$$\Delta(\phi) = 0$$

has the solution $\phi = 0$ ($p-1$ roots) or $\alpha + \beta\phi = 0$. The required non-zero root is thus $\phi_1 = -\alpha/\beta$.

Suppose we evaluate $\Delta(\phi)$ at $\phi = \pm 1$; this provides two points on the curve of $\Delta(\phi)$ and enables us to calculate $\phi_1$.

In this case

$$\alpha + \beta = \Delta(1) = |\mathrm{P} - \mathrm{W}| = \begin{vmatrix} -830{\cdot}65 & -921{\cdot}94 \\ -921{\cdot}94 & -1075{\cdot}50 \end{vmatrix} = 43390{\cdot}7$$

$$-\alpha + \beta = \Delta(-1) = |\mathrm{P} + \mathrm{W}| = \begin{vmatrix} 852{\cdot}25 & 954{\cdot}24 \\ 954{\cdot}24 & 1123{\cdot}80 \end{vmatrix} = 47184{\cdot}6$$

and thus

$$\phi_1 = -\frac{\alpha}{\beta} = \frac{3793{\cdot}9}{90575{\cdot}3} = 0{\cdot}041887$$

There is thus only one vector of $c$-components and, since $p = 2$, there are two components in the vector. Writing these two components as $a$ and $b$ the two equations are

$$[a \quad b]\begin{bmatrix} 10{\cdot}80 & 16{\cdot}15 \\ 16{\cdot}15 & 24{\cdot}15 \end{bmatrix} = (0{\cdot}041887)\,[a \quad b]\begin{bmatrix} 841{\cdot}45 & 938{\cdot}09 \\ 938{\cdot}09 & 1099{\cdot}65 \end{bmatrix}$$

Both lead to the same result so we write only the first one, namely

$$10{\cdot}80a + 16{\cdot}15b = 352{\cdot}46a + 392{\cdot}94b$$

i.e.,

$$341{\cdot}66a + 376{\cdot}79b = 0 \quad \text{or} \quad a = -1{\cdot}102\,82b$$

To supplement this

$$[a \quad b]\begin{bmatrix} 841{\cdot}45 & 938{\cdot}09 \\ 938{\cdot}09 & 1099{\cdot}65 \end{bmatrix}\begin{bmatrix} a \\ b \end{bmatrix} = 47$$

i.e.,

$$841{\cdot}45a^2 + 1876{\cdot}18ab + 1099{\cdot}65b^2 = 47$$

and substituting for $a$ this becomes

$$53{\cdot}943b^2 = 47$$

whence

$$b = \pm 0{\cdot}93343$$

and thus

$$a = \mp 1{\cdot}02941$$

Note that we may choose whether to make $b$ positive or negative but that the sign of $a$ is then fixed as the opposite to what we have chosen.

The canonical transformation has thus become

$$y = [1{\cdot}02941 \quad -0{\cdot}93343]\,\mathrm{x}$$

This axis clearly passes through the origin of the $x$-measurements since

$$\mathbf{x} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \text{ implies } y = 0$$

It is, however, more helpful to move the origin of the $y$'s to the grand mean of all $N = 49$ observations. This may be done by writing

$$y = [1{\cdot}02941 \quad -0{\cdot}93343] \begin{bmatrix} x_1 - 22{\cdot}563 \\ x_2 - 23{\cdot}953 \end{bmatrix}$$

On this axis the mean of the female cranial measurements is

$$\bar{y}_{\varphi} = [1{\cdot}02941 \quad -0{\cdot}93343] \begin{bmatrix} 0{\cdot}297 \\ 0{\cdot}444 \end{bmatrix} = -0{\cdot}1087$$

while

$$\bar{y}_{\delta} = [1{\cdot}02941 \quad -0{\cdot}93343] \begin{bmatrix} -0{\cdot}742 \\ -1{\cdot}110 \end{bmatrix} = 0{\cdot}2723$$

The single linear function $\mathbf{c}_1'(\mathbf{x} - \bar{\mathbf{x}})$ obtained when $h = 2$ is known as the discriminant function. Having calculated $\mathbf{c}_1'$ and $\bar{\mathbf{x}}$ from two (large) groups of observations a new observational vector $\mathbf{x}$ may be inserted into the discriminant function and the observational unit allocated to one or other of the two groups (universes) depending on whether the result is positive or negative, respectively.

## Example (V.2)

This example concerns four samples of two species of Rana from four localities in Europe. Here we then have four groups.
The various stages in the calculations are supplied here below

MEANS

| | | |
|---|---|---|
| 59.26000 | 30.77000 | 2.93000 |
| 60.30000 | 31.15000 | 2,91000 |
| 76.89000 | 41.34000 | 3,93000 |
| 66.08000 | 33.47000 | 4.22000 |

WITHIN GROUPS SSCP MATRIX

| | | |
|---|---|---|
| 5734.25376 | 3436.53560 | 372.87610 |
| 3436.53560 | 2103.70678 | 193.50070 |
| 372.87610 | 193.50070 | 45.56860 |

AMONG GROUPS SSCP MATRIX,A(I,J)

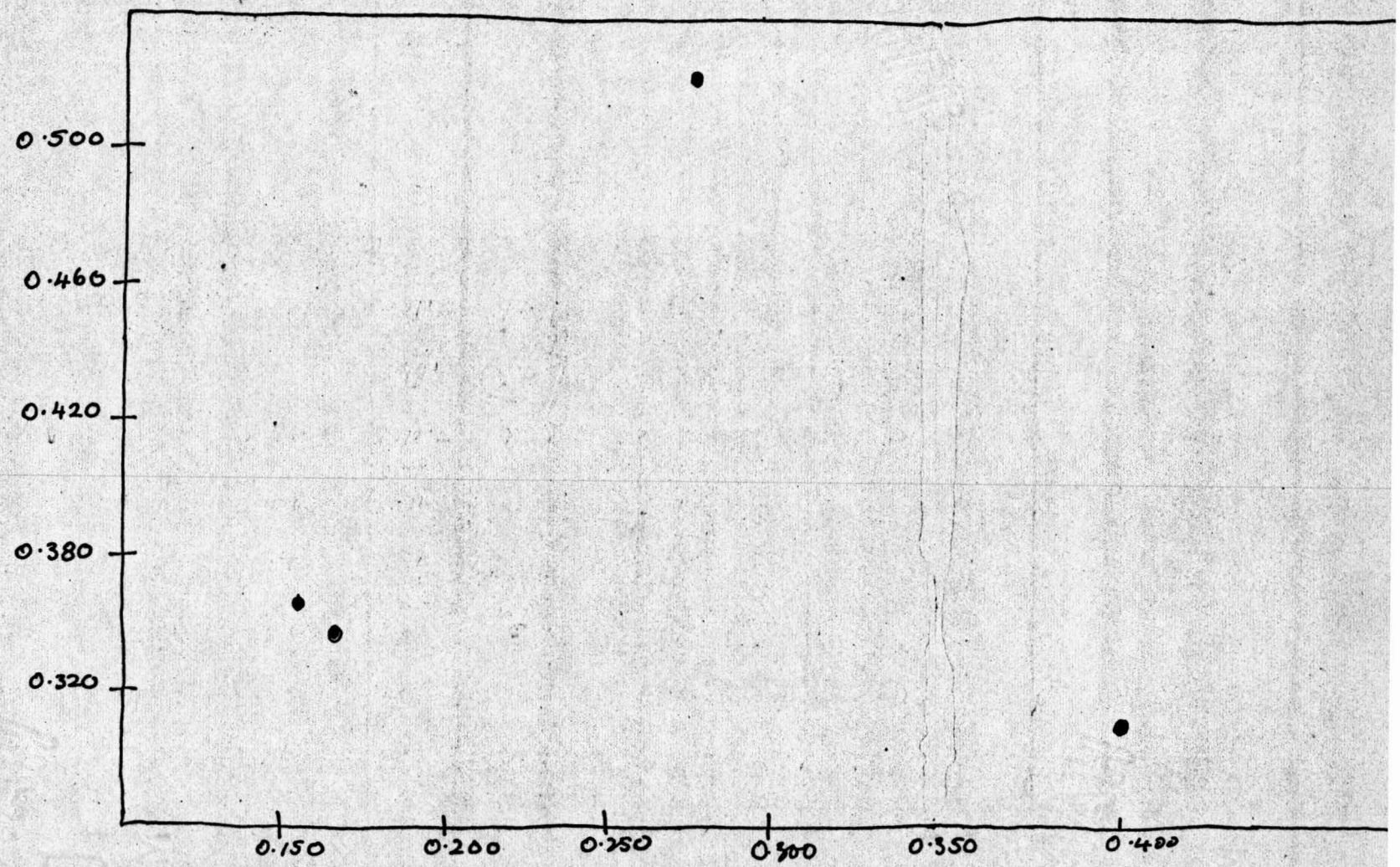
| | | |
|---|---|---|
| 17773.14448 | 8916.65952 | 763.11190 |
| 8916.65952 | 4986.78728 | 370.83890 |
| 763.11190 | 370.83890 | 59.55860 |

| | | |
|---|---|---|
| SOLUTION OF GENERALIZED DETERMINANTAL EQUATION | | |
| 22383.9561728 | 409.1178368 | 26.4152414 |
| 0.8889501 | -0.4545612 | -0.0560515 |
| 0.4564274 | 0.8893765 | 0.0261398 |
| 0.0379687 | -0.0488204 | 0.9980856 |
| SPUR OF B(-1/2)AB(-1/2) = | 1.3958787 | |
| 0.9536221 | 0.4397442 | 0.0025123 |
| EIGENVALUE 1 | 0.9536221 | |
| EIGENVALUE 2 | 0.4397442 | |
| EIGENVALUE 3 | 0.0025123 | |
| TRACE OF EIGENVALUES = | 1.3958786 | |
| EIGENVECTORS | | |
| -0.0059098 | -0.0069884 | -0.0239902 |
| 0.0000888 | 0.0277070 | 0.0346539 |
| 0.1855064 | -0.0300519 | 0.0490131 |
| TEST FOR Z(I,J) | | |
| -0.0059098 | -0.0069884 | -0.0239902 |
| 0.0000888 | 0.0277070 | 0.0346539 |
| 0.1855064 | -0.0300519 | 0.0490131 |
| CANONICAL VARIATES | | |
| -0.0318414 | -0.1685222 | -0.3711192 |
| 0.0004786 | 0.6681476 | 0.5360817 |
| 0.9994928 | -0.7246924 | 0.7582129 |

## Classification and canonical variates

From the foregoing discussion, it may have become clear, that the method of canonical variates would seem to be well suited to the problem of classification, as occurring in numerical taxonomy The NT experts claim, that canonical variates have no physical significance, like the indices and categories they advocate and a classification achieved by means of CVA cannot be explained or interpreted in familiar terms.

CANONICAL VARIATE PLOT



CANONICAL VARIATE MEAN NO. 1

| | | |
|---|---|---|
| 0.19605 | 0.35036 | -0.21175 |

CANONICAL VARIATE MEAN NO. 2

| | | |
|---|---|---|
| 0.18623 | 0.35423 | -0.22451 |

CANONICAL VARIATE MEAN NO. 3

| | | |
|---|---|---|
| 0.27831 | 0.48997 | -0.21940 |

CANONICAL VARIATE MEAN NO. 4

| | | |
|---|---|---|
| 0.39529 | 0.33875 | -0.21857 |

TESTS OF SIGNIFICANCE

LAMBDA NO 1

-3.6528100

CHISQUARE = 445.6428160

LAMBDA NO 2

-0.5818772

CHISQUARE = 70.9890232

LAMBDA NO 3

-0.0025155

CHISQUARE = 0.3068852

Any difference between individuals may be described of SIZE and SHAPE in biometric connexions. These are, however, not easy to define in statistical-mathematical terms, and there is no satisfactorily objective way of measuring these.

Effective number of dimensions

In general, in CVA, the EFFECTIVE NUMBER OF DIMENSIONS cannot be more than the minimum number of variables and ONE LESS than the NUMBER OF GROUPS.

This reciprocity between the variables and the degrees of freedom of the groups can affect the way in which the analysis is carried out. With only two groups, and thus one degree of freedom representing the contrast between the groups, it may be convenient to introduce a pseudovariable to represent this contrast, and carry out the analysis on such a variable. With several groups, this method is naturally possible, although it is not the best way available, particularly, if the number of variables is less than the number of group contrasts.

Usually, the standard analysis of variance and covariance technique is preferable to separate the degrees of freedom for the group contrasts. In terms of the general statistical problem involved, the CVA technique is one of correlating one set of p variables with another set of q variables, and this may, of course, be done, starting from either set. We shall run into this concept again in the Chapter dealing with the subject of canonical correlation.

Applications of the generalized eigenvalue problem

In mathematical terms, we may look at the foregoing procedure as a generalization of the standard eigenvalue problem, in the form with which we have become familiar in connexion with the method of principal component analysis. We shall briefly refer to this in non-statistical terms in order to put our

discussion on a par with the previous sections. In particular, in Applied Mathematics (for example, Oscillation Theory, the theory of Differential Equations), we have the generalized eigenvalue problem:

$$Ax = \lambda Bx; \quad (A - \lambda B)x = 0,$$

with two, n-rowed square matrices, A and B. This problem is sometimes known as the eigenvalue procedure for the matrix pair A and B. The nontrivial solution of the equation is in terms of the generalized characteristic equation:

$$|A - \lambda B| = 0.$$

This is an algebraic equation in $\lambda$, the absolute member of which equals $|A|$ and the coefficients of $\lambda^n$ equal $|B|$. Hence, it is an equation of the n-th order in $\lambda$, when B is nonsingular. In this case, the generalized condition may be put in terms of the special by multiplying through with $B^{-1}$, as follows:

$$B^{-1}Ax = Sx = \lambda x.$$

Hence, it is possible to carry out the solution in terms of matrix S and there are n roots $\lambda_i$ for the characteristic equation.
This point is of considerable importance for computer applications, because it allows one to produce a readily programmable procedure for producing the eigenvalues and eigenvectors of the generalized charactersitic equation in this special case.
In other situations, where this particular structure does not pertain, the methods of solution are vastly more tedious indeed.

What happens if B is singular? In this case, the characteristic equation will lack the highest power $\lambda^n$ and, in certain situations, also some of the lower powers.

If B is singular, but A is nonsingular, the transformation, $\mu = 1/\lambda$, allows us to treat the problem in the form.-

# CHAPTER VI
# DISCRIMINANT FUNCTION ANALYSIS

We shall now be concerned with the classical topic of discriminant functions. These are among the most widely talked about sections of layman statistics, superceded only by factor analysis; although the roots of the concept are predated by Hotelling's researches in 1930, R. Fisher, by the mode of treatment of the biologic problem with he was confronted, gained the support of popular fantasy. The discriminan function was invented in order to solve a problem of identification, which may be put into the following terms. Given a specimen, or a homogeneous sample of specimens, deriving from one of k populations, it is desired, with a minimum risk of error, to identify the specimen or sample of specimens with one of these k populations. The two principal basic points connected herewith are:

1) —That the specimen or sample actually does belong to one of the populations:

2) That the covariance matrices of the k populations are, in statistical terms, equal.

I shall not here take up points referred to in my article in Computer Contributions 7, which are of a general or introductory nature, but will expect the seminar participant to be familiar with the underlying points raised therein. I refer in particular to the topics reviewed on page 5 of this paper.

The problem of classification may be considered as a problem of statistical decision functions. There are a number of hypotheses: each of these is that the distribution of the observation is a given one. One of these must be accepted, the others rejected.

For only two populations the problem is elementary and requires only testing the hypothesis of a specified distribution against another.

In constructing a procedure of classification, it is necessary to minimize the probability of misidentification, that is, it is desirable to minimize, on the average, the bad effects of misidentification.

We shall consider the simple case posed by the two-populational situation.

Suppose we have a specimen, represented by a vector of p, hopefully diagnostic, observations, and it is known to derive from either of the populations $\pi_1$ or $\pi_2$. We may set up an identificationary rule such that if the specimen is characterized by certain sets of values of $x_1, \ldots, x_p$, it will be identifie with $\pi_1$, but if it has other values we shall regard it as having come from $\pi_2$.
It is natural to think of the vector of observations typifying our specimen as a point in p-space. Consider this hyperspace divided intwo two regions: If the specimen falls into region $R_1$, it is identified as belonging to $\pi_1$ and if it falls into the region of space $R_2$ it is identified as coming from $\pi_2$.

In this given elementary situation it is logically only possible for our geostatistician to make two kinds of errors in identification. Firstly, that of putting a specimen from (1) into (2) and conversely. The geostatistician will be concerned to a considerable extent with the cost of misidentification = how bad it will be if his procedure is wrong. We may express the cost of the first type error as $C(2|1) > 0$ and that of the second type error as $C(1|2) (>0)$, which costs may be measured in any kinds of units, for, it is the ratio of the costs that turns out to be important and not their actual magnitudes on their own.

A good classification procedure will, therefore, be one that

minimizes the costs of misidentification.

Let the probability that an observation comes from population $\pi_1$ be $q_1$ and from population $\pi_2$ be $q_2$. The probability properties of the first population are specified by a distribution function and likewise for the second population. Since the probability of drawing a specimen from $\pi_1$ is $q_1$, the probability of drawing a specimen from this population and correctly identifiying it is $q_1P(1|1,R)$. Similarly, the probability of misidentifying the specimen in this situation is $q_1P(2|1,R)$, and likewise for the second population we have the probabilities; $q_2P(1|2,R)$ and $q_2P(2|2,R)$.

The statistician will also want to know that the average or expected loss from costs of misidentification is? It turns out to be the sum of the products of costs of each misidentification, multiplied by the probability of its occurrence: Thus,

$$C(2|1)P(2|1,R)q_1 + C(1|2)P(1|2,R)q_2. \qquad (VI:1)$$

It is this average that one desires to minimize. In other words, one wishes to divide the space into regions $R_1$ and $R_2$, such that the expected loss is as small as possible. A procedure that minimizes (1) for a given $q_1$ and $q_2$ is called a BAYES procedure. You should keep the notion of this in mind as Bayesian estimation procedures happens to be a term that is receiving the etiquette "catchy" presently, in computer-geologic circles.

Let us now regard the situation when we do not know anything about a priori probabilities. In this situation, the expected loss if the observation is from $\pi_1$ is

$$C(2|1)P(2|1,R) = r(1,R);$$

the expected loss if the observation is from $\pi_2$ is

$$C(1|2)P(1|2,R) = r(2,R).$$

For the treatment of this situation, one considers a procedure of identification, R, which is termed admissible if there is no other procedure which is better than R. If one considers the "entire class of admissible procedures" it may be shown, that under certain conditions, this class is the same as the class of Bayesian procedures. If every procedure outside of a class of procedures is less good than at least one procedure within the class, the class is said to be"complete" The class is said to be "essentially complete" if at least one procedure within it is at least as good as a procedure occurring outside the class. The class is said to be a "minimal complete class" if it is a complete class such that no proper subset is a complete class. Under certain conditions, the admissible class may be shown to be minimal complete.

A principle that usually leads to a unique procedure is the minimax principle, which is a principle that makes the maximum expected loss, r(i.R) minimum.

We shall first briefly consider the Bayesian example, that is, the case occurring when the probabilities are known.
For a given, observed specimen, from p-space, one minimizes the probability of misidentification by assigning it to the population with the higher conditional probability.
In symbols, then, if

$$\frac{q_1 p_1(x)}{q_1 p_1(x) + q_2 p_2(x)} + \geq \frac{q_2 p_2(x)}{q_1 p_1(x) + q_2 p_2(x)}$$

one chooses the first population. Otherwise, the second population is the one selected for assignation of the specimen. If the two are equal, then the specimen could be equally as well be identified with either of the populations.

In such a case, practicially speaking, one may wish to consider another choice of variables to be measured on the specimen, etc.

The foregoing may be put in the form of a theorem which goes as follows:

If $q_1$ and $q_2$ are a priori probabilities of drawing an observation from population $\pi_1$ with density $p_1(x)$ and $\pi_2$ with density $p_2(x)$ respectively, and if the cost of misclassifying an observation from $\pi_1$ as from $\pi_2$ is $C(2|1)$ and an observation from $\pi_2$ as from $\pi_1$ is $C(1|2)$, then the regions of classification $R_1$ and $R_2$, defined as follows:

$$R_1: \quad \frac{p_1(x)}{p_2(x)} + \geq \quad \frac{C(1|2)q_2}{C(2|1)q_1},$$

$$R_2: \quad \frac{p_1(x)}{p_2(x)} < \quad \frac{C(1|2)q_2}{C(2|1)q_1}.$$

minimize the expected cost of misidentification.

In the case where the statistician does not know anything about a priori probabilities, it is necessary to seek for the class of admissible procedures, notably, that set of procedures upon which improvement is not possible. It may be shown that a Bayes procedure is admissible. In the form of a theorem this may be stated as: If Probability$\{p_2(x) = 0|\pi_1\} = 0$ and Probability$\{p_1(x) = 0|\pi_2\} = 0$, then every Bayes procedure is admissible. It is also readily possible to prove the coverse of this theorem, i.e., that every admissible procedure is a Bayes procedure. The proof of this theorem shows that the class of Bayes procedures is complete and it may also be shown, moreover that the class of Bayes procedures is minimal complete since it is identical with the class of admissible procedures.

We shall now use the foregoing discussion to review the philosophy of "classification" = identification by means of the technique of discriminant function analysis (hereinafter referred to as DCA).

Consider two multivariate normal populations with equal covariance matrices:

$N(\mu^{(1)},\Sigma)$ and $N(\mu^{(2)},\Sigma)$. We approach the problem by considering the i-th density (i=1,2) expression, which may be expressed as

$$p_i(x) = \frac{1}{(2\pi)^{\sqrt{p}}|\Sigma|^{\sqrt{2}}} \exp[-\tfrac{1}{2}(x-\mu^{(i)})'\Sigma^{-1}(x-\mu^{(i)})]. \quad \text{(VI:2)}$$

The next step is to examine the ratio of the densities for i=1,2, which results in the following expression:

$$\frac{p_1(x)}{p_2(x)} = \frac{\exp[-\tfrac{1}{2}(x-\mu^{(1)})'\Sigma^{-1}(x-\mu^{(1)})]}{\exp[-\tfrac{1}{2}(x-\mu^{(2)})'\Sigma^{-1}(x-\mu^{(2)})[} \quad \text{(VI:3)}$$

The region of classification into Population (1), $R_1$, is the set of values of x for which equation (VI:3) is greater than, or equal to, a specially chosen value, say, k.
This inequality may be written in terms of the logarithm of (VI:3) (as the logarithmic function is monotonic increasing).
Thus,

$$-\tfrac{1}{2}[x-\mu^{(1)})'\Sigma^{-1}(x-\mu^{(1)} - (x-\mu^{(2)})'\Sigma^{-1}(x-\mu^{(2)}) \geq \log k. \quad \text{(VI:4)}$$

The left hand side of this equation may be rearranged and expanded to yield the ensuing expression:

$$x'\Sigma^{-1}(\mu^{(1)} - \mu^{(2)}) - \tfrac{1}{2}(\mu^{(1)} + \mu^{(2)})'\Sigma^{-1}(\mu^{(1)} - \mu^{(2)}). \quad \text{(VI:5)}$$

The first term of this expression is the discriminant function of Fisher.
Relating, now, this expression, which includes the discriminant function, to the problem of identification, we have the following:

If the i-th population possesses the density given in (VI:2) the best regions of classification are yielded by:

$$R_1: \text{Expression (4)} \geq \log_e k$$
$$R_2: \text{Expression (4)} < \log_e k. \quad \text{(VI:6)}$$

If we happen to know the a priori probabilities, $q_1$ and $q_2$, then the value of "k" is given by:

$$k = \frac{q_2 C(1|2)}{q_1 C(2|1)} .$$

In very many classification computations, one will assume that the two populations are equally likely to be correct, hence, $q_1 = q_2$ and the costs are equal. In my experience this is usually not a good assumption, and I prefer to allow some form of weighting, if the available information is based on samples of different size. In the case of equal probabilities, k becomes unity, hence $\log_e k = 0$.

If there are no a priori probabilities, one may select $\log_e k = c$, say, on the basis of making the expected losses due to misclassification equal. If X is a random observation, distributed in accord with the multivariate Normal, then replacing x by X in (VI:5) and calling this U, say, it may be shown that U is normally distributed with mean,

$$E_1 U = \tfrac{1}{2}(\mu^{(1)} - \mu^{(2)})'\Sigma^{-1}(\mu^{(1)} - \mu^{(2)})$$

and variance,

$$\mathrm{Var}_1(U) = (\mu^{(1)} - \mu^{(2)})'\Sigma^{-1}(\mu^{(1)} - \mu^{(2)}) \qquad \text{(VI:7)}$$

This expression for the variance of U will be seen to be the Mahalanobis' generalized distance. This is a way of defining the "distance" between our two populations.

The methods of computing the linear discriminant function, as outlined in the foregoing, is well known, and programs are available in the KU Program Library.

We shall now take a look at a pitfall that may occur in conjunction with DFA.

## The case of $\Sigma_1 \neq \Sigma_2$ - a possible pitfall

There are various possibilities in connexion herewith of solving the problem. We shall consider two of these:

It is possible to define a statistical distance term by means of an expression of the following kind, as has been done by T.W. Anderson and Bahadur:

$$\frac{b'd}{(b'S_1b)^{\frac{1}{2}} + (b'S_2b)^{\frac{1}{2}}}, \qquad \text{(VI:8)}$$

where d is the difference vector between sample mean vectors, $S_1$ and $S_2$ are sample covariance matrices, and b is a vector which has to be estimated by some method or other. It is of the following form:

$$|tS_1 + (1-t)S_2|^{-1}d, \quad 0<t<1. \qquad \text{(VI:9)}$$

the vector b has to be estimated by means of iteratively obtained values of the scalar t. Vector b, when obtained, is an analog of the discriminant function coefficients in the case of equal covariance matrices. When $S_1 = S_2$, twice the maximum of (8) is the normal generalized statistical distance between samples. The estimation is made iteratively by finding t from (after an initial guess at t):

$$b'[t^2S_1 - (1-t)^2]b = 0 \qquad \text{(VI:10)}$$

and solving

$$(tS_1 + (1-t)S_2)b = d. \qquad \text{(VI:11)}$$

for vector b.

This procedure has been programmed by the writer and is available in the library of programs of the Kansas Geological Survey.

This method is a good one, providing computational facilities are available, but for machine computation, the amount of work is considerable, involving as it does, a matrix inverse for each iteration.

If it is desired to test the resulting distance for significanc

this approach is not satisfactory, as the statistic yielded by this procedure is not related to a test. If it is required to test the significnace in the case of unequal covariance matrices, it is necessary to employ the method devised by the writer. These topics will receive treatment in the next chapter.

Normally, there will not be any point in carrying out a DFA study, if the means of the two groups are not significantly different. Therefore, logically, the calculation of the Hotelling $T^2$ should precede the proposed DFA, and be the decision-making point for this. However, with respect to the actual calculating sequence, it is just as well to output the discriminant coefficients, whether or not the distance value is significant. Again, this is a topic for the next chapter.

## Discriminant functions for more than two populations

By and large, the theory pertaining here is nought but an outgrowth of the foregoing. We can approach this problem in two ways, from the point of view of the calculationary procedures. For the sake of convenience, I shall refer to the two models as the (a) matrix inversion model , and (b) the generalized determinantal equation model.

In my experience, it is model (b), that is the most widely employed one in computer applications, while model (a) is the concept most likely to be found in textbook discussions of the technique.

### (a) The Matrix Inversion Model

Let $P_1$, ....., $P_m$ be m populations with density functions

$p_1(x). \ . \ . \ ., p_m(x)$ respectively. It is now desired to divide this space into m mutually exclusive regions $R_1, \ . \ . \ . \ ., R_m$. I do not propose going through the arguments connected with the development of the identifacotry procedure, as it is virtually the same as for the two-populational case already studied in some detail.

We shall for our purposes assume that the costs of misclassification are equal. We shall use the functions:

$$u_{jk}(x) = \log \frac{p_j(x)}{p_k(x)} = [x - \tfrac{1}{2}(\mu^{(j)} + \mu^{(k)}]'\Sigma^{-1}(\mu^{(j)} - \mu^{(k)}) \qquad \text{(VI:12)}$$

If a priori probabilities are known, the region $R_j$ is defined by those x satisfying:

$$R_j: \ u_{jk}(x) > \log_e \frac{q_k}{q_j}, \ k = 1, \ . \ . \ . \ ., m; \ k \neq j. \qquad \text{(VI:13)}$$

If q is then the a priori probability of drawing an observation from one of our populations, then the regions of classification, $R_1, \ . \ . \ . \ . \ . \ ., R_m$, that minimize the expected costs of misidentification are defined by the inequality (VI:13), where $u_{jk}(x)$ is given by equation (VI:12).

The probabilities of correct classification

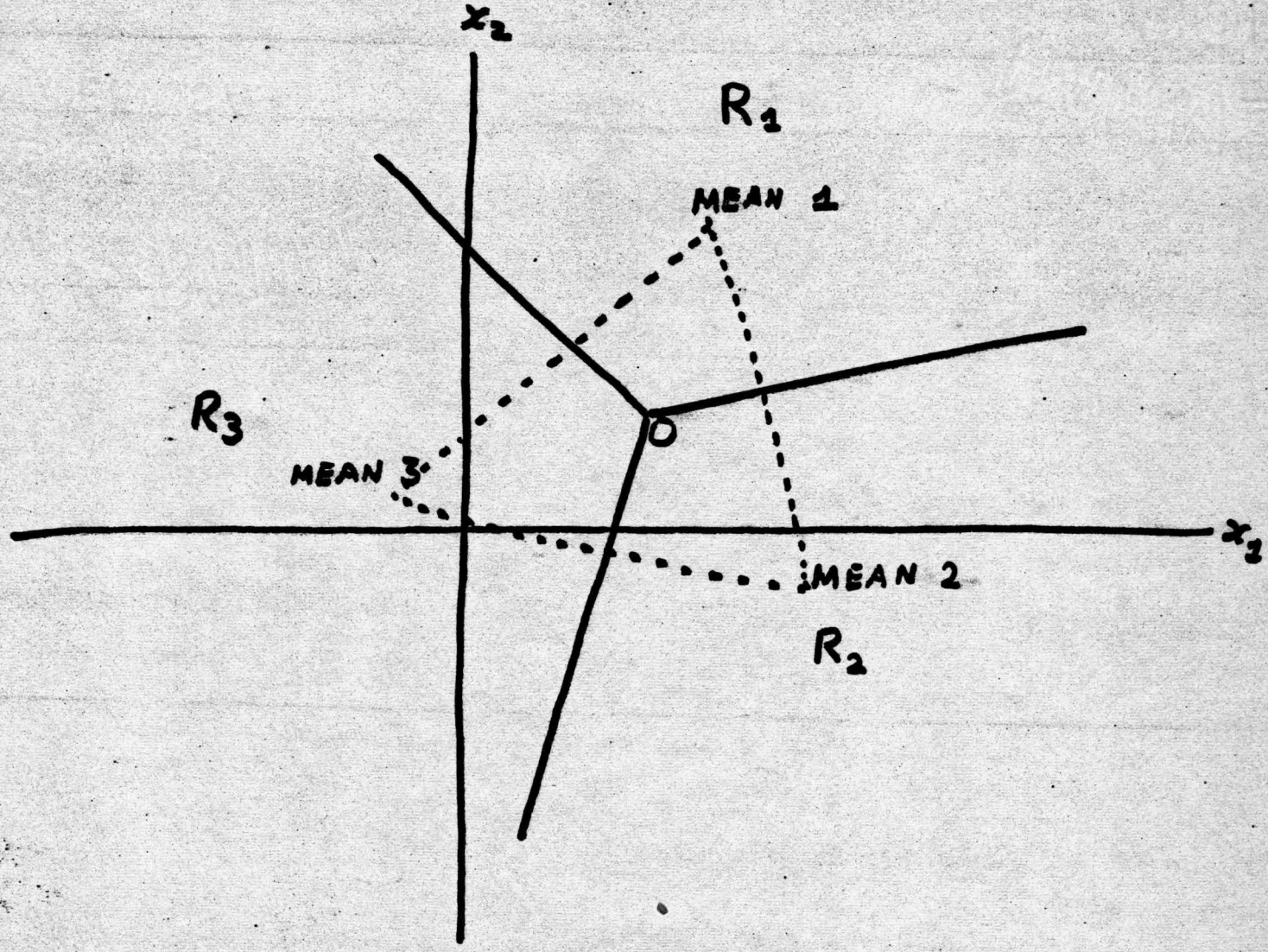
If X is a random observation, let us consider the random variables:

$$U_{ji} = [X - \tfrac{1}{2}(\mu^{(i)} + \mu^{(j)})]'\Sigma^{-1}(\mu^{(j)} - \mu^{(i)}) \qquad \text{(VI:14)}$$

Here, $U_{ji} = -U_{ij}$. It is necessary to use m(m-1)/2 identification functions if the means span an (m - 1)-dimensional space.

For the case of m=3 and number of variables = p = 2,

Fig. 2



The regions $R_1$, $R_2$, $R_3$ are marked out by the three lines meeting at O. If the minimax procedure is valid here, these lines cannot be shifted to any other position and still retain the equality:

$$P(1|1,R) = P(2|2,R) = P(3|3,R).$$

This equality of probabilities uniquely determines

$$u_{jk}(x) \geq c_j - c_k, \quad k = 1,\ldots,m, k \neq j. \qquad \text{(VI:15)}$$

The situation is shown in Fig. 2.

To do this in an actual problem in which we have the numerical values of the components of the mean vectors and the covariance matrix, one would consider the three ($\leq p+1$) joint distributions each of $2U_{ij}$ ($j \neq 1$). One might begin by trying the values of $c_i = 0$, and using the tables published by Pearson in 1931 in the Biometrika series of the bivariate normal distribution, compute $P(i|i,R)$. By trial and error, a value of $c_i$ is finally found which approximates to the above condition. The above theory may be applied to the case where all parameters are estimated, providing the sample sizes are large.

## Multiple Discriminant Analysis: Determinantal Equation Model

This is another procedure occurring in the literature and, as already observed earlier on in this chapter, it is one that is likely to be encountered in computer work, as it is often more economic to program than the first model. Here we use the equation:

$$|W^{-1}A - \lambda I| = 0. \qquad \text{(VI:16)}$$

Here, I is the identity matrix, W is the pooled within cross-products matrix, A is the between cross products matrix. Thus, if T is the matrix of deviations from the grand mean:

$$T = A + W.$$

The required discriminant functions come from the solution of the equation

$$(W^{-1}A - \lambda I)v = 0 \qquad \text{(VI:17)}$$

Equation (VI:17) derives from the maximization of the ratio:

$$\lambda = \frac{v_i' A v_i}{v_i' W v_i}, \qquad \text{(VI:18)}$$

where i = 1, . . ., p (p is the number of variables). If g denotes the number of groups, and g-1>p, there will be p eigenvalues; if g-1<p, there will be g-1 eigenvalues. Equation (VI:18) also indicates a further useful point, notably, that the eigenvalues of A and W are the same.

It is often stated (e.g., Cooley and Lohnes, 1962), that the magnitude of $\lambda_i$ shows how good the particular "functions", with corresponding eigenvector elements as coefficients, is at distinguishing the g groups. This is, however, not necessarily true (Kullback, 1959).

The Wilk's lambda-criterion for testing the significance of eigenvalues is:

$$\Lambda = \prod_{i=1}^{r} \left[ \frac{1}{(1+\lambda_i)} \right]. \qquad \text{(VI:19)}$$

As regards which of the two "discriminator models" is the "right" one, it may be mentioned, that under certain conditions, model "2" converts algebraically to model "1". In effect, one could compare the two by saying that the second model attempts a stepwise approach tothe problem. Instead of using the entire covariance matrix, it brings about a stepwise breakup of the matrix in the hope, that one of the linear combinations will be more effective than a "blanket" discriminator in separating the groups from the point of view of emphasizing their unlikeness to the maximum.

With these few remarks, I think I have presented the main features of the concept of multivariate identification. The subject is, however, one that has attracted much attention among mathematicians and the corpus of pertinent literature is large indeed. There are numerous developments of the theory, some of which involve exciting new slants on the problem, but in all of these, the level of sophistication of the mathematics exceeds that usually encountered in statistics.

Example VI.1

As a first example of the calculations we shall consider a trivariate study of an ecologic-sedimentologic problem. pH, Eh and free oxygen have been measured on the interstitial pore water of nearshore sediments from the Ivory Coast and the Niger Delta. The first step ih the calculations was to ascertain whether the covariance matrices of the samples differ sufficiently to invalidate the usual method of calculating the discriminant function. The results of these computations are shown below:

SECTION (1). Homogeneity of covariance matrices.

| Niger Delta ($N=37$) | | | Ivory Coast ($N=20$) | | |
|---|---|---|---|---|---|
| pH | Eh | [O] | pH | Eh | [O] |
| 7.329 | -140.000 | 6.000 | 7.610 | -188.000 | 58.000 |
| 7.740 | -150.000 | 39.000 | 8.110 | -208.000 | 24.000 |
| 7.690 | 80.000 | 50.000 | 7.460 | -180.000 | 20.000 |
| 7.830 | 160.000 | 38.000 | 7.790 | 121.000 | 58.000 |
| 7.308 | -135.000 | 42.000 | 7.420 | -167.000 | 70.000 |
| 7.460 | -230.000 | 40.000 | 7.280 | -167.000 | 30.000 |
| 7.300 | -210.000 | 40.000 | 7.420 | -156.000 | 32.000 |
| 7.510 | -210.000 | 48.000 | 7.410 | -80.000 | 26.000 |
| 7.480 | -171.000 | 52.000 | 7.410 | -90.000 | 32.000 |
| 7.514 | -155.000 | 32.000 | 7.510 | -115.000 | 72.000 |
| 7.438 | -200.000 | 32.000 | 7.450 | -166.000 | 42.000 |
| 7.557 | -175.000 | 42.000 | 7.690 | 45.000 | 49.000 |
| 7.474 | -195.000 | 32.000 | 7.930 | 93.000 | 72.000 |
| 7.600 | -206.000 | 80.000 | 7.880 | 159.000 | 77.000 |
| 7.390 | 120.000 | 48.200 | 7.680 | -142.000 | 52.000 |
| 7.670 | -40.000 | 20.000 | 7.760 | 65.000 | 75.000 |
| 7.480 | -162.000 | 28.000 | 7.460 | -95.000 | 69.000 |
| 7.606 | -85.000 | 48.000 | | | |
| 7.300 | -205.000 | 84.000 | 7.500 | -120.000 | 68.000 |
| 7.365 | -230.000 | 19.000 | 7.610 | -155.000 | 61.000 |
| 7.540 | -158.000 | 25.000 | 7.430 | -180.000 | 46.000 |
| 7.520 | -163.000 | 38.000 | | | |
| 7.425 | -112.000 | 0.000 | | | |
| 7.700 | -189.000 | 31.000 | | | |
| 7.443 | -139.000 | 41.000 | | | |
| 7.510 | -110.000 | 42.000 | | | |
| 7.594 | -200.000 | 59.000 | | | |
| 7.524 | -100.000 | 23.000 | | | |
| 7.630 | -198.000 | 71.000 | | | |
| 7.560 | -218.000 | 32.000 | | | |
| 7.560 | -196.000 | 41.000 | | | |
| 7.650 | -177.000 | 23.000 | | | |
| 7.518 | -172.000 | 46.000 | | | |
| 7.560 | -203.000 | 53.000 | | | |
| 7.294 | -219.000 | 66.000 | | | |
| 7.668 | -135.000 | 43.000 | | | |
| 7.554 | -205.000 | 30.000 | | | |

The calculations were made in accordance with a previous chapter and give, for the information-theoretic equation, $\chi^2 = 11.9$, which for 6 degrees of freedom is not significant. We shall go ahead as though there is nothing unusual in the data. As a matter of fact, some further tests on the covariance matrices suggest that larger samples might have shown up the existence of a certain degree of unlikeness in the covariance matrices.

SECTION (2) The discriminant functions.

| MEAN VECTOR 1 | MEAN VECTOR 2 | MEAN VECTOR OF DIFFERENCES ADAPTED |
|---|---|---|
| 7.5213776 | 7.5904994 | -0.0691218 |
| -146.8378368 | -86.2999992 | -60.5378376 |
| 40.1135128 | 51.6499996 | -11.5364866 |

COVARIANCE MATRIX 1

| | | |
|---|---|---|
| 0.0168720 | 4.0870768 | 0.0045641 |
| 4.0870768 | 8363.8063104 | -172.5078656 |
| 0.0045675 | -172.5079184 | 310.8545440 |

COVARIANCE MATRIX 2

| | | |
|---|---|---|
| 0.0457997 | 12.4006797 | 1.1217780 |
| 12.4006732 | 13278.1158400 | 1102.6788608 |
| 1.1217780 | 1102.6788608 | 362.4500384 |

INVERSE OF POOLED COVARIANCE MATRIX

| | | |
|---|---|---|
| 45.6182420 | -0.0307767 | -0.0291041 |
| -0.0307767 | 0.0001224 | -0.0000632 |
| -0.0291045 | -0.0000632 | 0.0031286 |

DISCRIMINATOR COEFFICIENTS

| | | |
|---|---|---|
| -0.954301 | -0.004551 | -0.030255 |

| D | D SQUARE |
|---|---|
| 0.8309612 | 0.6904965 |

SIGNIFICANCE FOR D SQUARE AND T SQUARE

T SQUARE = 8.9643410

F = 2.8794550

These results, for 3 and 53 degrees of freedom, are below the level of significance which I believe should be adopted in this connexion, notably the 1% level, and there is no strong evidence, theat the two groups are different with respect to the multivariate test used. This example is also the one used in the next chapter. Note, that the $T^2$ is the statistie devised by Hotelling in 1931.

## Example VI(2)

This example is concerned with the analysis of POTTER's data. It deals with 15 variables, expressed in uncorrected proportions and refers to two samples of shales. It is desired to set up a discriminant function that will be able to distinguish between marine and freshwater shales. For a discussion of the various chemical variables, I refer to Potter's published paper.

| N1 = 39 MEAN VECTOR 1 | N2 = 43 MEAN VECTOR 2 | DIFFERENCE MEAN VECTOR |
|---|---|---|
| 111.7692288 | 62.6511624 | 49.1180672 |
| 72.4102560 | 65.4418600 | 6.9683961 |
| 13.1282050 | 12.0232557 | 1.1049494 |
| 40.1282048 | 29.9534880 | 10.1747164 |
| 83.1282040 | 98.5581384 | -15.4299344 |
| 36.2948696 | 26.1279044 | 10.1669654 |
| 20.6538448 | 15.1232540 | 5.5305907 |
| 17.4358972 | 29.5348832 | -12.0989861 |
| 14.1025640 | 8.8372092 | 5.2653549 |
| 33.7179484 | 33.9534880 | -0.2355399 |
| 32.8205124 | 21.1627904 | 11.6577222 |
| 1.9230769 | 6.5116279 | -4.5885509 |
| 3.8212816 | 2.0027905 | 1.8184911 |
| 8.6646144 | 4.5553482 | 4.1092662 |
| 1.8099999 | 1.7541859 | 0.0558139 |

COVARIANCE MATRIX 1

| | | | | | | |
|---|---|---|---|---|---|---|
| 979.1297152 | 696.2288192 | 37.9251124 | 530.6619968 | 737.1621056 | 448.0698432 | 220.5206336 |
| 7.9453638 | -41.7509764 | -51.4371912 | -8.7550562 | 29.9003196 | 68.6660656 | -7.4523716 |
| 696.2288192 | 1253.4588672 | 16.4986506 | 775.4197120 | 578.5776000 | 250.2995200 | 86.2825600 |
| 42.4831412 | -26.5654296 | -80.3981024 | -36.3360316 | 51.3423428 | 117.7980112 | -6.1626362 |
| 37.9251124 | 16.4986634 | 13.9568158 | 3.9831350 | 38.3252460 | 29.7085728 | 19.3639556 |
| -12.7766513 | -8.1207789 | 13.5762552 | -3.0161942 | -4.5070098 | -10.1429723 | 1.1318420 |
| 530.6619968 | 775.4197376 | 3.9831350 | 897.4305152 | 332.9305312 | 284.3848256 | 100.7744640 |
| 19.7233564 | -46.0154960 | -2.7395276 | -22.0951408 | 66.2982440 | 151.1035680 | -2.2763182 |
| 737.1621056 | 578.5776576 | 38.3252460 | 332.9305312 | 876.8516608 | 382.5217120 | 227.2455008 |
| -44.3555840 | -1.8049445 | -18.5289882 | -33.0161916 | -5.1249004 | -11.1003481 | -8.8613152 |
| 448.0698432 | 250.2995456 | 29.7085728 | 284.3848256 | 382.5217120 | 329.4111040 | 171.7276608 |
| -64.9652408 | -64.5988376 | 70.5938136 | -28.5293480 | 18.4749816 | 42.0773892 | -1.9926300 |
| 220.5206336 | 86.2825720 | 19.3639556 | 100.7744640 | 227.2455008 | 171.7276608 | 128.6846832 |
| -58.5425152 | -70.4949560 | 85.5941088 | -15.4483782 | -2.8709074 | -6.4047450 | 1.7903438 |
| 93.9979952 | 100.8164824 | 10.3373830 | 51.1268628 | 97.7058104 | 87.4996712 | 58.8916912 |
| -41.1774608 | -32.3211792 | -33.3670652 | -10.0708496 | -3.0768874 | -6.9378501 | -1.3460529 |
| 7.9453510 | 42.4831412 | -12.7766530 | 19.7233500 | -44.3555968 | -64.9652480 | -58.5425152 |
| 74.8313088 | 42.8981160 | -103.3232096 | 26.7712550 | 6.8340759 | 15.3845118 | -3.6368420 |
| -41.7509764 | -26.5654296 | -8.1207789 | -46.0154960 | -1.8049445 | -64.5988376 | -70.4949560 |
| 42.8981160 | 127.2604720 | -152.2098304 | 14.3724700 | -6.0548915 | -13.6715600 | -5.2460520 |
| -51.4371912 | -80.3981024 | 13.5762488 | -2.7395276 | -18.5289882 | 70.5938000 | 85.5941088 |
| -103.3232096 | -152.2098432 | 353.6774624 | -64.7773272 | 5.0607637 | 11.4603334 | 11.3157877 |
| -8.7550594 | -36.3360304 | -3.0161942 | -22.0951408 | -33.0161916 | -28.5293480 | -15.4483782 |
| 26.7712554 | 14.3724700 | -64.7773264 | 33.7044528 | -2.7630557 | -6.2354231 | -1.0868419 |
| 29.9003196 | 51.3423428 | -4.5070098 | 66.2982440 | -5.1249004 | 18.4749802 | -2.8709074 |
| 6.8340767 | -6.0548915 | 5.0607637 | -2.7630557 | 15.0271064 | 33.9056424 | -0.4957044 |
| 68.6660536 | 117.7980112 | -10.1429740 | 151.1035648 | -11.1003545 | 42.0773892 | -6.4047450 |
| 15.3845118 | -13.6715600 | 11.4603334 | -6.2354231 | 33.9056424 | 76.5220048 | -1.0939839 |
| -7.4523716 | -6.1626362 | 1.1318417 | -2.2763182 | -8.8613152 | -1.9926300 | 1.7903438 |
| -3.6368420 | -5.2460520 | 11.3157877 | -1.0868419 | -0.4957046 | -1.0939839 | 0.9375106 |

COVARIANCE MATRIX 2

| | | | | | | |
|---|---|---|---|---|---|---|
| 1051.7087744 | -26.6516926 | 32.1987880 | 164.0786352 | -167.0327376 | 447.5931840 | -224.3964224 |
| 75.5371024 | 195.8167296 | -127.4418464 | 16.0160608 | 11.5750397 | 26.3509634 | -3.4939837 |
| -26.6516926 | 912.5382144 | 66.3704416 | 406.3543968 | 1149.3189888 | 62.8183120 | 65.3752312 |
| 28.5022200 | -207.2646480 | 81.2596944 | -52.9457304 | -23.2112568 | -52.7374004 | 8.9440568 |
| 32.1987940 | 66.3704416 | 15.2613524 | 41.4296840 | 124.4867216 | 39.1969568 | 21.5280194 |
| -2.4723133 | -9.3798362 | 12.9485095 | -8.2502760 | -1.3226841 | -3.0051247 | -0.9848613 |
| 164.0786352 | 406.3543712 | 41.4296812 | 243.5692192 | 577.3123072 | 95.1251272 | 65.4534968 |
| 40.1827244 | -26.0022086 | 16.3648970 | -24.9280160 | -5.0896265 | -11.5378329 | 4.8192465 |
| 167.0327376 | 1149.3189888 | 124.4867216 | 577.3123072 | 2481.1097344 | 261.6841024 | 153.4724576 |
| 23.6406598 | -205.5923984 | 121.3593744 | -119.4352080 | -22.6563486 | -51.3523176 | -8.2502498 |
| 447.5931840 | 62.8183120 | 39.1969568 | 95.1251272 | 261.6840800 | 422.3120800 | 208.4688576 |
| 2.3070301 | 2.6608305 | -23.8784584 | -10.8884131 | 6.7133774 | 15.2391922 | -5.9516641 |
| 224.3964352 | 65.3752312 | 21.5280194 | 65.4535024 | 153.4724576 | 208.4688576 | 122.4718496 |
| 7.6824573 | -13.1298187 | -14.7895680 | -6.9050292 | 3.1695073 | 7.1885477 | -2.8248829 |
| -159.9279888 | 150.4485200 | 7.1539364 | -5.6173735 | 180.0277152 | 29.7989904 | 27.1420428 |
| -80.9108472 | -255.2602272 | -139.3272336 | -20.7087458 | -13.6236680 | -31.0188740 | 5.6579437 |
| 75.5371024 | 28.5022200 | -2.4723133 | 40.1827272 | 23.6406598 | 2.3070330 | 7.6824573 |
| 97.4252496 | 45.1827272 | -42.0681032 | -19.6290136 | 4.3330843 | 9.8611315 | 0.6728408 |
| 195.8167296 | -207.2646704 | -9.3798420 | -26.0022086 | -205.5924464 | 2.6608305 | -13.1298246 |
| 45.1827244 | 275.6644544 | -114.2303392 | 48.6434120 | 4.9458473 | 11.2807340 | -0.9967056 |
| -127.4418464 | 81.2597072 | 12.9485095 | 16.3649028 | 121.3593744 | -23.8784584 | -14.7895680 |
| -42.0681032 | -114.2303328 | 430.7585856 | -135.1328864 | 11.4835836 | 26.1138770 | -8.0097445 |
| 16.0160664 | -52.9457304 | -8.2502760 | -24.9280132 | -119.4352080 | -10.8884116 | -6.9050278 |
| -19.6290136 | 48.6434148 | -135.1328864 | 126.8272416 | -7.1388412 | -16.2368452 | 2.6756652 |
| 11.5750411 | -23.2112554 | -1.3226841 | -5.0896257 | -22.6563486 | 6.7133774 | 3.1695073 |
| 4.3330843 | 4.9458480 | 11.4835836 | -7.1388412 | 5.6016061 | 12.7424460 | -0.6929448 |
| 26.3509634 | -52.7374004 | -3.0051255 | -11.5378315 | -51.3523176 | 15.2391922 | 7.1885477 |
| 9.8611311 | 11.2807354 | 26.1138770 | -16.2368454 | 12.7424460 | 28.9863784 | -1.5761217 |
| -3.4939822 | 8.9440568 | -0.9848615 | 4.8192465 | -8.2502498 | -5.9516641 | -2.8248829 |
| 0.6728407 | -0.9967056 | -8.0097445 | 2.6756651 | -0.6929448 | -1.5761217 | 2.4926821 |

POOLED COVARIANCE MATRIX

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| 1017.233704 | 316.716548 | 34.918792 | 338.205728 | 437.844180 | 447.819592 | 222.555420 | -39.313147 | 43.431026 | 82.972068 |
| -91.339633 | 4.249780 | 20.279547 | 46.450636 | -5.374218 | | | | | |
| 316.716548 | 1074.475520 | 42.681341 | 581.660408 | 878.216824 | 151.871882 | 75.306211 | 126.873294 | 35.143157 | -121.432518 |
| 4.472241 | -45.056123 | 12.201703 | 28.266920 | 1.768378 | | | | | |
| 34.918795 | 42.681347 | 14.641697 | 23.642573 | 83.560020 | 34.689974 | 20.500089 | 8.666075 | -7.366874 | -8.781784 |
| 13.246689 | -5.764087 | -2.835239 | -6.395602 | 0.020573 | | | | | |
| 338.205728 | 581.660408 | 23.642571 | 554.153320 | 461.230960 | 185.023482 | 82.230955 | 21.336133 | 30.464524 | -35.508520 |
| 7.290295 | -23.582400 | 28.819612 | 65.716832 | 1.448853 | | | | | |
| 437.844180 | 878.216856 | 83.560020 | 461.230960 | 1719.087120 | 319.081964 | 188.514650 | 140.924814 | -8.657556 | -108.793355 |
| 54.912402 | -78.386174 | -14.328911 | -32.232632 | -8.540506 | | | | | |
| 447.819592 | 151.871894 | 34.689974 | 185.023482 | 319.081952 | 378.184112 | 191.016788 | 57.206808 | -29.647299 | -29.287512 |
| 20.995871 | -19.267857 | 12.300139 | 27.987335 | -4.071123 | | | | | |
| 222.555426 | 75.306217 | 20.500089 | 82.230957 | 188.514650 | 191.016788 | 125.422945 | 42.223127 | -23.774405 | -40.378258 |
| 32.892678 | -10.963120 | 0.300310 | 0.731734 | -0.632650 | | | | | |
| -39.313147 | 126.873301 | 8.666073 | 21.336139 | 140.924808 | 57.206814 | 42.223125 | 316.053592 | -62.037488 | -149.364176 |
| -88.996153 | -15.655745 | -8.613947 | -19.580387 | 2.331045 | | | | | |
| 43.431020 | 35.143157 | -7.366875 | 30.464523 | -8.657562 | -29.647300 | -23.774405 | -62.037488 | 86.693126 | 44.097537 |
| -71.164278 | 2.411114 | 5.521055 | 12.484737 | -1.374258 | | | | | |
| 82.972068 | -121.432529 | -8.781787 | -35.508520 | -108.793380 | -29.287512 | -40.378262 | -149.364182 | 44.097535 | 205.172560 |
| -132.270597 | 32.364714 | -0.279504 | -0.571606 | -3.015145 | | | | | |
| -91.339633 | 4.472247 | 13.246686 | 7.290298 | 54.912402 | 20.995864 | 32.892678 | -88.996153 | -71.164278 | -132.270599 |
| 394.145048 | -101.713994 | 8.432744 | 19.153443 | 1.169883 | | | | | |
| 4.249782 | -45.056122 | -5.764087 | -23.582398 | -78.386174 | -19.267856 | -10.963119 | -15.655745 | 2.411114 | 32.364716 |
| -101.713994 | 82.593915 | -5.060343 | -11.486170 | 0.888474 | | | | | |
| 20.279548 | 12.201704 | -2.835239 | 28.819613 | -14.328911 | 12.300139 | 0.300310 | -8.613947 | 5.521056 | -0.279503 |
| 8.432744 | -5.060343 | 10.078719 | 22.794964 | -0.599256 | | | | | |
| 46.450631 | 28.266920 | -6.395603 | 65.716831 | -32.232635 | 27.987335 | 0.731734 | -19.580388 | 12.484737 | -0.571605 |
| 19.153443 | -11.486170 | 22.794964 | 51.565800 | -1.347106 | | | | | |
| -5.374217 | 1.768378 | 0.020573 | 1.448853 | -8.540506 | -4.071123 | -0.632650 | 2.331045 | -1.374259 | -3.015145 |
| 1.169883 | 0.888474 | -0.599256 | -1.347106 | 1.753976 | | | | | |

INVERSE OF POOLED COVARIANCE MATRIX

| | | | | | | |
|---|---|---|---|---|---|---|
| 0.0030286 | -0.0004704 | 0.0013243 | -0.0001658 | 0.0000150 | -0.0026079 | -0.0021781 |
| 0.0009732 | -0.0007408 | -0.0008673 | 0.0008988 | 0.0003261 | 0.1577346 | -0.0701140 |
| -0.0009788 | | | | | | |
| -0.0004704 | 0.0037256 | -0.0020636 | -0.0030340 | -0.0010630 | 0.0008049 | 0.0017913 |
| 0.0017757 | 0.0013897 | 0.0039044 | 0.0024897 | 0.0023929 | 0.1627174 | -0.0711222 |
| -0.0017913 | | | | | | |
| 0.0013244 | -0.0020636 | 0.1298255 | -0.0008671 | -0.0029313 | -0.0129475 | 0.0026522 |
| -0.0239431 | -0.0224803 | -0.0270365 | -0.0286828 | -0.0228666 | -0.4265303 | 0.2118744 |
| -0.0225927 | | | | | | |
| -0.0001658 | -0.0030340 | -0.0008671 | 0.0064227 | -0.0001045 | -0.0009132 | -0.0008571 |
| 0.0039416 | 0.0033986 | 0.0028283 | 0.0035938 | 0.0028730 | 0.1845624 | -0.0879206 |
| -0.0117169 | | | | | | |
| 0.0000150 | -0.0010630 | -0.0029313 | -0.0001045 | 0.0015008 | -0.0001290 | -0.0008811 |
| -0.0022542 | -0.0022244 | -0.0024393 | -0.0022624 | -0.0015548 | -0.1213554 | 0.0553731 |
| 0.0083519 | | | | | | |
| -0.0026079 | 0.0008049 | -0.0129475 | -0.0009132 | -0.0001290 | 0.0202617 | -0.0232865 |
| 0.0090923 | 0.0138275 | 0.0086907 | 0.0115662 | 0.0113832 | 0.1324579 | -0.0688008 |
| 0.0227290 | | | | | | |
| -0.0021781 | 0.0017913 | 0.0026523 | -0.0008571 | -0.0008811 | -0.0232865 | 0.0497329 |
| -0.0134317 | -0.0147429 | -0.0070044 | -0.0160679 | -0.0154945 | -0.3720814 | 0.1785091 |
| -0.0254030 | | | | | | |
| 0.0027644 | 0.0029944 | -0.0191592 | 0.0052209 | -0.0031843 | 0.0087496 | -0.0160094 |
| 8595.0266368 | 8595.0232576 | 8595.0271488 | 8595.0284800 | 8595.0212096 | 12.1504907 | -5.3945298 |
| 0.0099833 | | | | | | |
| 0.0010504 | 0.0026084 | -0.0176963 | 0.0046779 | -0.0031546 | 0.0134848 | -0.0173206 |
| 8595.0244864 | 8595.0395392 | 8595.0236672 | 8595.0300160 | 8595.0236672 | 11.9848316 | -5.3257802 |
| 0.0198030 | | | | | | |
| 0.0009239 | 0.0051231 | -0.0222525 | 0.0041076 | -0.0033694 | 0.0083480 | -0.0095820 |
| 8595.0259200 | 8595.0212096 | 8595.0359552 | 8595.0298112 | 8595.0208000 | 12.3121527 | -5.4645017 |
| 0.0186788 | | | | | | |
| 0.0026900 | 0.0037085 | -0.0238988 | 0.0048731 | -0.0031925 | 0.0112235 | -0.0186456 |
| 8595.0272512 | 8595.0275584 | 8595.0298112 | 8595.0350336 | 8595.0277632 | 12.3146790 | -5.4705673 |
| 0.0114125 | | | | | | |
| 0.0021173 | 0.0036116 | -0.0180826 | 0.0041522 | -0.0024849 | 0.0110406 | -0.0180721 |
| 8595.0217216 | 8595.0228480 | 8595.0225408 | 8595.0295040 | 8595.0367744 | 12.0490471 | -5.3475179 |
| 0.0074771 | | | | | | |
| 0.1577295 | 0.1627191 | -0.4265506 | 0.1845614 | -0.1213549 | 0.1324644 | -0.3720779 |
| 6.9346200 | 6.7689618 | 7.0962860 | 7.0988091 | 6.8331766 | 546.6747712 | -242.3516608 |
| 0.4852358 | | | | | | |
| -0.0701118 | -0.0711230 | 0.2118834 | -0.0879201 | 0.0553729 | -0.0688037 | 0.1785075 |
| -3.0849723 | -3.0162230 | -3.1549460 | -3.1610102 | -3.0379605 | -242.3516512 | 107.4785264 |
| -0.1922351 | | | | | | |
| -0.0009788 | -0.0017913 | -0.0225927 | -0.0117169 | 0.0083519 | 0.0227290 | -0.0254030 |
| 0.0054998 | 0.0153195 | 0.0141953 | 0.0069290 | 0.0029936 | 0.4852477 | -0.1922404 |
| 0.7042432 | | | | | | |

We note, that owing to the closure in the data, there will be a spurious negative correlation of approximately .07. This was not corrected for in Potter's material.

DISCRIMINATOR COEFFICIENTS

| | | | | |
|---|---|---|---|---|
| 0.098642 | 0.010696 | 0.155543 | -0.005372 | -0.036095 |
| -0.064107 | -0.022584 | 0.196136 | 0.248257 | 0.161565 |
| 0.246297 | 0.168085 | 11.181265 | -4.957331 | -0.033441 |

| D | D SQUARE |
|---|---|
| 2.4027799 | 5.7733515 |

SIGNIFICANCE FOR D SQUARE AND T SQUARE

T SQUARE = 118.0720752

F = 6.4939641

DF1 = 15 DF2 = 66

The significance test indicates a high level of significance and there is not much doubt, that Potter's data indicate that on the basis of the chemical variables involved, it should be possible to discriminate between shales from the two environments.

In personal discussions, Potter has proclaimed his preoccupation with redundant variables, and at the metting in December we had several discussions on the topic. Clearly, closely correlated variables are going to add very little to the value of the Hotelling $T^2$. In fact, a simple practical test will show, that you can increase the number of variables in a generalized distance study, virtually without affecting the value of $T^2$, providing the new variables are highly correlated. My advice to Potter was, in attempting to eliminate redundancy in his data, was to remove highly correlated variables. In other words, these variables, add little or no information.

In this example, I did not calculate the homogeneity of covariance matrices statistics. We can, however, look at this in another manner and I decided to do so for this particular case. It is possible, when the covariance matrices are not unequal, to produce a rough generalized distance by means of a rough multivariate approximation to the so-called Welch procedure for unequal variances, well known from univariate statistics. This is simply done by producing a covariance matrix, which is the half of the sum of the component sample covariance matrices. This yields a $D^2$ of 5.86, which differs only slightly from the value obtained under the hypothesis of equal covariance matrices. The results are different enough to make me suspect some sort of structural difference in the covariance matrices, but not different enough to jeopardize the conclusions as regards differences in the two environments.

## CHAPTER VII Generalized Statistical Distance

Several of the points pertinent to this chapter have already been raised in Chapter VI and shall only be touched upon again in the interest of continuity of presentation. For the purposes of this chapter I am going to present the material in a way slightly different from the foregoing chapters, as I wish to bring out certain relationships between multivariate tests that otherwise might not seem apparent. The need for this may perhaps be brought out by mentioning that the method of canonical variates and Model 2 of multiple linear discriminant function analysis actually overlap almost entirely. Multivariate statistical analysis is not made up of a number of watertight compartments but rather of a more or less unified body of knowledge with "output points" located conveniently for extracting various features of importance for a particular set of problem conditions. I also wish to bring in the basic theorem of multivariate statistics, that concerning the Wishart distribution. I believe, that if the computer user realizes these interconnexions, pitfalls and unnecessary repetitions will be avoided in the development of programs concerned with data based on many variables.

### Multivariate tests for two populations

We shall consider to multivariate populations in the variables:

$$x = \begin{pmatrix} x_1 \\ \vdots \\ x_p \end{pmatrix}$$

with, as is our wont, the mean vectors indicated as:

$\mu_1$

and $\mu_2$. The difference mean vector will as usual be denoted by $\delta$. For the present purposes, we shall regard the covariance matrices as equal, thus,

$$\Sigma_1 = \Sigma_2 = \Sigma.$$

Each of the populations is normally distributed with respective frequency function:

$$f_i = \text{const.} e^{-\frac{1}{2}(x - \mu_i)'\Sigma^{-1}(x-\mu_i)}$$

$i=1,2$; the constant term is

$$\text{const.} = (|\Sigma|^{\frac{1}{2}}(2\pi)^{p/2})^{-1}.$$

The Wilk's $\Lambda$-criterion

Firstly, we shall consider the Wilk's lambda, already referred to in the foregoing, but as yet unexplained in any detail.

Consider now that we have p-variate samples, based on $n_1$, respectively $n_2$, degrees of freedom, drawn from the two populations described in the above section. The observations are independent. Our sample mean vectors are $\bar{x}_1$ and $\bar{x}_2$, the mean vector of differences will be written $\bar{d}$ and the sample covariance matrix may be written as:

$$z^{-1}S = z^{-1}(S_1 + S_2),$$

where $z = n_1 + n_2 - 2$. Vector $\bar{d}$ has expectation equal to $\delta$ and covariance matrix

$$(n_1^{-1} + n_2^{-1})\Sigma = c^{-1}\Sigma.$$

Since $\bar{d}$, in accord with our assumptions concerning the two populations, is multivariate-normally distributed, its frequency function will be equal to:

$$\text{const.} e^{-\frac{1}{2}c(\bar{d} - \delta)'\Sigma^{-1}(\bar{d} - \delta)}$$

Since $S = S_1 + S_2$ follow a Wishart distribution with z degrees of freedom, it may be shown, as was originally done by Sam Wilks, that the variable:

$$\Lambda = \frac{|S|}{|S + \bar{d}c\bar{d}'|} = \frac{|S|}{|S + Q|} = \frac{|S|}{|T|}, \qquad \text{(VII:1)}$$

which is a determinantal ratio of a kind often occurring in multivariate statistics and of which we already have had occasion to observe. In the above, B is the #between# matrix and T is the #total# matrix, reckoned from the grand vector of means.
This determinantal equation, (1) follows the non-central β-distribution.

Under the null hypothesis:

$$H_o: \quad \delta = 0,$$

the parameter for non-centrality reduces to zero and the distribution of $\Lambda$ becomes the common type of β-distribution.

We have thus here an exact test for the hypothesis that the populations have a homogeneous set of mean vectors. P.C. Tang has worked out a set of tables which permit one to work out the power of the test.

The Wilks lambda criterion was produced in 1932 as a generalization of the one-way model of the analysis of variance. Inasmuch as one may transform from the beta distribution to the distribution of the variance ratio,

it can be shown, that (1) may be tested for significance by means of the expression (in the null case):

$$\frac{n_1 + n_2 - p - 1.}{p} \cdot \frac{1 - \Lambda}{\Lambda} \qquad \text{(VII:2)}$$

This is distributed as F, with p and $n_1+n_2 - p - 1$ degrees of freedom. Inasmuch as this is a generalization of ANOVA, it is normally used with more than two groups. I have discussed the criterion in the present connexion in order to bring out the fact, that in the case of two populations, it reduces to the same thing as the Hotelling $T^2$. This was not realized in the book by Miller and Kahn (1962).

## Hotellings $T^2$

Further to what I have just said, the two population version of the lambda criterion is also equivalent to the $D^2$ attributed to P. C. Mahalanobis. Using the same terminology as earlier on in this chapter, Hotelling defined his generalization of the <u>student</u> t-test is:

$$T^2 = cz\bar{d}'S^{-1}\bar{d} \ , \qquad \text{(VII:3)}$$

which, in the null case, has the distribution:

$$\text{const.}(T^2)^{(p/2-1)}(1 + z^{-1}T^2)^{-(n_1+n_2-1)/2}$$

It may be shown (by a transformation involving the term-

$$T^2/z = pF/(n_1+n_2-p-1))$$

that $T^2$ may be related to the variance ratio. It is also a matter of straightforward algebra to show, that Wilks' lambda and the $T^2$ are connected by means of the expression:

$$T^2/z = (1 - \Lambda)/\Lambda.$$

## The generalized distance of Mahalanobis

The well known generalized distance was defined as, in terms of population quantities:

$$\Delta^2 = \delta'\Sigma^{-1}\delta;$$

geometrically, this is the square of the length of a vector in a p-dimensional space (defined be a set of oblique axes). $D^2$ is easily transformable to $T^2$ by a simple multiplication by $c^{-1}$.

One point perhaps worth noting at this juncture is that $D^2$ is, moreover, not independent of the respective sample sizes, which may be easily shown by a simple experiment. This lack of invariance with respect to the sample size is not a desirable feature. The same remark applies to $T^2$.

---

Examples VII.
Reference is made to the examples given in the previous chapter, as both include $D^2$ and $T^2$ computations.

---

General Remarks on the techniques reviewed in Chapter VII

It should perhaps be mentioned, that the generalized distance is not the "actual" distance between the mean vector tips in the variable space. If all our component variables are uncorrelated, it is a straightforward measure of distance between the points, but where the variables are correlated, this is clearly no longer applicable.

It is usual, in a study based on generalized distances, to produce a model, devised by joining, topologically, the points represented by the samples by the lengths of the generalized distances. Such a model is, in appearance, not unlike a crystal structure model.

There are numerous examples of the application of the $D^2$ to geologic topics. Papers by Shaw, Kudo, Griffiths, and many others have appeared treating problems of geochemistry, mineralogy, sedimentology and many taxonomic questions in paleontology.

Reyment (1962) has dealt with the problem of computing a generalized distance, related to a statistical test (a $T^2$ translatable to a variance ratio) in the case when the covariance matrices are demonstratably unequal. This procedure treats both the case of equal sample size and the case of unequal sample size by a generalization of a procedure devised by Scheffe for a univariate problem appearing in the analysis of variance.

Burnaby (1966) has been concerned with a particular problem arising in the study of organisms in which growth is continuous. His problem concerns estimating a growth-invariant discriminant function and a growth invariant $D^2$ and relies heavily on an a set of assumptions concerning vectors of direction numbers, known a priori, the effects of which are successively removed from the covariance matrix. The weakness of the procedure lies in finding the vector of direction numbers, as there is no satisfactory way of obtaining an objective estimate of these.

Both of the above-discussed procedures have been programmed by the writer and are available in the StateGeological Survey library of programs.

The problem of growth invariance has recently also been taken up by the German mathematician Udo Rempe along the same lines as suggested by me recently in a study of variation in forams.

In the special case where the variables in a generalized distance study are uncorrelated, with the same unit variance;

$$S = I,$$

the covariance matrix is the unit matrix; $D^2$ reduces to

$$D^2 = \bar{d}'\bar{d},$$

and, in particular, if we consider two variables ($p = 2$), and we write $d_i$ for the difference between two scalar mean values, $d_i = u_1^{(i)} - u_2^{(i)}$, $D^2$ becomes (remembering that $p=2$),

$$D^2 = d_1^2 + d_2^2 = d_1^2(1 + f^2),$$

where $f = |d_2/d_1|$. Why did I bring f into the picture? The reason for expressing the function in terms of f is to point out that the situation becomes more complex with the addition of variables and it is no longer readily possible to be sure whether the "distance" D increases with the inclusion of further variables. This point related back to a remark I made in the Chapter on Discriminant Functions. Let us now see what happens if we consider the element of correlation between our two random variables, $x^{(1)}$ and $x^{(2)}$. If these variables are correlated to the extent $\psi$, the above expression for $D^2$ (for positive differences in the mean vectors) becomes, as has been shown by Cochran:

$$D^2 = d_1^2 + \frac{d_1^2(f - \psi)}{1 - \psi^2}.$$

This is greater than the expression given immediately above for no correlation if

$$(f - \psi)^2 > f^2(1 - \psi^2).$$

Thus, as a corollary, we may remark, that negative correlation will always be helpful in increasing distance, but positive correlation will have a non-advantageous effect unless,

$$\psi > 2f/(1 + f^2).$$

This, I feel, is a rather important point to the computer worker inasmuch as it could provide a source of in the development of a computer program, particularly in the case of highly correlated variables, such as occur in biology, and he may wish to bolster against deleterious effects of positive correlation by some sort of branching procedure.

The case of highly multivariate small samples

This represents a class of problems of much potential importance in geology. The problem arises in a situation where the worker has measured a large number of variables but the sample size is small, smaller than the number of variables.

It is oftennot always realized, particularly by users of icompletely buffered computer programs, that many multivariate statistics are indeterminate for $p > N - 1$ ($p$= numberof variables; $N$ = number of observations vectors), and it is not uncommon to observe, that correlation matrices are computed, based on a number of variables exceeding the number of observation vectors.

This pitfall also occurs in $D^2$-computations, discriminant functions, etc. and is one that must be kept vividly in mind by the tyro.

What can be done with this type data ? Dempster has taken up the question for the problem of computing a significance test for the separation of two highly multivariate small samples, producing thereby a sort of $T^2$-type statistic, although not an exact analog of it.

As far as I am aware, this kind of approach has not been developed any further by Dempster nor by any other workers, owing probably to the difficulties connected with the distribution theory.

As a basic model we may regard an experiment (in general terms), that is carried out on, say, 12 subjects, involving 60 chemical analyses. The same set of tests are performed on a second small sample.

In general a type of individual or object is contemplated on each example of which a large number k of different characteristics may be measured, and it is supposed that 2 groups of such individuals can be distinguished by means apart from the k measured items. Suppose a small sample to be available from each group so that the basic data consists of k items measured on each member of 2 samples. Now many questions may be asked about the kind and degree of the relationship between the k measured variables on one hand and on the other hand the two-valued variable which assigns each individual to his proper group. Our concern is with a situation in which the samples do not show a clear and meaningful relation between the group of an individual and his measurements on a single item or small set of items but where it seems reasonable that all k characteristics might be used to define a relationship of sufficient strength to produce statistical significance on a test with fairly small samples. This may be described in statistical terms as a multivariate 2-sample problem where the air is a significance test to distinguish between the populations sampled, and it should not be confused with the apparently more difficult problem of patterning where a single sample is presented as an unknown mixture of 2 groups and the objective is to find a grouping which predominates in some sense. (This may be solved sometimes using PCA).

The test is based on similar theory and directed at the same kind of difference as the usual 2-sample t-test or its classical multivariate generalization using essentially Hotelling's $T^2$. The 2 groups are populations whose means are 2 points in k-space and

whose scatter about these means is largely described by the within-population variances and covariances of the k measures. The question is whether the population means are sufficiently separated relative to the scatter about the means to be shown significant from samples of sizes $n_1$ and $n_2$. The theory assumes homogeneity of variances and covariances within populations and also multivariate normal distributions, but in practice the method may be expected to share certain robustness qualities with analysis of variance techniques. The present method is a substitute for $T^2$ made necessary because $T^2$ is undefined for $k > n_1 + n_2 - 2$ and in any case requires inversion of a matrix of order k which is impractical for large k. In avoiding these difficulties of $T^2$ we find it necessary to give up the desirable affineness property of $T^2$ whereby the same $T^2$ results from any k linear combinations of the k variables used in place of the k given variables. This necessitates more care in the choice of variables.

The imput for the analysis is taken to be k items, typically shrewdly chosen functions of measured variables, available on each of n individuals. This data may be represented by an n X k matrix $X = (x_{ri})$ where $x_{ri}$ is the value of item i for individual r. The rth row of X may be denoted by the vector $X_r$ and is the set of item values for individual r. An individual will be thought of as a single observation drawn randomly from a multivariate population and we will denote the first- and second-order moments of such a multivariate population by

$$\text{ave}\{X_r\} = M_r \quad \text{and} \quad \text{var}\{X_r\} = L_r$$

where $M_r$ is a 1 X k matrix $(m_{r1}, m_{r2}, \ldots, m_{rk})$ and $L_r$ is a k X k matrix $(l_{rii})$. This notation means

$$\text{ave}\{x_{ri}\} = m_{ri}, \qquad \text{var}\{x_{ri}\} = l_{rii}$$
$$\text{and} \quad \text{cov}\{x_{ri}, x_{ri'}\} = l_{rii'}.$$

We assume the first $n_1$ individuals are a sample from one population and the next $n_2$ are a sample from the second population where $n = n_1 + n_2$, and so we may write

$$\text{ave}\{X_r\} = M' \quad \text{for} \quad 1 \leq r \leq n_1$$

and

$$\text{ave}\{X_r\} = M'' \quad \text{for} \quad n_1 + 1 \leq r \leq n.$$

Also we assume both populations to have the same variances and covariances among items, and we we may write

$$\text{var}\{X_r\} = L \quad \text{for} \quad 1 \leq r \leq n.$$

Here M', M" and L are unknown matrices and our concern is to test the hypothesis that M' - M" is the zero vector.

Now each of the n individuals corresponds to a single degree of freedom (d.f.) and, as often done in univariate work, a coordinate change can be made to yield n new orthogonal single d.f., one corresponding to the overall mean, one to the difference between sample means, and the remainder to within-sample variation. This change of coordinates amount to finding $Y = AX$ where A is an n X n orthogonal matrix and the rows of n X k matrix Y, namely $Y_1, Y_2, \ldots, Y_n$, correspond to the n new orthogonal d. f. The first row of A produces the first new d.f. corresponding to the grand mean and therefore must be

$$\sqrt{n}\left(\frac{1}{n}, \frac{1}{n}, \frac{1}{n}, \cdots \frac{1}{n}\right)$$

and the second row corresponding to differences of sample means must be

$$\left(\frac{1}{n_1}, \frac{1}{n_1}, \cdots [n_1 \text{ terms}] \cdots, -\frac{1}{n_2}, -\frac{1}{n_2}, \cdots [n_2 \text{ terms}] \cdots\right) \Big/ \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}.$$

The remaining rows are arbitrary except that they must satisfy the conditions for orthogonality of A.

The new orthogonal vectors $Y_1, \ldots, Y_n$ have the first and second order moments

$$\text{ave}\{Y_1\} = (n_1 M' + n_2 M'')/\sqrt{n},$$

$$\text{ave}\{Y_2\} = (M' - M'') \Big/ \sqrt{\frac{1}{n_1} + \frac{1}{n_2}},$$

$$\text{ave}\{Y_r\} = 0 \quad \text{for} \quad 3 \leq r \leq n,$$

and

$$\text{var}\{Y_r\} = L \quad \text{for} \quad 1 \leq r \leq n.$$

And from these formulas a method of detecting nonzero M' - M" appears naturally, for, except for the shift in mean due to nonzero M' - M", $Y_2$ has the same mean and variance as each of $Y_3, \ldots, Y_n$ in some average sense. Accordingly a significance test may be based on

$$F = Q_2/[(Q_3 + \cdots + Q_n)/(n-2)]$$

where $Q_i$ is the squared length of $Y_i$, i.e. $Q_i = Y_i Y_i'$. By introducing the notion of length into the definition of the test we also introduce a type on nonuniqueness under linear transformation of the k variables.

For definite distribution theory we assume the $X_i$ to be samples from multivariate normal distributions determined by the means and variances as described. Then $Y_2, \ldots, Y_n$ are also independent and normally distributed with means and variances as described, so that, under the null hypothesis M' = M", $Q_2, \ldots, Q_n$ are independently distributed as a positive quadratic form in normal variables, which distribution depends on all the parameters in L. Fortunately, it is generally a good approximation to use a $\chi^2$-shaped distribution for Q, i.e., write $Q \sim m\chi_r^2$, meaning Q is approximately distributed as m times a $\chi^2$ random variable on r d.f. This results in

$$F \sim F_{r,(n-2)r}$$

where, under the null hypothesis, $F_{r,(n-2)r}$ denotes an F-type random variable on r and (n - 2) r d.f.

In this way the dependency on unknown parameters of the distribution of F is reduced from L to the single parameter r. It is known that $r \leq k$, and r may be thought of as a reduced dimensionality from an ideal dimensionality k which would hold if L were a unit matrix. However, r is unknown and so an exact significance test cannot be based on F. We avoid this difficulty by using an estimate $\hat{r}$ of r and testing F as

though it were $F_{f,(n-2)f}$. This inevitably results in distortion of significance levels but this distortion is slight.

Two methods of estimating r are as follows. Thie first uses only $Q_3, Q_4, \ldots, Q_n$. Supposing these Q-values to be a sample from $m\chi^2_r$, there exist [3] sufficient statistics for m and r, and on of these depends only on r,

$$t = (n-2)\left[\ln\left(\frac{1}{n-2}\sum_3^n Q_i\right)\right] - \sum_3^n \ln Q_i .$$

It can be shown that

$$t \sim \left[\frac{1}{r} + \frac{1+\frac{1}{n-2}}{3r^2}\right]\chi^2_{n-3}$$

is a good approximation even for small r so that a good estimator $\hat{r}_1$ can be defined from

$$t = \left[\frac{1}{\hat{r}_1} + \frac{1+\frac{1}{n-2}}{3\hat{r}_1^2}\right](n-3).$$

A more precise estimator of r can be constructed by making use of the angles among vectors $Y_3, \ldots, Y_n$. If $\theta$ is the angle between 2 such vectors, it can be shown [3] that, analogous to the method above,

$$-\ln\sin^2\theta \sim \left(\frac{1}{r} + \frac{3}{2r^2}\right)\chi^2_1$$

and that the $\binom{n-2}{2}$ angles are approximately pairwise independent. Thus if u is the sum of the natural logs of the squared sines of these angles,

$$u \sim \left(\frac{1}{r} + \frac{3}{2r^2}\right)\chi^2_{\binom{n-2}{2}} .$$

Thus a second estimator $\hat{r}_2$ may be defined from

$$t + u = \left[\frac{1}{\hat{r}_2} + \frac{1+\frac{1}{n-2}}{3\hat{r}_2^2}\right](n-3) + \left[\frac{1}{\hat{r}_2} + \frac{3}{2\hat{r}_2^2}\right]\binom{n-2}{2}.$$

There are certain disadvantageous properties associated with the test. The first concerns the concept of length and angle in k-dimensional space, the so-called nonaffineness property. This means, that if we use k linear combinations of k variables in place of the given k variables themselves, we will obtain different

lengths and angles and consequently different significance levels. The test is therefore nonunique. A result of this is, and I think it is an undesirable one, that a priori knowledge is required on the part of the person making use of the method plus a not insignificant portion of subjectivity.

A second property of the test which may seem a practical disadvantage is that the last (n - 2) rows of matrix A were partly arbitrary. It can be easily shown that this arbitrariness does not affect at all the value of F, but only $\hat{r}_1$ and $\hat{r}_2$ which are of secondary importance.

Computer application of the method

In the following, we shall review the most elegant way of presenting the computations from the point of view of a FORTRAN program. The numerical illustration is one concerning 62 biochemical analysis made on 12 subject and is only included in order to clarify the steps.

The first step is to read in the 12 X 62 matrix X and reduce immediately to W = XX' on which further computations are performed. The last 10 rows of A are determined using the method of random choice just described. The entries of $C_3$, ... , $C_{12}$ are 120 random normal deviates produced internally by the machine using a subroutine. From A and W the $Q_i$ are computed for $2 \leq i \leq 12$ from the formula

$$Q_i = A_i W A_i'$$

and from the $Q_i$ previously given formulas were used to find F and $\hat{r}_1$. In order to find $\hat{r}_2$, it is necessary to compute quantities $Q_{ij}$ for $3 \leq i \leq j \leq 12$ from the formula

$$Q_{ij} = A_i W A_j'$$

and from these the squared sine of the angle between $Y_i$ and $Y_j$ is given by

$$\sin^2 \theta = 1 - \frac{Q_{ij}^2}{Q_i Q_j}.$$

From these squared sines u and $\hat{r}_2$ are found directly. The random choice of A is made a total of 5 times to check empirically the variations in $\hat{r}_1$ and $\hat{r}_2$ which might be expected.

An alternative method of attaching a significance level to F is the randomization test similar to that proposed by Pitman for the univariate 2-sample t-test. Suppose F were computed for each of $\binom{n}{n_1}$ ways of dividing the n individuals into samples of size $n_1$ and $n_2$. Suppose these F values were ranked and the F corresponding to the true division into samples had rank S. Then, under the null hypothesis that the 2 populations have identical distributions, S would be equiprobably distributed over the integers 1 to $\binom{n}{n_1}$. Thus the formula gives a means of attaching a significance level to F. It turns out to be equivalent to rank the lengths of the vectors joining sample means, and this simplifies calculations.

$$\text{Prob}\,(S \leq s) = s \Big/ \binom{n}{n_1}$$

# CHAPTER VIII CANONICAL CORRELATION

Introductory comments.- The multivariate statistical procedure here to be discussed is one of potential importance in some geological applications, particularly in paleoecologic (and ecologic) studies. We shall consider two sets of variates with a joint distribution; it is desired to analyze the correlations of one set with the other. In connexion with the methodology of this analysis, we find a new coordinate system in the space of each set of variates, this being performed in such a way that the new coordinates display, without ambiguity, the system of correlation. In other words, we find the linear combinations of variables in each set that have maximum correlation and these linear combinations are the first coordinates in the new systems. Thereafter, a second linear combination is sought in each set, such that the correlations between these is the maximum of correlations between such linear combinations as are uncorrelated with the first linear combinations. The procedure is continued until the two new coordinate systems are completely specified. The theory of the method of canonical correlations was worked out by Hotelling in 1936 and it will already be clear to you, that much of the basic linear algebra is related to that occurring in the Hotelling procedure termed canonical variates.

It is also apparent that, in a way, canonical correlation represents a generalization of the regression model.

Consider the random vector X of p components, which has the covariance matrix $\Sigma$, assumed positive definite. Inasmuch as our particular problem is only concerned with variances and covariances, it will be assumed that $E(X) = 0$.

Consider now the random vector X partitioned into two subvectors, $X^{(1)}$ and $X^{(2)}$, of $p_1$ and $p_2$ components respectively. For the following presentation it will be assumed that

$p_1 \leq p_2$. The covariance matrix is partitioned into $p_1$ and $p_2$ rows and columns:

$$\Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{21} \\ \Sigma_{12} & \Sigma_{22} \end{pmatrix}.$$

We may write an arbitrary linear combination of the components of $X^{(1)}$ as

$$U = \alpha' X^{(1)},$$

and an arbitrary linear function of $X^{(2)}$ as

$$V = \gamma' X^{(2)}.$$

It is required to find the linear functions that have maximum correlation. The correlation of a multiple of U and a multiple of V is the same as the correlation of U and V, therefore, it is possible to make an arbitrary normalization of $\alpha$ and $\gamma$. It is thus required, that U and V have unit variance, which determines the nature of our two vectors.

$$1 = EU^2 = \alpha' \Sigma_{11} \alpha,$$

and,

$$1 = EV^2 = \gamma' \Sigma_{22} \gamma.$$

It is readily appreciated that the correlation between U and V is:

$$EUV = \alpha' \Sigma_{12} \gamma. \qquad \text{(VIII:1)}$$

The algebraic problem is then to find the two vectors so as to maximize (1), subject to the two foregoing equations.

The correlation, $\lambda$, between U and V is given by (VIII:1) when satisfy

$$\begin{pmatrix} -\lambda\Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & -\lambda\Sigma_{22} \end{pmatrix} \begin{pmatrix} \alpha \\ \gamma \end{pmatrix} = 0 \quad . \qquad \text{(VIII:2)}$$

We can take $\lambda = \lambda_1$ in finding the maximum correlation. A solution of (VIII:2) for $\lambda = \lambda_1$ may be written

$$\alpha^{(1)}, \gamma^{(1)}$$

Let

$$U_1 = \alpha^{(1)'} X^{(1)}$$

and

$$V_1 = \gamma^{(1)'} X^{(2)}$$

Then $U_1$ and $V_1$ are normalized linear combinations of $X^{(1)}$ and $X^{(2)}$ respectively, with maximum correlation.

The second linear combination is then found, and so on.

Consequently, at the rth step, the following linear combinations have been obtained:

$$U_1 = \alpha^{(1)'} X^{(1)}, V_1 = \gamma^{(1)'} X^{(2)}, \ldots, U_r = \alpha^{(r)'} X^{(1)},$$

$V_r = V^{(r)'} X^{(2)}$, with corresponding correlations given by:

$$\begin{vmatrix} -\lambda\Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & -\Sigma_{22} \end{vmatrix} = 0 \quad . \qquad \text{(VIII:3)}$$

Remembering our condition of $P_1 \leq P_2$, there will be $P_1$ canonical correlations. The components of

$$U = \begin{vmatrix} U_1 \\ \cdot \\ \cdot \\ \cdot \\ U_{P_1} \end{vmatrix} .$$

are one set of canonical variables and the components of

$$V = \begin{vmatrix} V_{P_i} + 1 \\ \vdots \\ V_{P_2} \end{vmatrix}$$

are the other set.

Another approach to canonical variables is used if the two sets of variables are not random. This will correspond to the generalized regression interpretation I mentioned in the introduction.

In sample terms this time, we could consider the matrix of regression coefficients

$$B = S_{12} S_{22}^{-1}, \text{ from } S$$

We are determinantal equation

$$\left| BS_{22} B' - \gamma \psi \right| = 0 \qquad \text{(VIII:4)}$$

where $\psi$ is a matrix

$$\psi = S_{11} - S_{12} S_{22}^{-1} S_{21}$$

Estimation Procedure

If $x_1, \ldots x_n$ are N obsevational vectors from the multivariate Normal $N(\mu, \Sigma)$. Each observational vector is partitioned into $P_1$ and $P_2$ components, as observed in the foregoing section:

$$x_\alpha = \begin{pmatrix} x_\alpha^{(1)} \\ x_\alpha^{(2)} \end{pmatrix} .$$

It is possible to proceed via either the sample covariance matrix, or its correlation matrix.

Using still the covariance matrix, the maximum likelihood estimates of the canonical correlations, $\Lambda$, are the roots of

$$\begin{vmatrix} -\lambda S_{11} & S_{12} \\ S_{21} & -\lambda S_{22} \end{vmatrix} = 0 \qquad \text{(VIII:5)}$$

The maximum likelihood estimates of the coefficients of the jth canonical components satisfy

$$\begin{pmatrix} -\lambda_i S_{11} & S_{12} \\ S_{21} & -\lambda_i S_{22} \end{pmatrix} \begin{pmatrix} a^{(j)} \\ g^{(j)} \end{pmatrix} = 0 , \qquad \text{(VIII:6)}$$

$$a^{(j)\prime} S_{11}\, a^{(j)} = 1 \qquad \text{(VIII:7)}$$

$$g^{(j)\prime} S_{22}\, g^{(j)} = 1 \qquad \text{(VIII:8)}$$

also,

$$a^{(j)\prime} S_{11}\, a^{(j)} = 1,$$

and

$$g^{(j)\prime} S_{22}\, g^{(j)} = 1.$$

The sample canonical variates may be defined as:

$$a^{(j)\prime} x_{\alpha}^{(1)} \text{ and } c^{(j)\prime} x_{\alpha}^{(2)} .$$

It may be useful to consider a geometrical interpretation of canonical correlations. The rows of the (P x N) data matrix, $X = (x_1 , \ldots , x_n)$ may be regarded as p vectors in N-dimensional space. If we consider the deviations from the mean vector $(n_1 - \bar{x}, x_2 - \bar{x}, \ldots , x_n - \bar{x})$, these are the p vectors projected on the (N - 1)-dimensional subspace orthogonal to the equiangular line. For convenience, these may be written as $n_1^* , \ldots , x_p^*$ . Any vector $U^*$ with the components $a'(x_1^{(1)} - \bar{x}^{(1)}, \ldots x_N^{(1)} - \bar{x}^{(1)}) = a_1 x_1^* + \ldots a_{p_1} x_{p_1}^*$ is located in the $p_1$- space spanned by $x_1^* , \ldots , x_{p_1}^*$ and a vector $V^*$ with components

$$g\,(x_1^{(2)} - \bar{x}^{(2)} \quad \ldots , x_N^{(2)} - \bar{x}^{(2)}) = g_1 x_{p_1+1}^* + \ldots g_{p_2} x_p^*$$

is in the $p_2$ - space spanned by $x^*_{p_1+1}, \ldots, x^*_P$.

The cosine of the angle between these two vectors is the correlation between $u_d = a' x_\alpha^{(1)}$ and $V_\alpha = g' x_\alpha^{(2)}$ $(\alpha = 1, \ldots, N)$. Thus, finding a and g to maximize this correlation is equivalent to finding the vectors in the $P_1$ - space and the $P_2$ - space such that the angle between them is least -- i.e., the cosine is maximum. The second canonical variates correspond to vectors orthogonal to the first canonical variates and with the angle minimized.

## General Observations

If we compare the technique here described with multiple regression, we see that it is possible to regard the latter as a special case of canonical correlation when $P_1 = 1$, and $P_2 > 1$. In canonical correlation, both multiple criteria and multiple predictors occur.

In my opinion, the method has its main geological applicability in ecologic studies. In connection herewith, it is a useful way of analyzing the interrelationship between a set of variables of, say, electrochemical type, and a set relating to organic elements.

## Computer Program Steps

The program available at K.U. uses the matrix of correlations between variables.

$$R = \begin{bmatrix} R_{11} & \vdots & R_{12} \\ \cdots & \vdots & \cdots \\ R_{21} & \vdots & R_{22} \end{bmatrix}$$

$R_{11}$: consists of the correlations between the variables of set I.

$R_{22}$: consists of the correlations between the variables of set II.

$R_{12}$, $R_{21}$: the correlations between the variables of set I and those of set II.

The determinantal equation is then

$$\left| R_{22}^{-1} R_{21} R_{11}^{-1} R_{12} - \lambda I \right| = 0 \qquad \text{(VIII:9)}$$

If the vectors of the other set are required first for some reason, the matrices in (VIII:9) are changed accordingly. If a vector b has been obtained for one set, the vector a corresponding to the second set is given by:

$$a_i = (R_{11}^{-1} R_{12} b_i) / \sqrt{\lambda_i} \qquad \text{(VIII:10)}$$

The $\lambda_i$ are the squares of the canonical correlations; thus $R_{CAN(i)} = \sqrt{\lambda_i}$

The method of canonical correlations has been generalized to the situation of more than two sets of variables. I have not myself attempted to apply this to any of my research and am therefore unable to express an opinion as to the overall usefulness of the generalized version in geology. Canonical correlations have not been given much attention at all by biologists and geologists.

The program referred to above also includes a test of significance for eigenvalues, developed by M.S. Bartlett. The test criterion, $\psi$ , is:

$$\psi = \prod_{i=1}^{P_2} (1 - \lambda_i), \quad (P_2 \leqslant P_1) \qquad \text{(VIII:11)}$$

$\psi$ has an approximate $\chi^2$ - distribution and provides a large-sample test for the null hypothesis that the set of $P_1$ variates is not significantly connected with the set of $P_2$ variates.

$$\chi^2 \approx - [N - 0.5 (P_1 + P_2 + 1)] \ln \psi \qquad \text{(VIII:12)}$$

with $P_1 P_2$ degrees of freedom.

If the null hypothesis is rejected, the significance of the remaining $P_2 - 1$ generalized eigenvalues may be tested:

$$\psi' = \prod_{i=2}^{P_2} (1 - \lambda_i),$$

and,

$$\chi^2 \approx - [N - 0.5 (p + q + 1)] \ln \psi' ,$$

with $(P_1 - 1)(P_2 - 1)$ degrees of freedom, and so on.

A possible pitfall in the present connotation is the selection of a suitable canonical correlation model. The majority of the studies in the literature that have come to my attention appear to prefer to rely on the information obtainable from the first eigenvalues alone. Experience demonstrates, however, that others of the eigenvalues may be more important in disclosing meaningful relationships between sets.

If it is decided to experiment with the FORTRAN programs in Cooley's and Lohnes' book (1962), it should be pointed out that the two methods given agree poorly. First, the iterative method for extracting eigenvalues and eigenvectors is inaccurate, on the IBM 7040 at K.U. at least, secondly, running the same data in both gives different results. Furthermore, the second program, using the Jacobi method of root extraction, is incomplete, and does not agree with the flow diagram.

Multivariate Regression

Consider variables x, linearly related to a set of z's, regarded as fixed, by

$$x = \beta + E \ , \qquad \text{(VIII:13)}$$

where $\beta$ is a p x q matrix of coefficients and E is a prior matrix of errors.

If the subvectors $E_1 , E_2 , \ldots , E_p$ were independent there would then be a set of p independent regressions, one for each vector x. However, generally independence will not be assumed.

As a concrete presentation of the ideas behind this, we could consider four variables; $x_1 , x_2 , x_3 , x_4$. One hypothesis that may be of inherest could be that the data are homogeneous with respect to, say, the means of series. An appropriate ratio of determinants test is available for this.

If there is a high degree of correlations between variables, we may ask whether differences between means of series are due to say only $x_1$ and $x_2$, and $x_3$ and $x_4$ only contribute by virtue of correlation with $x_1$ and $x_2$. To answer this, we determine the

regressions of $x_3$ and $x_4$ and $x_1$ and $x_2$, extract them from the total variation and test the residual matrices.

Thus, $x_3$ and $x_4$ act as a matrix of dependent variables (the 'x's of [1]), and $x_1$ and $x_2$ as a matrix of independent variables (the z's of [1]).

EXAMPLE VIII:1

Example employing the procedure of canonical correlation in accord with the program supplied in the set of worksheets.

The data derive from five ecological variables measured on pore water from the interstitial sedimentary environment. These variables are: pH (= $x_1$), Eh (= $x_2$), free, dissolved oxygen (= $x_3$), content of carbonates from all sources (= $x_4$) and content of chemically oxidizable matter (= $x_5$).

The various steps in the calculations and intermediate results are given below.

CANONICAL CORRELATION FOR SEDIMENTARY DATA

NO. VARIABLES ON LEFT = M1 = 3

NO. VARIABLES ON RIGHT = M2 = 2

NO. OBSERVATIONS = N = 37

MEANS FOR ALL VARIABLES

| PH | EH | Dissolved oxygen | Carbonates | Oxidizable organic matter |
|---|---|---|---|---|
| 7.5206 | -146.8378 | 39.4595 | 1.7361 | 2.2189 |

STANDARD DEVIATIONS FOR ALL VARIABLES

| | | | | |
|---|---|---|---|---|
| 0.1280 | 90.2095 | 17.5293 | 1.6121 | 1.2831 |

R11 MATRIX OF CORRELATIONS

| | | |
|---|---|---|
| 1.000000 | 0.345783 | 0.038894 |
| 0.345783 | 1.000000 | -0.216512 |
| 0.038894 | -0.216512 | 1.000000 |

R22 MATRIX OF CORRELATIONS

| | |
|---|---|
| 1.000000 | -0.214395 |
| -0.214395 | 1.000000 |

R12 MATRIX OF CORRELATIONS

| | |
|---|---|
| 0.136944 | 0.037357 |
| 0.059086 | 0.049451 |
| -0.020746 | -0.160082 |

INVERSE OF MATRIX R11

| | | |
|---|---|---|
| 1.1535942 | -0.4287048 | -0.1376878 |
| -0.4287048 | 1.2085005 | 0.2783289 |
| -0.1376878 | 0.2783289 | 1.0656167 |

DETERMINANT = 0.8262200 OF MATRIX R11

FIRST SQUARED CANONICAL CORRELATION = 0.0423447

FIRST CANONICAL CORRELATION = 0.20578

RIGHT-HAND WEIGHTS
-0.627001 -0.779019

LEFT-HAND WEIGHTS
-0.643978 -0.020049 0.764781

SECOND SQUARED CANONICAL CORRELATION = 0.0111160

SECOND CANONICAL CORRELATION = 0.10543

RIGHT-HAND WEIGHTS
0.813996 -0.580870

LEFT-HAND WEIGHTS
0.751721 0.054136 0.657255

## TESTS OF SIGNIFICANCE

LAMBDA ONE = 0.9470100

CHI SQUARE = 1.8511514

DF = 6.

LAMBDA TWO = 0.9888840

CHI SQUARE = 0.3800613

DF = 2.

Neither of the canonical correlations are significant and it may therefore be concluded, that the a priori considered highly plausible model of correlation between a set of electrochemical variables and a set of organic variables, is not a useful and valid one.

# CHAPTER IX

## SOME OBSERVATIONS ON TYPES OF CLASSIFICATORY AND IDENTIFICATORY PROCEDURES

In this chapter it is my intention to review the philosophy of classification, as opposed to quantitative methods of identification, and to give my personal opinions as to what I consider to be a useful approach to the subject for geological problems I have encountered in my own work. I am perfectly aware that the problem met with by others may require a different mode of attack, possibly non-statistical.

Before entering into the main theme of this chapter I wish to examine the meaning pertaining to the word "classification". There is a section of Multivariate Statistical Analysis which bears the title "Classification". This is concerned with what is referred to as (Multiple) Discriminant Function Analysis.

As Sokal has pointed out, this is really not a matter of classification but rather one of identification. At least, this is the interpretation that must be given as soon as this technique is applied to taxonomic situations.

A moments reflection will show why this must be so. We shall consider the statistical discriminantion problem for two populations, compatible with respect to k variables. Call these populations $\Pi_1$ and $\Pi_2$.

We have an observation vector

$$X = (x_1 , \ldots , x_k)$$

It is required to find which of $\Pi_1$ and $\Pi_2$ it belong to. This is thus a question of IDENTIFICATION with either of $\Pi_1$ and $\Pi_2$ and not really one of CLASSIFICATION. Vector X is actually derived from either of $\Pi_1$ or $\Pi_2$. If this is not true, then it is wrong to employ the DISCRIMINANT FUNCTION model.

## The Classification Problem

It is well known that visual methods of traditional stamp, in classification are deeply influenced by subjectivity. It is enough to compare contemporary publications in one's own field of specialization. This is perhaps really strongly brought out in comparing East Block and "Western" works. I call this the first kind of classificatory dilemma. That is, where one already has a form of TAXONOMIC classification in existence, but it suffers from defects of various kinds, introduced in varying degrees by subjectivity. It is desirable to be able to better this, by some means or other, where the element of REPEATABILITY becomes incorporated. Thus, if our model is a reliable one, it should be possible for any other person in the field to be able to take my material, and without a priori knowledge of what I have drawn for classificatory conclusions, end up with the same result.

The second kind of classificatory dilemma I see in particular in non-biologic areas, in which several people have felt the need to have a try at the procedures of NT. From Permo-Carboniferous cyclothems to mineral deposits. Here, the application is concerned with producing a classification, on quantitative grounds, on little or no a priori information. The biological models seem to have been employed, and the resolution needs be only as good as the underlying model.

## Nonstatistical Numerical Taxonomy

Modern numerical taxonomists, such as Rohlf and Sokal, have departed largely from the statistical corpus because of a large number of problems occurring in connection therewith. This is, in my opinion, not always such a good idea, as this approach is adopted, almost always, to get around something unpleasant in the data. Such a thing as the homogeneity in variances and co-variances of two groups (and more than two groups,) is one which may create problems for the numerical taxonomist.

## ADANSONIAN TAXONOMY

As is well know, the concept of numerical taxonomy appears to have been born with Adanson, the French naturalist, who formulated the concept in conjunction with his taxonomic studies on the Recent marine molluscs of Senegal, West Africa.

One of the principles he postulated was the principle of equal weighting of all the characters selected by the zoologist as diagnostic of his material [the question of what is diagnostic is, of course, a moot point.] This principle of equal weighting appears to be the most widely applied one among numerical taxonomists of today. The obvious logic employed in support hereof is, that subjective elements would be introduced in that the quantitative zoologist would be exercising personal opinions and prejudice in the choice-making procedure. However, if one regards, for example, the SIMILARITY COEFFICIENT of numerical taxonomy, in the garb presented by Sokal and Sneath, it soon becomes apparent that this is an area in which personal prejudice is allowed full rein and in actual fact, it turns out that some of these similarity coefficients may only owe their differentness to some form of character weighting or other.

What does the critic of equal character weighting have against it? The most available complaint would seem to be that any definition of what is a diagnositc and useful character (often termed a "unit character") lies in the mind of the person carrying out a particular study and, of necessity, will be subjectively flavored. Different workers will view a certain situation in an unlike manner from others. Hence, the claim of objectivity in conjunction with the principle of equal weighting is one that should invite a certain measure of kindly skepticism.

It is a naturally occurring question, whether it is not possible to produce a character-weighting coefficient that will in some manner compensate for the lack of pertinence in a chosen character.

The generalized statistical distance of Mahalanobis presents a method of character-weighting, whereby the introduction of a new character to a set of characters causes little or no change in the "distance", if this character (or characters) is (are) strongly correlated with characters existing in the set. We may state this in other terms, notably, that if a character conveys no new information for separating between two samples (or populations) the pertinent elements of the inverse covariance matrix of the quadratic form will be very small, to use rather nonprecise language. Like reasoning is applicable to what I shall have to say further on concerning the subject of CANONICAL VARIATES.

The general applicability of the $D^2$ - method is to a degree limited by the difficulty of satisfactorily using it in conjunction with discrete characters, and there are some other problems, such as the a priori establistment of the basic groups.

Suggestions have been made that a possible approach is by means of the information - theoretic quantity of $-\log_e p$. In this presentation, the characters of a taxonomic unit may be regarded as a group of messages which carry information on the taxonomic relationships of this unit. The idea behind this point of view seems well worth looking into.

Another opinion has been put forward by Smirnov with respect to the weighting dilemma. This requires the estimation of prior probabilities for character states, these determining the weights of the states. I only mention this in order to bring out the fact that several points of view in the non-Adansonian sphere are developing. The Smirnov weighting device is, however, far from ideal, for various reasons.

## Application of Canonical Variate Analysis

We shall, from now on, mainly be concerned with continuously varying variables, although, to a certain extent it is possible to analyze discrete data in terms of the statistical procedure to be reviewed. It does not appear advisable,

in my opinion, to attempt a union of discrete and continuous variables in conjunction with the procedure here to be discussed, notably, that of canonical variates.

We shall first briefly review the basic principles of this method. We consider g populations, each of multivariate - normally distributed p- component variables. The theory, which was developed by H. Hotelling in 1936, requires the g covariance matrices of the g K-variate populations to be equal; $\Sigma_1 = .... = \Sigma_g = \Sigma$. If we write $\Omega_w$ for the "within groups" sums of squares and cross products matrices ( $\Omega_i = n_i\Sigma_i$) and $\Omega_B$ for the "between groups" sums of squares and cross products matrix, one wishes to find the generalized eigenvalues

$$| \Omega_B - \lambda \Omega_w | = 0$$

and the associated vectors,

$$( \Omega_B - \lambda \Omega_w )b = 0 .$$

This is nothing more than the generalized determinantal equation problem of linear algebra. The mean vectors of the original variables are transformed into a set of mean vectors in the new space. For k eigenvalues to be defined the condition $g - 1 > k$ is required. When $k > g - 1$, there are $(k - g - 1)$ zero roots $MX\ B = D$ and $g - 1$ other roots.

$$(g \times k)\ (k \times k)\ (g \times k)$$

illustrates the transformation from the original means, M(g x k) to the means in the new space, D (g x k). Statistically this is said to produce a transformation that will emphasize the differences between the means of the estimates of g populations. It is often illuminating to plot the first two transformed means on a bivariate diagram (scatter diagram). This may indicate useful clusterings or groupings in the material, which may be of taxonomic significance. Figure 2 gives a simplified illustration of these for samples from population of European

frogs of the species Rana esculenta and R temporaria. When g = 2, there will only be one now - zero eigenvalue, the corresponding vector of which, $y = b'(n - \bar{x})$ is the discriminant function for 2 populations.

The question of whether the method of canonical variate analysis can be applied when the basic requirement,

$$\Sigma_1 \neq \Sigma_2 \neq \ldots \Sigma_p ,$$

is not fulfilled, may deserve some consideration. Although I am not aware of any detailed analysis of the subject, it is conceivable that this method might be robust toward moderate departures from normality and from equality of variances and covariances. An approximate means of reducing some of the effects of deviations of this kind is by means of a procedure found empirically useful in generalized distance studies, by forming a pooled covariance matrix, $S = 1/2 (S_1 + S_2)$, without regard to the respective degrees of freedom on which $S_1$ and $S_2$ are based. Expressing this for cononical variates in a MANOVA table, we have.

| | | Degrees of Freedom |
|---|---|---|
| "Between" matrix | $Q = T - W$ | $k - 1$ |
| "Within" matrix | $W = (N-K) \sum_{i=1}^{k} S_{wi})/k$ | $N - k$ |
| "Total" matrix | $T = (N - 1) S_T$ | $N - 1$ |

[k = numer of groups; N = total number of observations]

## The Question of Scale Invariance

It is clear that the model of canonical variate analysis, accounted for in the foregoing, has a rather severe limitation imposed upon it by the necessity of having to use like variables; in other words, we are constrained to restrict ourselves to the use of, say, continuous variables, measured on the same scale. It is also not feasible to mix continuous and discontinuous characters, which is a

drawback in almost all numerical taxonomic studies, which usually consider mixed characters. If it is not unrealistic to consider correlations between continuous and discontinuous characters, it would not seem unreasonable to regard a correlation matrix as a legitimate representation of the interrelationships between continuous and most discontinuous characters. Useful as this might appear to be, it finds no application in the model under consideration, inasmuch as it is not designed for extension to the case of g correlation matrices.

A possible, approximate means of approaching a solution to the problem would perhaps be by means of Joereskog's matrix, employed in a version of Factor Analysis, to produce the condition of scale invariance. If R denotes the correlation matrix of a sample, and D is a diagonal matrix,

$$D = \text{diag } R^{-1},$$

the Joereskog matrix is found by the relationship

$$C = D^{1/2} R D^{1/2}.$$

The interesting feature about matrix C is that is is yielded by both the correlation matrix as well as by the covariance matrix corresponding.

We could suggest an approximate type of generalized eigenvalue study by replacing the normal matrices of canonical variates by those based on the Joereskog type of matrix, as shown in the following MANOVA table.

| | | Degrees of Freedom |
|---|---|---|
| "Between" matrix | $Q = T - R$ | $k - 1$ |
| "Within" matrix | $R = (N - k) C_W$ | $N - k$ |
| "Total" matrix | $T = (N - 1) C_T$ | $N - 1$ |

[k = number of groups; N = total number of observations]

The generalized determinantal equation is then

$$|Q - \phi R| = 0$$

and the usual methods of solutions are then applied.

## PRODUCING GROUPINGS FOR CANONICAL ANALYSIS

We are now entering the realm of what Tukey sometimes calls "rough and dirty" statistics. The procedure I am going to discuss is not one which can be given a very convincing statistical raison d'etre but the proof of the pudding is in the eating and practical and experimental experience has provided me with sufficient a posteriori assurance to make me reasonably certain of the viability of the technique.

### The Model

Consider a homogeneously constituted sample of N specimens upon each of which p characters have been measured. Denoting the matrix of sums of squares and cross products of these observation vectors as A, the eigenvalues and eigenvectors hereof will be given by

$$|A - \lambda I| = 0$$

$$(A - \lambda I)\, b = 0$$

respectively. Regarding now the first two eigenvectors of A, $b^{(1)}$ and $b^{(2)}$, substitution of the observation vectors into the vectorial equation will yield a (2 x N) matrix of transformed values, which when plotted as a bivariate scatter diagram, will form a close cluster of points; for multivariate - normally distributed variables, this cluster will approximate to the form of an ellipse.

The possibility of the application of this procedure to "mixed data" has a natural appeal. A sample comprising material of mixed origin will fail to adopt a homogeneous pattern when plotted as a bivariate diagram of transformed observations. In effect, if the transformations have succeeded in magnifying the

degree of unlikeness, a pattern of clusters will be produced instead of the homogeneous concentration of points that result when the observations have been drawn from a homogeneous source. The groupings so obtained may be usefully analyzed further by the method of canonical variates.

The advantage of the clustering technique, using the eigen-method, over plotting the raw data is that the latter is often insufficient to bring out the eventual existence of discrete groups, if the chosen characters, scrutinized in pairs, are not very diagnostic. The plotted transformed values are linear combinations of the input variables and thus, to a considerable extent, take all of these into account (to a degree corresponding to $\lambda_1 + \lambda_2$). It is sometimes possible to produce more distinct groupings by using combinations of eigenvectors other than the first two.