The Use and Interpretation of Principal Component Analysis in Applied Research

Author(s): C. Radhakrishna Rao

Source: *Sankhyā: The Indian Journal of Statistics, Series A (1961-2002)*, Dec., 1964, Vol. 26, No. 4 (Dec., 1964), pp. 329-358

Published by: Indian Statistical Institute

Stable URL: https://www.jstor.org/stable/25049339

THE USE AND INTERPRETATION OF PRINCIPAL COMPONENT ANALYSIS IN APPLIED RESEARCH*

By C. RADHAKRISHNA RAO

Indian Statistical Institute

Visiting at Stanford University

SUMMARY. The paper provides various interpretations of principal components in the analysis of multiple measurements. A number of generalizations of principal components have been made specially in the study of a set of variables in relation to another set known as instrumental variables. The use of generalized principal components in applied research has been indicated. Finally, the difference between factor analysis and principal component analysis has been explained.

1. Introduction

During the last few years, I had the opportunity of surveying some of the multivariate statistical techniques and evaluating their usefulness in practical research work (Rao, 1960, 1961). I undertook to do this mainly because the superiority of the multivariate methods over the simpler univariate analysis has been questioned by some statisticians. It was said that an examination of individual measurements by simpler univariate methods is an essential step which can be supplemented by more elaborate multivariate analysis when this appears feasible and necessary. It was also thought that multivariate methods have only provided new outlets for mathematical theory without materially assisting scientific research.

It is, indeed, necessary to criticise uncritical and inappropriate applications, which are perhaps natural when new techniques are put forward. But a criticism of multivariate methods as such is not justifiable. I have considered a number of practical examples specially in one of the papers (Rao, 1961) to demonstrate the use of multivariate analysis and to emphasize the need for exploring the potentialities of the existing techniques and for further research.

The purpose of the present paper is to examine, in some detail, the role of principal component analysis in applied research. When a large number of measurements are available, it is natural to enquire whether they could be replaced by a fewer number of the measurements or of their functions, *without loss of much information*, for convenience in the analysis and in the interpretation of data. Principal components, which are linear functions of the measurements, are suggested for this purpose. It is, therefore, relevant to examine in what sense principal components provide a reduction of the data without much loss of *information we are seeking from the data.*

The following are some typical statements found in the literature on applications of principal component analysis. 'The transformed variables (the first few principal components) are useful in connection with preliminary investigation of a large number of samples of species from different localities.' 'For an orientatory

* Technical Report No. 9. Prepared Under Contract OE 2-10-065 with U.S. Office of Education.

type of investigation, it would probably suffice with the variable provided by the first principal component (out of four) for this represents 3/4 of the total variation.' 'At first principal components are most likely to be regarded by the nonmathematicians as a highly arbitrary set of manipulations. Such a reaction should be dismissed, as soon as the geometrical meaning is considered : principal component analysis merely leads to new angles of viewing data, analysis best suited to disclose the nature of size and shape variation.' In no case are the statements made substantiated by statistical analysis to show that the information neglected does not lead to misleading conclusions. Often, the principal component analysis is first undertaken without any clear objective and then an attempt is made to interpret the derived results.

The discussion in the present paper has the following aims : (a) To provide various interpretations of principal components. (b) To examine the situations where the principal component analysis can be undertaken with a definite purpose or in an exploratory way in the earlier stages of investigation of a research problem. (c) To indicate the difference between the principal component analysis and the factor analysis. (d) To generalize the principal component analysis in a number of directions useful in applied research.

2. EIGEN VALUES AND VECTORS OF MATRICES

For a theoretical development of the principal component analysis and its interpretation it is necessary to use some results on the canonical reduction of matrices, which are summarized in this section for use in the later sections.

(i) *Eigen values and vectors of a matrix.* Let $\boldsymbol{\Sigma}$ be a non-negative (i.e., positive definite or positive semi-definite) matrix of order $p \times p$. The roots of the determinantal equation

$$|\boldsymbol{\Sigma} - \lambda \boldsymbol{I}| = 0 \qquad \ldots \quad (2.1)$$

are called the eigen values of $\boldsymbol{\Sigma}$. The equation (2.1) has p real non-negative roots, $\lambda_1 \geqslant \lambda_2 \geqslant \ldots \geqslant \lambda_p$. Corresponding to each root λ_i of (2.1), there exists a column vector $\boldsymbol{P}_i$, such that

$$\boldsymbol{\Sigma}\, \boldsymbol{P}_i = \lambda_i \boldsymbol{P}_i, \qquad \ldots \quad (2.2)$$

which is called an eigen vector. Then the following results hold :

(a) $\boldsymbol{P}_1, \ldots, \boldsymbol{P}_p$ can be chosen to be orthonormal whether $\lambda_1, \ldots, \lambda_p$ are distinct or not.

(b)
$$\boldsymbol{\Sigma} = \lambda_1 \boldsymbol{P}_1 \boldsymbol{P}_1' + \ldots + \lambda_p \boldsymbol{P}_p \boldsymbol{P}_p'.$$
$$\boldsymbol{I} = \boldsymbol{P}_1 \boldsymbol{P}_1' + \ldots + \boldsymbol{P}_p \boldsymbol{P}_p' \qquad \ldots \quad (2.3)$$

(c) Let $\boldsymbol{L}_1, \ldots, \boldsymbol{L}_q$ be any set of orthonormal vectors. Then

$$\sum_{i=1}^{q} \boldsymbol{L}_i' \boldsymbol{\Sigma} \boldsymbol{L}_i \leqslant \sum_{1}^{q} \boldsymbol{P}_i' \boldsymbol{\Sigma} \boldsymbol{P}_i = \lambda_1 + \ldots + \lambda_q \qquad \ldots \quad (2.4)$$
$$q = 1, \ldots p,$$

and the maximum is attained at $\boldsymbol{L}_i = \boldsymbol{P}_i$.

(d) Let $\boldsymbol{B}$ be any matrix of rank q. Then

$$\min_{\boldsymbol{B}} \|\boldsymbol{\Sigma}-\boldsymbol{B}\| = (\lambda_{q+1}^2+\ldots+\lambda_p^2)^{\frac{1}{2}} \quad \ldots \quad (2.5)$$

and the minimum is attained when

$$\boldsymbol{B} = \lambda_1 \boldsymbol{P}_1\boldsymbol{P}_1'+\ldots+\lambda_q\boldsymbol{P}_q\boldsymbol{P}_q'. \quad \ldots \quad (2.6)$$

In (2.5), the symbol $\|\boldsymbol{A}\|$ denotes the Euclidean norm, which is the square root of the sum of squares of the elements of $\boldsymbol{A}$. Or in other words, the choice of $\boldsymbol{B}$ as in (2.6) is the best fitting matrix of given rank q to $\boldsymbol{\Sigma}$.

(ii) *Eigen values and vectors associated with a pair of matrices.* Let $\boldsymbol{\Sigma}$ and $\boldsymbol{\Gamma}$ be two symmetric matrices of order $(p\times p)$ such that $\boldsymbol{\Gamma}$ is positive definite. Consider the determinantal equation

$$|\boldsymbol{\Sigma}-\lambda\boldsymbol{\Gamma}| = 0. \quad \ldots \quad (2.7)$$

The equation (2.7) has p roots, $\lambda_1 \geqslant \lambda_2 \geqslant \ldots \geqslant \lambda_p$ which may be called eigen values of $\boldsymbol{\Sigma}$ with respect to $\boldsymbol{\Gamma}$. Corresponding to each root λ_i there is a vector $\boldsymbol{P}_i$ such that

$$\boldsymbol{\Sigma}\,\boldsymbol{P}_i = \lambda_i\,\boldsymbol{\Gamma}\,\boldsymbol{P}_i,$$

which is called an eigen vector. Then the following results hold :

(a) $\boldsymbol{P}_1, \ldots, \boldsymbol{P}_p$ can be chosen such that

$$\boldsymbol{P}_i'\,\boldsymbol{\Gamma}\,\boldsymbol{P}_i = 1,\ \boldsymbol{P}_i'\,\boldsymbol{\Gamma}\,\boldsymbol{P}_j = 0, \quad i \neq j \quad \ldots \quad (2.8)$$
$$i, j = 1, \ldots, p.$$

(b)
$$\boldsymbol{\Sigma} = \lambda_1\boldsymbol{P}_1\,\boldsymbol{P}_1'+\ldots+\lambda_p\boldsymbol{P}_p\boldsymbol{P}_p'$$
$$\boldsymbol{\Gamma} = \boldsymbol{P}_1\,\boldsymbol{P}_1'+\ldots+\boldsymbol{P}_p\,\boldsymbol{P}_p'. \quad \ldots \quad (2.9)$$

(c) Let $\boldsymbol{L}_1, \ldots, \boldsymbol{L}_q$ be any set of vectors satisfying the same conditions as $\boldsymbol{P}_i$ in (2.8). Then

$$\sum_{i=1}^{q} \boldsymbol{L}_i'\boldsymbol{\Sigma}\boldsymbol{L}_i \leqslant \sum_{i=1}^{q} \boldsymbol{P}_i'\boldsymbol{\Sigma}\boldsymbol{P}_i = \lambda_1+\ldots+\lambda_q \quad \ldots \quad (2.10)$$
$$q = 1, \ldots, p.$$

(iii) *The Eigen values and vectors under restrictions.* Consider a symmetric $p\times p$ matrix $\boldsymbol{\Sigma}$ and a $p\times k$ matrix $\boldsymbol{C}$ of rank k. Let $\boldsymbol{L}_1, \ldots, \boldsymbol{L}_q$ be p dimensional vectors satisfying the conditions

$$\left.\begin{array}{ll} \text{(a)} & \boldsymbol{L}_i'\,\boldsymbol{L}_i = 1,\ \boldsymbol{L}_i'\,\boldsymbol{L}_j = 0 \quad \text{for } i \neq j \\ \text{(b)} & \boldsymbol{L}_i'\,\boldsymbol{C} = 0, \quad i = 1, \ldots, q. \end{array}\right\} \quad \ldots \quad (2.11)$$

Then the maximum of

$$\boldsymbol{L}_1'\,\boldsymbol{\Sigma}\,\boldsymbol{L}_1+\ldots+\boldsymbol{L}_q'\,\boldsymbol{\Sigma}\,\boldsymbol{L}_q \quad \ldots \quad (2.12)$$

is attained when $\boldsymbol{L}_i = \boldsymbol{R}_i$, the i-th eigen vector of the matrix $(\boldsymbol{I}-\boldsymbol{C}(\boldsymbol{C}'\boldsymbol{C})^{-1}\boldsymbol{C}')\boldsymbol{\Sigma}$. The maximum value of (2.12) is then $\nu_1+\ldots+\nu_q$ where ν_i is the i-th eigen value of the matrix $(\boldsymbol{I}-\boldsymbol{C}(\boldsymbol{C}'\boldsymbol{C})^{-1}\boldsymbol{C}')\boldsymbol{\Sigma}$. It is easy to see that if ν_i, $\boldsymbol{R}_i$ are the eigen values and vectors

of $(\boldsymbol{I}-\boldsymbol{C}(\boldsymbol{C}'\boldsymbol{C})^{-1}\boldsymbol{C}')\boldsymbol{\Sigma}$ and μ_i, $\boldsymbol{M}_i$ are the eigen values and vectors of the symmetric matrix $\boldsymbol{\Sigma}^{1/2}(\boldsymbol{I}-\boldsymbol{C}(\boldsymbol{C}'\boldsymbol{C})^{-1}\boldsymbol{C})\boldsymbol{\Sigma}^{1/2}$, then $\lambda_i = \mu_i$ and $\boldsymbol{R}_i = \gamma_i\ \boldsymbol{\Sigma}^{-1/2}\boldsymbol{M}_i$, where $\gamma_i = \boldsymbol{M}_i'$ $\boldsymbol{\Sigma}^{-1}\boldsymbol{M}_i$. Thus the eigen vectors $\boldsymbol{R}_i$ are obtained from the eigen vectors of a symmetric matrix.

(iv) *Relation between the eigen values and vectors of the matrices* $\boldsymbol{AA}'$ *and* $\boldsymbol{A}'\boldsymbol{A}$. Let $\boldsymbol{A}$ be a matrix of order $n\times p$, and rank r. Then $\boldsymbol{AA}'$ and $\boldsymbol{A}'\boldsymbol{A}$ are both symmetric and have the same non-zero eigen values $\lambda_1, \ldots, \lambda_r$. The multiplicity of the zero root is $n-r$ for $\boldsymbol{AA}'$ and $p-r$ for $\boldsymbol{A}'\boldsymbol{A}$. Further, let $\boldsymbol{P}_i$ be the eigen vector of $\boldsymbol{AA}'$ and $\boldsymbol{Q}_i$ be the eigen vector of $\boldsymbol{A}'\boldsymbol{A}$ corresponding to the same non-zero eigen value λ_i. Then $\boldsymbol{Q}_i = \lambda_i^{-1/2}\ \boldsymbol{A}'\boldsymbol{P}_i$ and $\boldsymbol{P}_i = \lambda_i^{-1/2}\,\boldsymbol{A}\boldsymbol{Q}_i$, so that $\boldsymbol{P}_i$ can be obtained from $\boldsymbol{Q}_i$ and vice versa.

(v) Let $\boldsymbol{A}$ be $n\times p$ matrix of rank r. To determine a matrix $\boldsymbol{B}$ of order $n\times p$ and of rank $k \leqslant r$ such that $\|\boldsymbol{A}-\boldsymbol{B}\|$ is a minimum.

Let $\boldsymbol{B} = \boldsymbol{CD}$ where $\boldsymbol{C}$ is $n\times k$ matrix with orthonormal columns. For given $\boldsymbol{C}$, it is easily shown that $\|\boldsymbol{A}-\boldsymbol{B}\|$ is a minimum when $\boldsymbol{D} = \boldsymbol{C}'\boldsymbol{A}$. With such a choice of $\boldsymbol{D}$

$$\begin{aligned}\|\boldsymbol{A}-\boldsymbol{B}\|^2 &= \|(\boldsymbol{I}-\boldsymbol{CC}')\boldsymbol{A}\|^2\\ &= \text{trace}\ [(\boldsymbol{I}-\boldsymbol{CC}')\boldsymbol{A}\ \boldsymbol{A}'(\boldsymbol{I}-\boldsymbol{CC}')]\\ &= \text{trace}\ [\boldsymbol{AA}'(\boldsymbol{I}-\boldsymbol{CC}')]\\ &= \text{trace}\ \boldsymbol{AA}'-\text{trace}\ \boldsymbol{AA}'\ \boldsymbol{CC}'\\ &= \text{trace}\ \boldsymbol{AA}'-\text{trace}\ \boldsymbol{C}'\boldsymbol{A}\ \boldsymbol{A}'\boldsymbol{C}.\end{aligned}$$

We need choose $\boldsymbol{C}$ such that trace $\boldsymbol{C}'\boldsymbol{A}\ \boldsymbol{A}'\boldsymbol{C}$ is a maximum. Then using the result (2.4), the columns of $\boldsymbol{C}$ are the first k eigen vectors of $\boldsymbol{AA}'$.

(vi) Let $\boldsymbol{L}_1, \ldots, \boldsymbol{L}_q$ be p dimensional vectors satisfying the conditions

(a) $\boldsymbol{L}_i'\ \boldsymbol{\Lambda}\ \boldsymbol{L}_i = 1, \qquad \boldsymbol{L}_i'\ \boldsymbol{\Lambda}\ \boldsymbol{L}_j = 0 \quad \text{for} \quad i \neq j$

(b) $\boldsymbol{L}_i'\ \boldsymbol{C} = 0, \qquad i = 1, \ldots, q$

where $\boldsymbol{\Lambda}$ is a positive definite matrix and $\boldsymbol{C}$ is $p\times k$ matrix of rank k. Then the maximum of

$$\boldsymbol{L}_1'\ \boldsymbol{\Sigma}\ \boldsymbol{L}_1 + \ldots + \boldsymbol{L}_q'\ \boldsymbol{\Sigma}\ \boldsymbol{L}_q$$

is attained when $\boldsymbol{L}_i = \boldsymbol{R}_i$, the i-th eigen vector of $\boldsymbol{\Sigma}-\boldsymbol{C}(\boldsymbol{C}'\boldsymbol{\Lambda}^{-1}\ \boldsymbol{C})^{-1}\ \boldsymbol{C}'\boldsymbol{\Lambda}^{-1}\ \boldsymbol{\Sigma}$ with respect to $\boldsymbol{\Lambda}$.

The problem is reduced to that of (iii) by considering the vectors $\boldsymbol{M}_i = \boldsymbol{G}'\boldsymbol{L}_i$ where $\boldsymbol{GG}' = \boldsymbol{\Lambda}$, $i = 1, \ldots, q$. In terms of $\boldsymbol{M}_i$, the problem is the same as in (iii).

3. Fitting a subspace to a set of points in a higher dimensional space

Principal component analysis involving the application of eigen values and vectors of a matrix was first encountered by Karl Pearson (1901) and Frisch (1929) in the problem of fitting a line, a plane or in general a subspace to a scatter of points in a higher dimensional space. Let

$$\mathscr{X} = \begin{pmatrix} X_{11} \dots X_{1n} \\ \dots \dots \\ X_{p1} \dots X_{pn} \end{pmatrix} \quad \dots (3.1)$$

represent n points (each column vector representing a point) in a p dimensional space. Any typical point is represented by $\boldsymbol{X}$. The i-th point is $\boldsymbol{X}_i$ and the centre of gravity of the points is $\bar{\boldsymbol{X}}$, with the s-th coordinate

$$\bar{X}_s = \sum_{j=1}^{n} X_{sj} \div n. \quad \dots (3.2)$$

The points $\boldsymbol{X}_i$ as measured from the centre of gravity are

$$\mathscr{X}_d = \begin{pmatrix} X_{11}-\bar{X}_1 \dots X_{1n}-\bar{X}_1 \\ \dots \quad \dots \quad \dots \\ X_{p1}-\bar{X}_p \dots X_{pn}-\bar{X}_p \end{pmatrix}. \quad \dots (3.3)$$

Let us define

$$\boldsymbol{\Sigma} = \mathscr{X}_d \, \mathscr{X}_d' = (\sigma_{ij}) \quad \dots (3.4)$$

as the total dispersion matrix (or scatter) of the points. In (3.4) the element σ_{ij} is computed by the formula

$$\sigma_{ij} = \sum_{r=1}^{n} (X_{ir} - \bar{X}_i)(X_{jr} - \bar{X}_j) = \sum_{r=1}^{n} X_{ir}X_{jr} - n\,\bar{X}_i\bar{X}_j.$$

Now a q dimensional subspace is specified by a point (origin) and q orthogonal axes (each axis specified by its direction cosines) passing through it. Pearson defined a best fitting subspace as that for which the sum of squares of the perpendiculars from the points to the subspace is a minimum. It is easily shown that such a subspace passes through the centre of gravity of the points. Further, the sum of squares of the perpendiculars from the points to the subspace defined by the centre of gravity and q orthogonal vectors $\boldsymbol{L}_1, \dots, \boldsymbol{L}_q$ is

$$\sum_{i=1}^{p} \sigma_{ii} - \sum_{i=1}^{q} \boldsymbol{L}_i' \boldsymbol{\Sigma} \boldsymbol{L}_i. \quad \dots (3.5)$$

Minimizing (3.5) is the same as maximizing $\Sigma\ \boldsymbol{L}_i' \boldsymbol{\Sigma} \boldsymbol{L}_i$ and an application of the result (2.4) shows that the maximum is attained when

$$\boldsymbol{L}_i = \boldsymbol{P}_i, \quad i=1, \dots, q \quad \dots (3.6)$$

where $\boldsymbol{P}_1, \dots, \boldsymbol{P}_q$ are the first q eigen vectors of the matrix $\boldsymbol{\Sigma}$.

What we need in practice is an actual representation of points in the best fitting lower dimensional subspace. This is simply done by computing for each $\boldsymbol{X}$ the q coordinates

$$\boldsymbol{P}_1' \boldsymbol{X}, \dots, \boldsymbol{P}_q' \boldsymbol{X}. \quad \dots (3.7)$$

Thus the points in the projected space referred to q orthogonal axes are

$$\mathcal{Y} = \begin{pmatrix} \boldsymbol{P}_1'\boldsymbol{X}_1 \ldots \boldsymbol{P}_1'\boldsymbol{X}_n \\ \cdot \quad \ldots \quad \cdot \\ \boldsymbol{P}_q'\boldsymbol{X}_1 \ldots \boldsymbol{P}_q'\boldsymbol{X}_n \end{pmatrix}. \qquad \ldots \quad (3.8)$$

The first row of $\mathcal{Y}$ gives the best one dimensional representation, the first two rows of $\mathcal{Y}$ give the best two dimensional representation, and so on.

In practice, it is also necessary to know how good the representation of points in a lower dimensional space is. A suitable criterion for this purpose is the sum of squares of the distances between the original and the projected points. For the best q space, the criterion has the value

$$\lambda_{q+1}+\ldots+\lambda_p \qquad \ldots \quad (3.9)$$

using the result (2.4). Instead of the absolute value (3.9), we may choose as a measure of goodness of fit the ratio

$$\frac{\lambda_{q+1}+\ldots+\lambda_p}{\lambda_1+\ldots+\lambda_p}. \qquad \ldots \quad (3.10)$$

The choice of q, then depends on the smallness of (3.10) or the largeness of the ratio

$$\frac{\lambda_1+\ldots+\lambda_q}{\lambda_1+\ldots+\lambda_p}. \qquad \ldots \quad (3.11)$$

It may happen that the overall measure (3.11) has a small value but the configuration of the points as a whole is distorted in the projected space due to some isolated points being far away from the best fitting space. The examination of such isolated points is of interest in practical work and may provide an interpretation of the nature of heterogeneity in the data. For this purpose it is necessary to compute the length of the perpendicular of each point on the best fitting space. The square of the perpendicular from $\boldsymbol{X}_i$, the i-th point is

$$d_i^2 = (\boldsymbol{X}_i-\bar{\boldsymbol{X}})'(\boldsymbol{X}_i-\bar{\boldsymbol{X}})-[\boldsymbol{P}_1'(\boldsymbol{X}_i-\bar{\boldsymbol{X}})]^2-\ldots-[\boldsymbol{P}_q'(\boldsymbol{X}_i-\bar{\boldsymbol{X}})]^2$$
$$i = 1, \ldots, n.$$

An examination of the values $d_1^2, \ldots, d_n^2$ will enable us to find out the outliers if any.

4. OTHER INTERPRETATIONS OF THE TRANSFORMATION (3.8)

The problem of representing the points (3.1) in a lower dimensional space may be posed as the determination of a transformation matrix $\boldsymbol{T}$ of order $p \times q$, and rank q transforming a p-vector $\boldsymbol{X}$ into a q-vector $\boldsymbol{Y}$,$\boldsymbol{Y} = \boldsymbol{T}'\boldsymbol{X}$, and satisfying some optimum properties. The transformation from $\boldsymbol{Y}$ to $\boldsymbol{X}$ is not one to one when $q < p$ and consequently certain properties of the configuration of the points in the original space (which is invariant when $q = p$) are altered. Our aim is to choose $\boldsymbol{T}$ which preserves the configuration of the points to the maximum possible extent. We shall show that some intuitive measures of closeness between the configurations in the original space and the transformed subspace, when minimized with respect to the transformation matrix $\boldsymbol{T}$ lead to the transformation (3.8) as the optimum.

Let us observe that the columns of $\boldsymbol{T}$ can be chosen to be orthonormal, without loss of generality.

Maximizing the sum of squares of the distances. The sum of squares of the distances between all possible pairs of points in the p-space is

$$\sum_{i,j=1}^{n}\sum (\boldsymbol{X}_i-\boldsymbol{X}_j)'(\boldsymbol{X}_i-\boldsymbol{X}_j) = n \text{ trace } \boldsymbol{\Sigma}, \qquad \ldots \quad (4.1)$$

while the corresponding expresion in the q-space is

$$\begin{aligned}\sum_{i,j=1}^{n}\sum (\boldsymbol{Y}_i-\boldsymbol{Y}_j)'(\boldsymbol{Y}_i-\boldsymbol{Y}_j) &= n \text{ trace } \boldsymbol{T}' \boldsymbol{\Sigma} \boldsymbol{T} \\ &= n(\boldsymbol{T}_1' \boldsymbol{\Sigma} \boldsymbol{T}_1+\ldots+\boldsymbol{T}_q' \boldsymbol{\Sigma} \boldsymbol{T}_q) \qquad \ldots \quad (4.2)\end{aligned}$$

where $\boldsymbol{T}_1, \ldots, \boldsymbol{T}_q$ are the columns of $\boldsymbol{T}$. It is seen that (4.2) $\leqslant$ (4.1) and (4.2) = (4.1) when (not necessarily only when) $p = q$. Let us choose $\boldsymbol{T}$ such that (4.2) is a maximum. Applying the result (2.4), we find that the maximum is attained when $\boldsymbol{T}_i^* = \boldsymbol{P}_i$, $i = 1, \ldots, q$, which is the solution obtained in Section 3.

Closest fit to the distances and the angles of the lines joining the points to the centre of gravity. In the p-space the vectors joining the points to their centre of gravity are represented by the matrix $\mathcal{X}_d$ as defined in (3.3). It may be seen that in the matrix

$$\boldsymbol{\Lambda} = \mathcal{X}_d' \mathcal{X}_d \qquad \ldots \quad (4.3)$$

the i-th diagonal entry is the square of the distance of the i-th point from the centre of gravity and the (i, j) entry divided by the square root of the i-th and j-th entries is the cosine of the angle between the lines joining the i-th and j-th points to the centre of gravity. Thus the matrix $\boldsymbol{\Lambda}$ represents the distances and the angles between the lines joining the points to their centre of gravity. In the q-space, under the transformation $\boldsymbol{Y} = \boldsymbol{T}'\boldsymbol{X}$ the matrix corresponding to (4.3) is

$$\boldsymbol{B} = \mathcal{Y}_d' \mathcal{Y}_d = \mathcal{X}_d' \boldsymbol{T} \boldsymbol{T}' \mathcal{X}_d. \qquad \ldots \quad (4.4)$$

We wish to determine $\boldsymbol{T}$ such that the matrix $\boldsymbol{B}$ is as close as possible to $\boldsymbol{\Lambda}$. We may measure the closeness of $\boldsymbol{B}$ and $\boldsymbol{\Lambda}$ by the Euclidean norm $\|\boldsymbol{\Lambda}-\boldsymbol{B}\|$, which is the square root of the sum of squares of the elements in $(\boldsymbol{\Lambda}-\boldsymbol{B})$. Then the problem is one choosing $\boldsymbol{B}$ such that $\|\boldsymbol{\Lambda}-\boldsymbol{B}\|$ is a minimum. An application of the result (2.6) gives the optimum choice of $\boldsymbol{B}$ as

$$\mathcal{X}_d'\boldsymbol{T} \ \boldsymbol{T}'\mathcal{X}_d = \lambda_1\boldsymbol{Q}_1\boldsymbol{Q}_1'+\ldots+\lambda_q \ \boldsymbol{Q}_q \ \boldsymbol{Q}_q' \qquad \ldots \quad (4.5)$$

where $\boldsymbol{Q}_1, \ldots, \boldsymbol{Q}_q$ are the eigen vectors of $\boldsymbol{\Lambda} = \mathcal{X}_d'\mathcal{X}_d$. The optimum choice of $\boldsymbol{T}$ is obtained from (4.5) as

$$\begin{aligned}\mathcal{X}_d \ \boldsymbol{T}^* &= (\sqrt{\lambda_1}\boldsymbol{Q}_1 \ldots \sqrt{\lambda_q}\ \boldsymbol{Q}_q) \\ &= (\boldsymbol{X}_d' \boldsymbol{P}_1 \ldots \boldsymbol{X}_d' \boldsymbol{P}_q) \qquad \ldots \quad (4.6)\end{aligned}$$

using the result (iv) of Section 2. The choice

$$\boldsymbol{T}^* = (\boldsymbol{P}_1 \ldots \boldsymbol{P}_q) \qquad \ldots \quad (4.7)$$

satisfies the equation (4.6). Thus $\boldsymbol{T}_i^* = \boldsymbol{P}_i$, where $\boldsymbol{T}_i^*$ is the i-th column of $\boldsymbol{T}^*$, which is the solution obtained in Section 3. This means that the representation of the points in a q-dimensional space as in (3.8) preserves the distances and angles of the configuration to a large extent.

5. Reduction in the number of points

In Sections 3 and 4, a reduction in the number of dimensions of the space is secured through a transformation of the type $\boldsymbol{Y} = \boldsymbol{T}'\boldsymbol{X}$ by choosing an optimum $\boldsymbol{T}$. We now consider the dual problem of reducing the number of points to a few *typical points* keeping the number of dimensions the same. For this purpose we consider the matrix $\mathcal{X}_d$ which represents the points with the origin at the centre of gravity. The reduction in the number of points is secured by a transformation of the type $\mathcal{Y}_d = \mathcal{X}_d \boldsymbol{M}$ where $\boldsymbol{M}$ is $n \times q$ matrix.

Reduction providing the closest fit to the total dispersion matrix. The total dispersion matrix based on the reduced points is

$$\mathcal{Y}_d \mathcal{Y}_d' = \mathcal{X}_d \boldsymbol{M}\boldsymbol{M}' \mathcal{X}_d' \qquad \ldots \quad (5.1)$$

while that based on all points is $\boldsymbol{\Sigma}$. We wish to choose $\mathcal{Y}_d$ in such a way that

$$\| \boldsymbol{\Sigma} - \mathcal{Y}_d \mathcal{Y}_d' \| \qquad \ldots \quad (5.2)$$

is a minimum. Applying the result (2.6), the minimum of (5.2) is attained when

$$\mathcal{Y}_d \mathcal{Y}_d' = \lambda_1 \boldsymbol{P}_1 \boldsymbol{P}_1' + \ldots + \lambda_q \boldsymbol{P}_q \boldsymbol{P}_q'$$

or when

$$\mathcal{Y}_d = (\sqrt{\lambda_1}\boldsymbol{P}_1 \ldots \sqrt{\lambda_q}\boldsymbol{P}_q). \qquad \ldots \quad (5.3)$$

Thus the typical points summarizing the deviations from the centre of gravity of the original points are the eigen vectors scaled by the square roots of the corresponding eigen values.

To determine the transformation matrix (if necessary) let us observe that, using result (iv) of Section 2,

$$\begin{aligned} \mathcal{Y}_d &= (\sqrt{\lambda_1}\boldsymbol{P}_1 \ldots \sqrt{\lambda_q}\boldsymbol{P}_q) \\ &= (\mathcal{X}_d\boldsymbol{Q}_1 \ldots \mathcal{X}_d\boldsymbol{Q}_q), \end{aligned} \qquad \ldots \quad (5.4)$$

where $\boldsymbol{Q}_1, \ldots, \boldsymbol{Q}_q$ are the first q eigen vectors of the matrix $\mathcal{X}_d' \mathcal{X}_d$. But (5.4) is the same as $\mathcal{X}_d \boldsymbol{M}^*$ where $\boldsymbol{M}^*$ is the matrix with $\boldsymbol{Q}_1, \ldots, \boldsymbol{Q}_q$ as its columns.

An interpretation of the typical points. In Section 3, it is shown that the best fitting q-space is specified by $\bar{\boldsymbol{X}}$, the centre of gravity and the eigen vectors $\boldsymbol{P}_1, \ldots, \boldsymbol{P}_q$. Any point in this space referred to the p *original axes* can be represented by a vector of the form

$$\bar{\boldsymbol{X}} + a_1\boldsymbol{P}_1 + \ldots + a_q\boldsymbol{P}_q \qquad \ldots \quad (5.5)$$

where $a_1, \ldots, a_q$ are arbitrary. By a change in scale, (5.5) can be written as

$$\bar{\boldsymbol{X}} + b_1\sqrt{\lambda_1}\boldsymbol{P}_1 + \ldots + b_q\sqrt{\lambda_q}\boldsymbol{P}_q \qquad \ldots \quad (5.6)$$

involving the typical points $\sqrt{\lambda_1}\boldsymbol{P}_1, \ldots, \sqrt{\lambda_q}\boldsymbol{P}_q$. Let $\boldsymbol{X}_i^{(q)}$ be the projection of $\boldsymbol{X}_i$ on this subspace. Then there exist constants $b_{1i}, \ldots, b_{qi}$ such that

$$\boldsymbol{X}_i^{(q)} = \bar{\boldsymbol{X}} + b_{1i}\sqrt{\lambda_1}\boldsymbol{P}_1 + \ldots + b_{qi}\sqrt{\lambda_q}\boldsymbol{P}_q. \qquad \ldots \quad (5.7)$$

Multiplying both sides of (5.7) by $\boldsymbol{P}_j'$ and observing that $\boldsymbol{P}_j'\boldsymbol{X}_i^{(g)} = \boldsymbol{P}_j'\boldsymbol{X}_i$ we obtain the value of b_{ji} as

$$b_{ji} = \lambda^{-1/2}\boldsymbol{P}_j'(\boldsymbol{X}_i - \bar{\boldsymbol{X}}). \quad \ldots \quad (5.8)$$

With such a choice of b_{ji}

$$\boldsymbol{X}_i^{(g)} = \bar{\boldsymbol{X}} + b_{1i}\sqrt{\lambda_1}\boldsymbol{P}_1 + \ldots + b_{qi}\sqrt{\lambda_q}\boldsymbol{P}_q \quad \ldots \quad (5.9)$$

$$i = 1, \ldots, n.$$

The point $\boldsymbol{X}_i^{(g)}$ so determined may be called the graduated or the fitted point of $\boldsymbol{X}_i$. If $\boldsymbol{X}_i - \boldsymbol{X}_i^{(g)}$ is small for each i, the result (5.9) shows that the average point and the typical points

$$\bar{\boldsymbol{X}}, \sqrt{\lambda_1}\boldsymbol{P}_1, \ldots, \sqrt{\lambda_q}\boldsymbol{P}_q$$

form an *approximate basis* of the n observed points

$$\boldsymbol{X}_i \sim \bar{\boldsymbol{X}} + b_{1i}\sqrt{\lambda_1}\boldsymbol{P}_1 + \ldots + b_{qi}\sqrt{\lambda_q}\boldsymbol{P}_q \quad \ldots \quad (5.10)$$

$$i = 1, \ldots, n.$$

The representation (5.10) admits a statistical interpretation under the following model for the observed points $\boldsymbol{X}_i$:

$$\boldsymbol{X}_i = \boldsymbol{\mu} + \beta_{1i}\boldsymbol{\pi}_1 + \ldots + \beta_{qi}\boldsymbol{\pi}_q + \boldsymbol{\epsilon}_i \quad \ldots \quad (5.11)$$

where $\boldsymbol{\mu}, \boldsymbol{\pi}_1, \ldots, \boldsymbol{\pi}_q$ are fixed vectors, $\beta_{1i}, \ldots, \beta_{qi}$ are constants specific to the i-th point and $\boldsymbol{\epsilon}_i$ is a vector of random errors. The model (5.11) implies that any observed point when corrected for errors of measurement can be expressed as linear combination of $(q+1)$ basic vectors only. It can be shown that, when the dispersion matrix of $\boldsymbol{\epsilon}_i$ is of the form $\sigma^2\boldsymbol{I}$, consistent estimators of $\boldsymbol{\mu}$, $\boldsymbol{\pi}_i$ and β_{ji} are provided by $\bar{\boldsymbol{X}}$, $\sqrt{\lambda_i}\boldsymbol{P}_i$ and b_{ji} as defined in (5.10). Under the additional condition that the p components of $\boldsymbol{\epsilon}_i$ have independent normal distributions, $\bar{\boldsymbol{X}}$, $\sqrt{\lambda_i}\boldsymbol{P}_i$ and b_{ji} can be shown to be maximum likelihood estimators. Such a statistical interpretation is not true when the dispersion matrix of $\boldsymbol{\epsilon}_i$ is not of the form $\sigma^2\boldsymbol{I}$.

The typical points $\sqrt{\lambda_1}\boldsymbol{P}_1, \ldots, \sqrt{\lambda_q}\boldsymbol{P}_q$ used in the representation (5.10) may not admit any physical interpretation. But in special situations it may be possible to characterize each typical point as indicating some aspect of the measurements as a whole (see Simonds, 1964).

6. Fitting a subspace when the points are in an oblique space

In the analysis of Section 3, it is implicitly assumed that the original points are represented in a p-dimensional Euclidean space with orthogonal axes. There are, however, situations where oblique axes are chosen to represent the coordinates. In such a case, the distance between any two points $\boldsymbol{X}_i$, $\boldsymbol{X}_j$ is a quadratic form

$$(\boldsymbol{X}_i-\boldsymbol{X}_j)'\boldsymbol{\Gamma}^{-1}(\boldsymbol{X}_i-\boldsymbol{X}_j) \quad \ldots \quad (6.1)$$

using a positive definite matrix $\boldsymbol{\Gamma}$. The scatter of points analogous to (3.10) is

$$\sum_{i,j=1}^{n}\sum (\boldsymbol{X}_i-\boldsymbol{X}_j)'\boldsymbol{\Gamma}^{-1}(\boldsymbol{X}_i-\boldsymbol{X}_j). \quad \ldots \quad (6.2)$$

Let us transform $\boldsymbol{X}$ to $\boldsymbol{Y}$ with q coordinates

$$\boldsymbol{Y}' = (\boldsymbol{L}_1'\boldsymbol{X}, \ldots, \boldsymbol{L}_q'\boldsymbol{X})$$

where $\boldsymbol{L}_i'\boldsymbol{\Gamma}\boldsymbol{L}_i = 1$ and $\boldsymbol{L}_i'\boldsymbol{\Gamma}\boldsymbol{L}_j = 0$, $i \neq j$. The scatter of points in the q-space with $\boldsymbol{L}_1, \ldots, \boldsymbol{L}_q$ as orthogonal axes is

$$\Sigma\Sigma(\boldsymbol{Y}_i-\boldsymbol{Y}_j)'(\boldsymbol{Y}_i-\boldsymbol{Y}_j) = n(\boldsymbol{L}_1'\boldsymbol{\Sigma}\boldsymbol{L}_1+\ldots+\boldsymbol{L}_q'\boldsymbol{\Sigma}\boldsymbol{L}_q). \quad \ldots \quad (6.3)$$

An application of the result (2.10) shows that (6.3) is a maximum when $\boldsymbol{L}_i = \boldsymbol{P}_i$, $i = 1, \ldots, q$, where $\boldsymbol{P}_1, \ldots, \boldsymbol{P}_q$ are the first q eigen vectors associated with the determinantal equation $|\boldsymbol{\Sigma}-\boldsymbol{\Gamma}| = 0$. The adequacy of a q dimensional fit is judged by the largeness of the ratio

$$\frac{\lambda_1+\ldots+\lambda_q}{\lambda_1+\ldots+\lambda_p} \quad \ldots \quad (6.4)$$

where $\lambda_1, \ldots, \lambda_p$ are the roots of $|\boldsymbol{\Sigma}-\lambda\boldsymbol{\Gamma}| = 0$.

7. Principal components of a vector random variable

The concept of principal component analysis as applied to a random variable is due to Hotelling (1933, 1935, 1936a).

Let us consider a p dimensional random variable $\boldsymbol{X}$ with mean $\boldsymbol{\mu}$ and dispersion matrix $\boldsymbol{\Sigma}$. Consider a transformation, $\boldsymbol{Y} = \boldsymbol{T}'\boldsymbol{X}$, where $\boldsymbol{T}$ is a $p\times q$ matrix, so that $\boldsymbol{Y}$ is q dimensional. When $q < p$, there is loss of information in replacing $\boldsymbol{X}$ by $\boldsymbol{Y}$. For a given q, we wish to determine $\boldsymbol{T}$ such that there is minimum loss.

In defining a loss function we may be guided by the question as to what extent we can predict $\boldsymbol{X}$ knowing $\boldsymbol{Y}$. The pedictive efficiency of $\boldsymbol{Y}$ for $\boldsymbol{X}$ depends on the residual dispersion matrix of $\boldsymbol{X}$ after subtracting its best linear predictor in terms of $\boldsymbol{Y}$. Now the joint dispersion matrix of $\boldsymbol{X}$ and $\boldsymbol{Y}$ is

$$\begin{pmatrix} \boldsymbol{\Sigma} & \boldsymbol{\Sigma T} \\ \boldsymbol{T'\Sigma} & \boldsymbol{T'\Sigma T} \end{pmatrix} \quad \ldots \quad (7.1)$$

and the residual dispersion matrix is

$$\boldsymbol{\Sigma}-\boldsymbol{\Sigma T}(\boldsymbol{T}'\boldsymbol{\Sigma T})^{-1}\boldsymbol{T}'\boldsymbol{\Sigma}. \quad \ldots \quad (7.2)$$

The smaller the values of the elements in (7.2), the greater is the predictive efficiency.

We consider two overall measures, the trace and the Euclidean norms of (7.2),

$$\text{(a)} \quad \text{trace } (\boldsymbol{\Sigma}-\boldsymbol{\Sigma}\, \boldsymbol{T}(\boldsymbol{T}'\, \boldsymbol{\Sigma}\, \boldsymbol{T})^{-1}\boldsymbol{T}'\boldsymbol{\Sigma}) \qquad \ldots \quad (7.3)$$

$$\text{(b)} \quad \|\boldsymbol{\Sigma}-\boldsymbol{\Sigma}\, \boldsymbol{T}(\boldsymbol{T}'\, \boldsymbol{\Sigma}\boldsymbol{T})^{-1}\boldsymbol{T}'\boldsymbol{\Sigma}\|. \qquad \ldots \quad (7.4)$$

We shall show that any one of these measures when minimized leads to the same choice of $\boldsymbol{T}$ as the matrix of the first q eigen vectors of $\boldsymbol{\Sigma}$.

Let us observe that the column vectors $\boldsymbol{T}_1, \ldots, \boldsymbol{T}_q$ of $\boldsymbol{T}$ can be chosen, without loss of generality, to satisfy the conditions

$$\boldsymbol{T}_i'\, \boldsymbol{\Sigma}\, \boldsymbol{T}_j = 0, \quad i \neq j \qquad \ldots \quad (7.5)$$

which imply that the components of the transformed variable are uncorrelated.

Consider the measure (7.3)

$$\begin{aligned} &\text{trace } (\boldsymbol{\Sigma}-\boldsymbol{\Sigma}\, \boldsymbol{T}(\boldsymbol{T}'\, \boldsymbol{\Sigma}\, \boldsymbol{T})^{-1}\boldsymbol{T}'\boldsymbol{\Sigma}) \\ &\quad = \text{trace } \boldsymbol{\Sigma} - \text{trace } (\boldsymbol{T}'\, \boldsymbol{\Sigma}\, \boldsymbol{T})^{-1}\boldsymbol{T}'\boldsymbol{\Sigma}\, \boldsymbol{\Sigma}\, \boldsymbol{T}) \\ &\quad = \text{trace } \boldsymbol{\Sigma} - \left(\frac{\boldsymbol{T}_1'\boldsymbol{\Sigma}\, \boldsymbol{\Sigma}\, \boldsymbol{T}_1}{\boldsymbol{T}_1'\boldsymbol{\Sigma}\, \boldsymbol{T}_1}+\ldots+\frac{\boldsymbol{T}_q'\boldsymbol{\Sigma}\, \boldsymbol{\Sigma}\, \boldsymbol{T}_q}{\boldsymbol{T}_q'\, \boldsymbol{\Sigma}\, \boldsymbol{T}_q}\right) \qquad \ldots \quad (7.6) \end{aligned}$$

using the conditions (7.5). Minimizing (7.6) is the same as maximizing

$$\frac{\boldsymbol{T}_1'\boldsymbol{\Sigma}\, \boldsymbol{\Sigma}\, \boldsymbol{T}_1}{\boldsymbol{T}_1'\boldsymbol{\Sigma}\, \boldsymbol{T}_1}+\ldots+\frac{\boldsymbol{T}_q'\boldsymbol{\Sigma}\, \boldsymbol{\Sigma}\, \boldsymbol{T}_q}{\boldsymbol{T}_q'\boldsymbol{\Sigma}\, \boldsymbol{T}_q}.$$

Applying the result (2.10) the optimum choice of $\boldsymbol{T}_i$ is the i-th eigen vector $\boldsymbol{\Sigma}\boldsymbol{\Sigma}$ with respect to $\boldsymbol{\Sigma}$, which is the same as $\boldsymbol{P}_i$ the i-th eigen vector of $\boldsymbol{\Sigma}$. The minimum value of (7.6) is then

$$\lambda_{q+1}+\ldots+\lambda_p \qquad \ldots \quad (7.7)$$

the sum of the smallest $p-q$ eigen values of $\boldsymbol{\Sigma}$.

Consider the problem of minimizing

$$\|\boldsymbol{\Sigma}-\boldsymbol{\Sigma}\boldsymbol{T}(\boldsymbol{T}'\, \boldsymbol{\Sigma}\, \boldsymbol{T})^{-1}\boldsymbol{T}'\, \boldsymbol{\Sigma}\|. \qquad \ldots \quad (7.8)$$

The result (2.6) shows that the minimum of (7.8) is attained when

$$\boldsymbol{\Sigma}\boldsymbol{T}(\boldsymbol{T}'\, \boldsymbol{\Sigma}\, \boldsymbol{T})^{-1}\boldsymbol{T}\, \boldsymbol{\Sigma} = \lambda_1\boldsymbol{P}_1\boldsymbol{P}_1'+\ldots+\lambda_q\boldsymbol{P}_q\boldsymbol{P}_q' \qquad \ldots \quad (7.9)$$

and it is easy to verify that (7.9) holds when $\boldsymbol{T}_i^* = \boldsymbol{P}_i$. The minimum value of (7.8) is

$$\lambda_{q+1}^2\ldots+\lambda_p^2 \qquad \ldots \quad (7.10)$$

where $\lambda_{q+1}, \ldots, \lambda_p$ are as defined in (7.7).

The transformed variables $\boldsymbol{P}'_1\boldsymbol{X}, \ldots, \boldsymbol{P}_q'\boldsymbol{X}$ are called the first q principal components of the random variable $\boldsymbol{X}$ and are interpreted variously. The usual interpretation is as follows. Under a complete orthogonal transformation $\boldsymbol{Y} = \boldsymbol{0}'\boldsymbol{X}$, i.e., from a p dimensional variable to another p dimensional variable, the trace of the dispersion matrix, which is the sum of the variances of the variables, remains invariant,

Thus

$$\text{trace } \boldsymbol{\Sigma} = \text{trace } \mathbf{0}'\boldsymbol{\Sigma}\,\mathbf{0} = \mathbf{0}_1'\boldsymbol{\Sigma}\,\mathbf{0}_1+\ldots+\mathbf{0}_p'\,\boldsymbol{\Sigma}\,\mathbf{0}_p \qquad \ldots \quad (7.11)$$

where $\mathbf{0}'\,\boldsymbol{\Sigma}\,\mathbf{0}$ is the dispersion matrix of $\boldsymbol{Y}$ and $\mathbf{0}_i$ is the i-th column vector of $\mathbf{0}$. The total of the variances of the first q variables in $\boldsymbol{Y}$ is

$$\mathbf{0}_1'\,\boldsymbol{\Sigma}\,\mathbf{0}_1+\ldots+\mathbf{0}_q'\,\boldsymbol{\Sigma}\,\mathbf{0}_q \leqslant \text{trace } \boldsymbol{\Sigma}. \qquad \ldots \quad (7.12)$$

It is seen from (2.6) that the optimum choice of $\mathbf{0}_i$ is $\boldsymbol{P}_i$ for (7.12) to be a maximum, i.e., the first q transformed variables are $\boldsymbol{P}_1'\boldsymbol{X}$, ..., $\boldsymbol{P}_q'\boldsymbol{X}$, the q principal components of $\boldsymbol{X}$.

The maximum value of (7.12) is $\lambda_1+\ldots+\lambda_q$ while the total of the variances of the variables is $\lambda_1+\ldots+\lambda_p$, in which case the first q principal components are said to explain $100(\lambda_1+\ldots+\lambda_q)/(\lambda_1+\ldots+\lambda_p)$ percent of the total variance.

The interpretation of the principal components as the best predictors of $\boldsymbol{X}$ suggests extensions of the principal component analysis in many directions, which are considered in the later sections.

8. Principal components of instrumental variables

Let $\boldsymbol{X}$ be the vector of p main variables, and $\boldsymbol{Z}$ the vector of m instrumental variables. In theory $\boldsymbol{Z}$ may include some or all the elements of $\boldsymbol{X}$. Denote the joint dispersion matrix of $(\boldsymbol{X}, \boldsymbol{Z})$ by

$$\begin{pmatrix} \boldsymbol{\Sigma} & \boldsymbol{\Theta} \\ \boldsymbol{\Theta}' & \boldsymbol{\Gamma} \end{pmatrix}. \qquad \ldots \quad (8.1)$$

We wish to replace $\boldsymbol{Z}$ by a q dimensional random variable $\boldsymbol{Y} = \boldsymbol{M}'\boldsymbol{Z}$ in such a way that the predictive efficiency of $\boldsymbol{Y}$ for $\boldsymbol{X}$ is a maximum, (Rao, 1962a). The dispersion matrix of $(\boldsymbol{X}, \boldsymbol{Y})$ is

$$\begin{pmatrix} \boldsymbol{\Sigma} & \boldsymbol{\Theta}\,\boldsymbol{M} \\ \boldsymbol{M}'\boldsymbol{\Theta}' & \boldsymbol{M}'\boldsymbol{\Gamma}\boldsymbol{M} \end{pmatrix} \qquad \ldots \quad (8.2)$$

and the residual dispersion matrix of $\boldsymbol{X}$ subtracting its best linear predictor in terms of $\boldsymbol{Y}$ is

$$\boldsymbol{\Sigma}-\boldsymbol{\Theta}\boldsymbol{M}(\boldsymbol{M}'\boldsymbol{\Gamma}\,\boldsymbol{M})^{-1}\boldsymbol{M}'\boldsymbol{\Theta}. \qquad \ldots \quad (8.3)$$

As in Section 7, we may consider the two measures of predictive efficiency of $\boldsymbol{Y}$

$$\text{(a)} \quad \text{trace } (\boldsymbol{\Sigma}-\boldsymbol{\Theta}\,\boldsymbol{M}\,(\boldsymbol{M}'\,\boldsymbol{\Gamma}\,\boldsymbol{M})^{-1}\,\boldsymbol{M}'\boldsymbol{\Theta}') \qquad \ldots \quad (8.4)$$

$$\text{(b)} \quad \|\boldsymbol{\Sigma}-\boldsymbol{\Theta}\,\boldsymbol{M}(\boldsymbol{M}'\,\boldsymbol{\Gamma}\,\boldsymbol{M})^{-1}\boldsymbol{M}'\boldsymbol{\Theta}'\|. \qquad \ldots \quad (8.5)$$

Unfortunately, the solution seems to depend on which measure is chosen for minimization.

Minimizing (8.4) is the same as maximizing

$$\begin{aligned} &\text{trace } \boldsymbol{\Theta}\,\boldsymbol{M}(\boldsymbol{M}'\boldsymbol{\Gamma}\,\boldsymbol{M})^{-1}\boldsymbol{M}'\boldsymbol{\Theta}' \\ &\quad = \text{trace } (\boldsymbol{M}'\,\boldsymbol{\Gamma}\,\boldsymbol{M})^{-1}\boldsymbol{M}'\boldsymbol{\Theta}'\;\boldsymbol{\Theta}\,\boldsymbol{M} \\ &\quad = \frac{\boldsymbol{M}_1'\boldsymbol{\Theta}'\,\boldsymbol{\Theta}\boldsymbol{M}_1}{\boldsymbol{M}_1'\,\boldsymbol{\Gamma}\,\boldsymbol{M}_1} + \ldots + \frac{\boldsymbol{M}_q'\,\boldsymbol{\Theta}'\boldsymbol{\Theta}\,\boldsymbol{M}_q}{\boldsymbol{M}_q'\,\boldsymbol{\Gamma}\,\boldsymbol{M}_q} \end{aligned} \qquad \ldots \quad (8.6)$$

assuming that $\boldsymbol{M}_i\,\boldsymbol{\Gamma}\,\boldsymbol{M}_j = 0$, $i \neq j$, without loss of generality. Applying the result (2.10), the best choice of $\boldsymbol{M}_1$, ..., $\boldsymbol{M}_q$ is the set of the first q eigen vectors of the matrix $\boldsymbol{\Theta}'\,\boldsymbol{\Theta}$ with respect to Γ, i.e. associated with the determinantal equation

$$|\boldsymbol{\Theta}'\boldsymbol{\Theta}-\lambda\boldsymbol{\Gamma}| = 0. \qquad \ldots \quad (8.7)$$

It may be obsrved that when $\boldsymbol{X}$ is a proper subset of $\boldsymbol{Z}$, the optimum choice of $\boldsymbol{Y}$ consists of the first q principal components of $\boldsymbol{X}$ alone.

The present analysis is different from that of canonical correlations and canonical variates as developed by Hotelling (1936b) for studying the association between two vector random variables.

There seems to be no simple method of minimizing the measure (8.5). The solution would be, however, different from that obtained by minimizing (8.4). The relative merits of the two solutions are worth exploring.

9. Weighted principal components

The measures of predictive efficiency (7.3), (7.4), (8.4) and (8.5) are not invariant under change of scales of the elements of $\boldsymbol{X}$. This is an undesirable feature as different solutions to principal components can be found by choosing different scales. The difficulty can be overcome by defining a loss function associated with the prediction of each element of $\boldsymbol{X}$ by any chosen set of q linear functions $\boldsymbol{Y}$ of the instrumental variable $\boldsymbol{Z}$. A simple loss function for the prediction of $\boldsymbol{X}_i$, the i-th element of $\boldsymbol{X}$, may be chosen as w_i^2 times the residual variance, where w_i^2, $i = 1, \ldots, p$, are assigned quantities. Denoting by $\boldsymbol{W}$, the diagonal matrix of the elements $w_1, \ldots, w_p$ the total loss in the prediction of $\boldsymbol{X}$ is the weighted trace

$$\text{trace } \boldsymbol{W\Sigma W} - \text{trace } \boldsymbol{W\Theta M}(\boldsymbol{M'\Gamma M})^{-1}\boldsymbol{M'\Theta' W}. \quad \ldots \quad (9.1)$$

The problem is one of maximizing

$$\text{trace } \boldsymbol{W\Theta M M'\Theta' W} = \text{trace } \boldsymbol{M'\Theta' W^2\Theta M}, \quad \ldots \quad (9.2)$$

choosing $\boldsymbol{M'\Gamma M} = \boldsymbol{I}$, without loss of generality. The best choice of $\boldsymbol{M}$ is the matrix of the first q eigen vectors associated with the determinantal equation

$$|\boldsymbol{\Theta' W^2\Theta} - \lambda\boldsymbol{\Gamma}| = 0. \quad \ldots \quad (9.3)$$

In the special case when the instrumental variable is the same as (or contains) the main variable, the equation (9.3) reduces to

$$|\boldsymbol{\Sigma W^2\Sigma} = \lambda\boldsymbol{\Sigma}| = 0 \quad \ldots \quad (9.4)$$

which is equivalent to

$$|\boldsymbol{\Sigma} - \lambda\boldsymbol{W^{-2}}| = 0. \quad \ldots \quad (9.5)$$

10. Principal components with restrictions on individual variances

In the previous sections the predictive efficiency of a random variable $\boldsymbol{Y}$ for predicting $\boldsymbol{X}$ is judged by the sum (or weighted sum) of residual variances for the different components of $\boldsymbol{X}$. An optimum choice of $\boldsymbol{Y}$ (in a given class) was made by maximizing the predictive efficiency, i.e., by minimizing the sum of residual variances. But it may so happen that for such an optimum choice of $\boldsymbol{Y}$, the residual variances for some elements of $\boldsymbol{X}$ are above certain desired levels. In such a case it may be necessary to impose some restrictions on the individual variances and maximize the overall predictive efficiency.

Let $\boldsymbol{Y}$ be the vector of q linear functions, $\boldsymbol{L}_1'\boldsymbol{X}, \ldots, \boldsymbol{L}_q'\boldsymbol{X}$, of $\boldsymbol{X}$, where $\boldsymbol{L}_i'\boldsymbol{\Sigma}\boldsymbol{L}_i = 1, \boldsymbol{L}_i'\boldsymbol{\Sigma}\boldsymbol{L}_j = 0, i \neq j$. The residual variance in predicting X_i, the i-th element of $\boldsymbol{X}$, by $\boldsymbol{Y}$ is

$$\sigma_{ii} - (\boldsymbol{L}_1'\boldsymbol{\Sigma}_i)^2 - \ldots - (\boldsymbol{L}_q'\boldsymbol{\Sigma}_i)^2 \qquad \ldots \quad (10.1)$$

where $\boldsymbol{\Sigma}_i$ is the i-th column vector of $\boldsymbol{\Sigma}$. We impose the conditions

$$\begin{aligned} &\sigma_{ii} - (\boldsymbol{L}_1'\boldsymbol{\Sigma}_i)^2 - \ldots - (\boldsymbol{L}_q'\boldsymbol{\Sigma}_i)^2 \leqslant e_i^2 \text{ (given)} \\ \text{or} \quad &(\boldsymbol{L}_1'\boldsymbol{\Sigma}_i)^2 + \ldots + (\boldsymbol{L}'\boldsymbol{\Sigma}_i)^2 \geqslant \sigma_{ii} - e_i^2 = \delta_i^2 \text{ (say)} \qquad \ldots \quad (10.2) \\ &i = 1, \ldots, p. \end{aligned}$$

The sum of squares of residual variances is

$$\Sigma\,\sigma_{ii} - (\boldsymbol{L}_1'\boldsymbol{\Sigma}\boldsymbol{\Sigma}\boldsymbol{L}_1 + \ldots + \boldsymbol{L}_q'\boldsymbol{\Sigma}\boldsymbol{\Sigma}\boldsymbol{L}_q). \qquad \ldots \quad (10.3)$$

The problem is one of maximizing

$$\boldsymbol{L}_1'\boldsymbol{\Sigma}\boldsymbol{\Sigma}\boldsymbol{L}_1 + \ldots + \boldsymbol{L}_q'\boldsymbol{\Sigma}\boldsymbol{\Sigma}\boldsymbol{L}_q \qquad \ldots \quad (10.4)$$

subject to the conditions

$$\left.\begin{aligned} &\boldsymbol{L}_i'\boldsymbol{\Sigma}\boldsymbol{L}_i = 1, \quad i = 1, \ldots, q \\ &\boldsymbol{L}_i'\boldsymbol{\Sigma}\boldsymbol{L}_j = 0, \quad i \neq j, \\ &(\boldsymbol{L}_1'\boldsymbol{\Sigma}_i)^2 + \ldots + (\boldsymbol{L}_q'\boldsymbol{\Sigma}_i)^2 \geqslant \delta_i^2, \ i = 1, \ldots, p. \end{aligned}\right\} \qquad \ldots \quad (10.5)$$

We are thus led to a very complicated problem in non-linear programming.

When $q = 1$, there is only a single linear function $\boldsymbol{L}_1'\boldsymbol{X}$ to be determined. The expression to be maximized is

$$\boldsymbol{L}_1'\boldsymbol{\Sigma}\boldsymbol{\Sigma}\boldsymbol{L}_1 \qquad \ldots \quad (10.6)$$

subject to the conditions

$$\boldsymbol{L}_1'\boldsymbol{\Sigma}\boldsymbol{L}_1 = 1 \text{ and } (\boldsymbol{L}_1'\boldsymbol{\Sigma}_i)^2 \geqslant \delta_i^2, \quad i = 1, \ldots, p. \qquad \ldots \quad (10.7)$$

Making the substitution $\boldsymbol{\Sigma}\boldsymbol{L}_1 = \boldsymbol{U}$ or $\boldsymbol{L}_1 = \boldsymbol{\Sigma}^{-1}\boldsymbol{U}$, we need only maximize

$$\boldsymbol{U}'\boldsymbol{U} \qquad \ldots \quad (10.8)$$

subject to the conditions

$$\boldsymbol{U}'\boldsymbol{\Sigma}^{-1}\boldsymbol{U} = 1 \text{ and } U_i^2 \geqslant \delta_i^2, \quad i = 1, \ldots, p. \qquad \ldots \quad (10.9)$$

The solution is complicated even in this simple case. It has been shown elsewhere (Rao, 1962b, 1964) how to obtain a solution for small values of p. The general problem awaits solution.

11. Principal components of $\boldsymbol{X}$ uncorrelated with the instrumental variable $\boldsymbol{Z}$

In Section 8, the instrumental variable $\boldsymbol{Z}$ is analyzed into principal components on the basis of their predictive efficiency for elements of a variable $\boldsymbol{X}$. In some problems, it is of interest to determine the principal components of $\boldsymbol{X}$ in the class of linear functions of $\boldsymbol{X}$ uncorrelated with instrumental variables. The problem may be stated as one of determining q linear functions $\boldsymbol{L}_1'\boldsymbol{X}, \ldots, \boldsymbol{L}_q'\boldsymbol{X}$ such that

$$V(\boldsymbol{L}_1'\boldsymbol{X}) + \ldots + V(\boldsymbol{L}_q'\boldsymbol{X}) = \boldsymbol{L}_1'\boldsymbol{\Sigma}\boldsymbol{L}_1 + \ldots + \boldsymbol{L}_q'\boldsymbol{\Sigma}\boldsymbol{L}_q \qquad \ldots \quad (11.1)$$

is a maximum subject to the conditions

$$\left.\begin{aligned} &\boldsymbol{L}_i' \boldsymbol{L}_i = 1, \quad \boldsymbol{L}_i' \boldsymbol{L}_j = 0 \\ &\operatorname{cov}(\boldsymbol{L}_i' \boldsymbol{X}, \boldsymbol{Z}) = 0, \ i = 1, \ldots, q. \end{aligned}\right\} \quad \ldots \quad (11.2)$$

As before, let the dispersion matrix of$(\boldsymbol{X}, \boldsymbol{Z})$ be

$$\begin{pmatrix} \boldsymbol{\Sigma} & \boldsymbol{\Theta} \\ \boldsymbol{\Theta}' & \boldsymbol{\Gamma} \end{pmatrix}. \quad \ldots \quad (11.3)$$

In terms of the elements of the matrix (11.3), the problem is one of maximizing

$$\boldsymbol{L}_1' \boldsymbol{\Sigma} \boldsymbol{L}_1 + \ldots + \boldsymbol{L}_q' \boldsymbol{\Sigma} \boldsymbol{L}_q \quad \ldots \quad (11.4)$$

subject to the conditions

$$\left.\begin{aligned} &\boldsymbol{L}_i' \boldsymbol{L}_i = 1, \ \boldsymbol{L}_i' \boldsymbol{L}_j = 0, \quad i \neq j \\ &\boldsymbol{L}_i' \boldsymbol{\Theta} = 0, \\ &\qquad i, j = 1, \ldots, q. \end{aligned}\right\} \quad \ldots \quad (11.5)$$

An application of the result (2.12) shows that the maximum of (11.4) is attained when $\boldsymbol{L}_1, \ldots, \boldsymbol{L}_q$ are the first q right eigen vectors of the matrix

$$(\boldsymbol{I} - \boldsymbol{\Theta} (\boldsymbol{\Theta}'\boldsymbol{\Theta})^{-1} \boldsymbol{\Theta}')\boldsymbol{\Sigma}. \quad \ldots \quad (11.6)$$

The principal components so determined may be useful in the following situation.

Suppose we have a vector variable $\boldsymbol{X}$ representing p 'economic transactions' measured in terms of a common unit (like the cash value). The observed values of $\boldsymbol{X}$ over time constitute a multiple time series. Some years ago, Stone (1947) considered the problem of isolating linear functions of $\boldsymbol{X}$ which have an intrinsic economic significance from those which represent trend with time and those which measure random errors. For this purpose, he computed the dispersion matrix of the p-variables considering the observations at different points of time as repeated values of $\boldsymbol{X}$ and extracted the principal components of $\boldsymbol{X}$ from the estimated dispersion matrix, without any reference to the time factor. The problem was then posed as that of interpreting the dominant principal components which accounted for a high percentage of the total variance. The first principal component with the largest variance was interpreted as representing linear trend and the rest were interpreted in economic terms.

It is not clear why the trend, even if it is linear, should be reflected in the first principal component only. Then there is the difficulty of choosing more than one principal component to explain the trend, if it is non-linear in time.

To apply the present analysis, we consider an instrumental variable $\boldsymbol{Z}$ consisting of orthogonal functions of time, $\phi_1(t)$, $\phi_2(t)$, ..., representing the first, second,... degree polynomial trend with respect to time. Then we obtain the principal compoments of $\boldsymbol{X}$ in the class of linear functions not showing any trend, which may then be interpreted in economic terms.

12. SIZE AND SHAPE FACTORS

Biologists use functions of measurements representing size and shape of organisms in studying differences within and between groups. Size and shape are not, however, well-defined concepts and their measurement will be arbitrary to some extent. Let us examine some functions proposed for this purpose.

Penrose (1947) defines size as the linear function

$$\frac{X_1}{\sigma_1}+\ldots+\frac{X_p}{\sigma_p} \qquad \ldots \quad (12.1)$$

where σ_i is the standard deviation of X_i and shape as any linear function

$$c_1\frac{X_1}{\sigma_1}+\ldots+c_p\frac{X_p}{\sigma_p} \qquad \ldots \quad (12.2)$$

with $c_1+\ldots+c_p=0$. In situations where X_1, X_2, X_3 correspond, in some sense, to length, width and height of an organism, Mosimann (1950) proposed the product

$$X_1X_2X_3 \qquad \ldots \quad (12.3)$$

as a volumetric or a ponderal definition of size. It may be noted that a function of the type

$$X_1^{\beta_1}X_2^{\beta_2}X_3^{\beta_3} \qquad \ldots \quad (12.4)$$

studied by Rao and Shaw (1948) provides a better index of volume for a suitable choice of the exponents $\beta_1, \beta_2, \beta_3$. In terms of the logarithms of the measurements the size function (12.4) has the linear form

$$\beta_1 \log X_1+\beta_2 \log X_2+\beta_3 \log X_3. \qquad \ldots \quad (12.5)$$

It has been suggested by Jolicoeur and Mosimann (1960) that the first principal component, which has maximum variation, may be taken as size factor provided all the coefficients are positive and other principal components with positive and negative coefficients as shape factors. A justification for such an interpretation of principal components is as follows.

Consider the i-th element X_i of $\boldsymbol{X}$ and the j-th principal component $\boldsymbol{P}_j'\boldsymbol{X}$ of $\boldsymbol{X}$. The regression of X_i on $\boldsymbol{P}_j\boldsymbol{X}$ is simply P_{ji}, the i-th element in $\boldsymbol{P}_j$. Now, a unit increase in $\boldsymbol{P}_j'\boldsymbol{X}$ produces on the average an increase of P_{ji} in X_i. If all the coefficients $P_{j1}, \ldots, P_{jp}$ are positive, a unit increase in $\boldsymbol{P}_j'\boldsymbol{X}$ increases the value of each of the measurements, in which case $\boldsymbol{P}_j'\boldsymbol{X}$ may be called a size factor. If some of the coefficients are positive and others negative, then an increase in $\boldsymbol{P}_j'\boldsymbol{X}$, increases the values of some of the measurements and decreases the values of the others, in which case $\boldsymbol{P}_j'\boldsymbol{X}$ may be called a shape factor. Penrose's size and shape functions have similar properties. In either case, there seems to be no suitable interpretation of the relative magnitudes

of changes in the different measurements or of the distribution of positive and negative coefficients in the case of a shape factor. There seem to be too many arbitrary elements involved as the principal components are not explicitly derived to represent size and shape factors in any well-defined manner.

We shall first generalize the problem by attempting to measure size and shape variation with respect to a given set of measurements $X_1, \ldots, X_p$ in terms of instrumental variables $Z_1, \ldots, Z_m$, where the latter may include some or all of the former.

Let the dispersion matrix of $(\boldsymbol{X}, \boldsymbol{Z})$ where $\boldsymbol{X}' = (X_1, \ldots, X_p)$ and $\boldsymbol{Z}' = (Z_1, \ldots, Z_m)$, be

$$\begin{pmatrix} \boldsymbol{\Sigma} & \boldsymbol{\Theta} \\ \boldsymbol{\Theta}' & \boldsymbol{\Gamma} \end{pmatrix} \qquad \ldots \quad (12.6)$$

and consider a linear function of $\boldsymbol{Z}$, with unit variance,

$$\boldsymbol{B}'\boldsymbol{Z} = B_1 Z_1 + \ldots + B_m Z_m, \quad \boldsymbol{B}'\boldsymbol{\Gamma}\,\boldsymbol{B} = 1. \qquad \ldots \quad (12.7)$$

The regression coefficients of $X_1, \ldots, X_p$ on $\boldsymbol{B}'\boldsymbol{Z}$ are the components of the vector $\boldsymbol{\Theta}\,\boldsymbol{B}$. Let us specify the ratios of the regression coefficients, taking into account the signs, as the elements of a vector $\boldsymbol{R}$, so that

$$\boldsymbol{\Theta B} = \rho \boldsymbol{R} \qquad \ldots \quad (12.8)$$

where ρ is a constant. If all the components of $\boldsymbol{R}$ are chosen as positive, then any solution for $\boldsymbol{B}$ of the equation (12.8) provides a size function. If some components of $\boldsymbol{R}$ are positive and others negative, then a solution for $\boldsymbol{B}$ provides a shape function. In the case when $\boldsymbol{\Theta}$ is a square matrix of rank m, there is a unique $\boldsymbol{B}$ and ρ for a given $\boldsymbol{R}$, satisfying the equations

$$\boldsymbol{\Theta B} = \rho \boldsymbol{R}, \quad \boldsymbol{B}'\boldsymbol{\Gamma}\boldsymbol{B} = 1. \qquad \ldots \quad (12.9)$$

Otherwise there may be a multiplicity of solutions. In such a case $\boldsymbol{B}$ may be chosen to maximize ρ.

Introducing Lagrangian multipliers $\boldsymbol{C}_1$ (a p-dimensional vector) and c_2 (a constant), the function to be differentiated is

$$\rho + \boldsymbol{C}_1'(\boldsymbol{\Theta B} - \rho \boldsymbol{R}) + c_2(\boldsymbol{B}'\boldsymbol{\Gamma}\,\boldsymbol{B} - 1). \qquad \ldots \quad (12.10)$$

Differentiating with respect to $\boldsymbol{B}$, $\boldsymbol{C}_1$, ρ and c_2 we obtain the equations

$$\left.\begin{aligned} \boldsymbol{\Theta}'\,\boldsymbol{C}_1 + c_2\boldsymbol{\Gamma}\,\boldsymbol{B} &= 0 \\ \boldsymbol{\Theta B} &= \rho R \\ \boldsymbol{R}'\boldsymbol{C}_1 &= 1 \\ \boldsymbol{B}'\boldsymbol{\Gamma}\,\boldsymbol{B} &= 1. \end{aligned}\right\} \qquad \ldots \quad (12.11)$$

The first two equations can be written as

$$\left.\begin{aligned} \boldsymbol{\Theta}'\boldsymbol{C} + \boldsymbol{\Gamma}\rho^{-1}\boldsymbol{B} &= 0 \\ \boldsymbol{\Theta}\,\rho^{-1}\boldsymbol{B} &= \boldsymbol{R} \end{aligned}\right\} \qquad \ldots \quad (12.12)$$

where $\boldsymbol{C} = (c_2\rho)^{-1}\boldsymbol{C}_1$. Solving the equations (12.12) we have a solution for $\rho^{-1}\boldsymbol{B}$, i.e., the size or shape function is determined apart from a constant multiplier. If necessary, the function can be standardized by reducing its variance to unity. To obtain an explicit representation for $\rho^{-1}\boldsymbol{B}$, let

$$\begin{pmatrix} \boldsymbol{\Theta}' & \boldsymbol{\Gamma} \\ \boldsymbol{0} & \boldsymbol{\Theta} \end{pmatrix}^{-1} = \begin{pmatrix} \boldsymbol{A} & \boldsymbol{E} \\ \boldsymbol{E} & \boldsymbol{D} \end{pmatrix}. \qquad \ldots \quad (12.13)$$

Then $\rho^{-1}\boldsymbol{B} = \boldsymbol{DR}$. The size and shape functions so determined have greater flexibility, although there will still be some arbitrariness in the choice of the vector of ratios, $\boldsymbol{R}$, when it cannot be specified by other considerations.

For instance, anthropologists use the ratio of head breadth to head length, called the cephalic index, to measure the shape of the head. The present approach suggests that the shape function can be built out of a *number* of length and breadth measurements on the head, by choosing $\boldsymbol{R}$ such that the ratios for length measurements have a positive sign and those for breadth measurements have a negative sign. The shape function so determined has the property that an increase in its value increases the lengths and decreases the breadths. Further, other measurements on the body such as stature, chest girth etc., can be brought in to obtain a better measure of head shape. Examples of such functions are given in Rao (1961, 1962a). It has also been found that the choice of absolute values of the elements of $\boldsymbol{R}$ in proportion to the standard deviations $\sigma_1, \ldots, \sigma_p$ of the measurements $X_1, \ldots, X_p$ leads to reasonable results.

The size and shape functions determined by the above method may not be uncorrelated as in the case of the principal components, although this can be achieved by choosing the vectors $\boldsymbol{R}$ for size and shape functions suitably. But lack of correlation is not an important property if size and shape variations are examined individually. However, if size and shape functions are used jointly in any study, their correlation should be taken into account in the statistical analysis. For instance, if a chart representing the configuration of a set of groups with respect to mean size and shape is desired, one could use 'size' and 'shape corrected for size' (by regression) and represent them on orthogonal axes.

13. PRINCIPAL COMPONENT AND FACTOR ANALYSES

Some authors do not make a distinction between these two analyses and this has, no doubt, caused some amount of confusion. It is, therefore, of interest to examine the differences in the nature of information provided by these two techniques. It is shown that they provide distinct answers to two different enquiries concerning a hypothetical structure of the elements of a vector random variable.

Let us suppose that there exist hypothetical uncorrelated factors (variables), $F_1, F_2, \ldots$, presumably infinite in number, such that

$$\left.\begin{aligned} X_1 &= a_{11}F_1 + a_{12}F_2 + \cdots \\ &\cdots \\ X_p &= a_{p1}F_1 + a_{p2}F_2 + \cdots \end{aligned}\right\} \qquad \ldots \quad (13.1)$$

or in matrix notation

$$\boldsymbol{X} = \boldsymbol{A}\,\boldsymbol{F}. \qquad \ldots \quad (13.2)$$

It may be seen that the factor structure with common and specific factors, assumed in psychological work is a special (and perhaps an unrealistic !) case of (13.1). The representation (13.1) is not unique for by an orthogonal transformation on the hypothetical variable $\boldsymbol{F}$, the structural equation (13.2) can be written as

$$\boldsymbol{X} = \boldsymbol{B}\boldsymbol{G} \qquad \ldots \quad (13.3)$$

where $\boldsymbol{G}$ represents a new (transformed) factor variable.

The problem we wish to investigate is one of *approximating* $X_1, \ldots, X_p$ by linear functions of as few factors, which may be called dominant, as possible. An approximation to $\boldsymbol{X}$ in terms of q factors, to be denoted by $\boldsymbol{X}^{(q)}$ is obtained by considering a representation of $\boldsymbol{X}$, such as (13.1) and truncating the right hand side at the q-th term. In matrix notation

$$\boldsymbol{X}^{(q)} = \boldsymbol{A}_q\,\boldsymbol{F}^{(q)} \qquad \ldots \quad (13.4)$$

where $\boldsymbol{A}_q$ is the matrix of the first q columns of $\boldsymbol{A}$ and $\boldsymbol{F}^{(q)}$ is the vector of the variables $F_1, \ldots, F_q$. The different choices of $\boldsymbol{A}_q$ are obtained by making orthogonal transformations on $\boldsymbol{F}$ (the entire set of hypothetical factors), leading to different representations of $\boldsymbol{X}$ and applying the truncation procedure.

We have not yet specified the nature of the approximation we are seeking. It is seen that the dispersion matrix of $\boldsymbol{X}^{(q)}$ is $\boldsymbol{A}_q\boldsymbol{A}_q'$ (assuming without loss of generality that all F_i have unit variance), while the dispersion matrix of $\boldsymbol{X}$ is $\boldsymbol{\Sigma}$. Let us suppose that the approximation aims at the closeness of $\boldsymbol{\Sigma}$ and $\boldsymbol{A}_q\boldsymbol{A}_q'$. In such a case, we may determine $\boldsymbol{A}_q$ by minimizing the Euclidean norm of the residual dispersion matrix $\|\boldsymbol{\Sigma} - \boldsymbol{A}_q\boldsymbol{A}_q'\|$. Applying the result (2.6), the answer is

$$\boldsymbol{A}_q' = (\sqrt{\lambda_1}\,\boldsymbol{P}_1\ \sqrt{\lambda_2}\,\boldsymbol{P}_2 \ldots \sqrt{\lambda_q}\,\boldsymbol{P}_q) \qquad \ldots \quad (13.5)$$

i.e., the i-th column of $\boldsymbol{A}_q$ is $\sqrt{\lambda_i}\,\boldsymbol{P}_i$ where λ_i is the i-th eigen value and $\boldsymbol{P}_i$ is the i-th eigen vector of $\boldsymbol{\Sigma}$. Then it may be shown that the best estimates of the factors are

$$F_i^q = \lambda_i^{-1/2}\boldsymbol{P}_i'\boldsymbol{X}, \quad i = 1, \ldots, q \qquad \ldots \quad (13.6)$$

where $\boldsymbol{P}_i'\boldsymbol{X}$ is the i-th principal component of $\boldsymbol{X}$. We are thus led to principal component analysis as a definite answer to the problem posed.

Using estimated F_i^q, the q-th approximation to $\boldsymbol{X}$ as in (13.6), the residual is

$$\boldsymbol{X}^q - \boldsymbol{X} = (\boldsymbol{P}_{q+1}\boldsymbol{P}_{q+1}' + \ldots + \boldsymbol{P}_p\boldsymbol{P}_p')\boldsymbol{X}. \qquad \ldots \quad (13.7)$$

The dispersion matrix of the residual is

$$\begin{aligned}(\boldsymbol{P}_{q+1}\boldsymbol{P}'_{+1} + \ldots + \boldsymbol{P}_p\boldsymbol{P}_p')\boldsymbol{\Sigma}(\boldsymbol{P}_{q+1}\boldsymbol{P}_{q+1}' + \ldots + \boldsymbol{P}_p\boldsymbol{P}_p)'.\\ = \lambda_{q+1}\boldsymbol{P}_{q+1}\boldsymbol{P}_{q+1}' + \ldots + \lambda_p\boldsymbol{P}_p\boldsymbol{P}_p' \qquad \ldots \quad (13.8)\end{aligned}$$

which is small if $\lambda_{q+1}, \ldots, \lambda_p$ are small. In any problem we can examine the actual residuals (13.7) in judging the adequacy of a given order of approximation.

We have seen how the criterion of minimizing $\|\boldsymbol{\Sigma}-\boldsymbol{A}_q\boldsymbol{A}_q'\|$ led to the principal component analysis. We shall show that a different criterion leads to factor analysis.

The amount of correlation between X_i, X_j explained by the q factors $F_1, \ldots, F_q$ is

$$\begin{aligned}&\operatorname{cov}(X_i^q, X_j^q) \div \sqrt{V(X_i)V(X_j)}\\ &= (a_{i1}a_{j1}+\ldots+a_{iq}a_{jq})/\sigma_i\sigma_j\\ &= b_{i1}b_{j1}+\ldots+b_{iq}b_{jq}\end{aligned}$$

where σ_i is the standard deviation of X_i and $b_{ik} = a_{ik}/\sigma_i$. The off diagonal elements of $\boldsymbol{B}_q\boldsymbol{B}_q'$ where $\boldsymbol{B}_q$ is the matrix with its i-th row as $(b_{i1}, \ldots, b_{iq})$ give all possible correlations explained by the q factors. Let $\boldsymbol{\Lambda}$ be the actual correlation matrix of the variables $X_1, \ldots, X_p$. We wish to choose $\boldsymbol{B}_q$ such that the off diagonal elements of $\boldsymbol{\Lambda}-\boldsymbol{B}_q\boldsymbol{B}_q'$ are as small as possible. A single measure of difference is the sum of squares of the off diagonal elements of $\boldsymbol{\Lambda}-\boldsymbol{B}_q\boldsymbol{B}_q'$ and the mathematical problem is one of minimizing this measure subject to the condition that the diagonal entries of $\boldsymbol{\Lambda}-\boldsymbol{B}_q\boldsymbol{B}_q'$ are non-negative. The factor structure provided by the optimum $\boldsymbol{B}_q$ is known as *factor analysis*; the attempt in such a case is to explain the correlations among the measurements rather than the variances of the individual measurements. The problem of an optimum choice of $\boldsymbol{B}_q$ is not easy to solve, and no definite solution can be obtained. An iterative method is suggested by Rao (1955) but the convergence of the procedure has not been adequately studied. For further literature see Whittle (1953), and the references in Maxwell and Lawley (1963).

14. An application to a problem in multidimensional scaling

The data consist of $n(n-1)/2$ independent determinations (possibly subject to errors) of distances between n stimuli and the problem is to determine the smallest Euclidean space in which the stimuli can be represented as points (Torgerson, 1958). Let $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ be points representing the n stimuli in a Euclidean space, and $\bar{\boldsymbol{X}}$ be the centre of gravity of the points. Then

$$\begin{aligned}(\boldsymbol{X}_i-\bar{\boldsymbol{X}})'(\boldsymbol{X}_j-\bar{\boldsymbol{X}}) &= n^{-2}[(\boldsymbol{X}_i-\boldsymbol{X}_1)+\ldots+(\boldsymbol{X}_i-\boldsymbol{X}_n)]'[(\boldsymbol{X}_j-\boldsymbol{X}_1)+\ldots+(\boldsymbol{X}_j-\boldsymbol{X}_n)]\\ &= n^{-2}\sum_k\sum_m [(\boldsymbol{X}_j-\boldsymbol{X}_k)'(\boldsymbol{X}_j-\boldsymbol{X}_k)+(\boldsymbol{X}_i-\boldsymbol{X}_m)'(\boldsymbol{X}_i-\boldsymbol{X}_m)\\ &\qquad -(\boldsymbol{X}_i-\boldsymbol{X}_j)'(\boldsymbol{X}_i-\boldsymbol{X}_j)-(\boldsymbol{X}_k-\boldsymbol{X}_m)'(\boldsymbol{X}_k-\boldsymbol{X}_m)]\\ &= \frac{1}{2n^2}\sum_k\sum_m (\delta_{jk}+\delta_{im}-\delta_{ij}-\delta_{mk}) \qquad \ldots\ (14.1)\end{aligned}$$

where δ_{rs} denotes the square of the true distance between the points $\boldsymbol{X}_r$ and $\boldsymbol{X}_s$. Using the notation $\sum_m \delta_{im} = \delta_{i\cdot}$ and $\sum_k\sum_m \delta_{km} = \delta_{\cdot\cdot}$ the expression (14.1) can be written

$$\frac{1}{2}\left(-\delta_{ij}+\frac{\delta_{i\cdot}+\delta_{\cdot j}}{n}-\frac{\delta_{\cdot\cdot}}{n^2}\right). \qquad \ldots\ (14.2)$$

Thus, given the mutual distances we can express the configuration of the points by the matrix $\mathcal{X}_d'\mathcal{X}_d$ whose (i, j) entry is given by (14.2), where $\mathcal{X}_d$ is as defined in (3.3). The matrix $\mathcal{X}_d'\mathcal{X}_d$, as introduced in Section 4, represents the lengths of the lines joining the points to the centre of gravity and the angles between them. The best representation of the points in a Euclidean space of q dimensions is provided by the $q \times m$ matrix whose i-th row is $\sqrt{\lambda_i}\, Q_i$ where λ_i and Q_i are the i-th eigen value and vector of $\mathcal{X}_d'\mathcal{X}_d$.

However, such an analysis is not directly applicable when the exact distances δ_{ij} are not known, but only estimates d_{ij} of δ_{ij} are available. Substituting d_{ij} for δ_{ij} in (14.2) we have an estimate of $\mathcal{X}_d'\mathcal{X}_d$ which we may denote by $B = (b_{ij})$. The matrix $\boldsymbol{B}$ need not be non-negative definite. We may still compute the eigen values and vectors of $\boldsymbol{B}$. If $\hat{\lambda}_i$ and $\hat{Q}_i$ denote the i-th eigen value and vector of $\boldsymbol{B}$, an estimate of the best q dimensional representation is given by the $q \times m$ matrix whose i-th row is $\sqrt{\hat{\lambda}_i}\, \hat{Q}_i$, provided $\hat{\lambda}_1, \ldots, \hat{\lambda}_q$ are positive.

One drawback of the solution derived from $\boldsymbol{B}$ is that the estimated points may not have the origin as the centre of gravity. This may be secured by determining the eigen vectors of $\boldsymbol{B}$ with the restriction that the sum of the elements of each eigen vector is zero. Let $\boldsymbol{U}$ represent a column vector of n unities. Then it can be shown, using result (iii) of Section 2, that the eigen vectors of the matrix

$$\left(\boldsymbol{I} - \frac{\boldsymbol{U}\boldsymbol{U}'}{n}\right)\boldsymbol{B} \qquad \ldots \quad (14.3)$$

satisfy the required conditions. Let $\mu_1, \mu_2, \ldots$ be the eigen values and $\boldsymbol{B}_1, \boldsymbol{B}_2, \ldots$ the corresponding eigen vectors of the matrix (14.3). Then the rows of the estimated matrix of points (columns representing the points) in a q dimensional space are $\sqrt{\mu_1}\, \boldsymbol{B}_1', \ldots, \sqrt{\mu_q}\boldsymbol{B}_q'$.

15. Estimation of structural relationship, heterogeneity of data, etc.

A recent application by Wernimont (1963) demonstrates the use of principal component analysis in detecting heterogeneity of data. In his study there were 9 spectrophotometers and each instrument was used to obtain a series of six absorbance curves (as a discrete set of measurements at 20 different wave lengths) of solutions at three different concentrations (30, 60 and 90) and on two different days. The data consisted of 54 sets of 20 measurements and in the notation of the present paper $n = 54$ and $p = 20$. According to Beer's law, if the same wave lengths have been used for all the instruments, the 54 points in the 20 dimensional space should ideally be on a line through the origin when there are no errors in the absorbance readings. The direction cosines of such a line would provide the constants of Beer's law for a given set of wave lengths. If, indeed, the wave length settings differed from one instrument to another the configuration of points would not be confined to a straight line. The object of investigation was to examine whether the instruments had systematic errors in the wave length setting. If some of the instruments did not allow a

proper setting of the desired wave lengths, the data would be heterogeneous ; this could be detected by asking the question whether the observed points could be regarded as clustering round a straight line.

This is exactly what the principal components can detect provided the errors of absorbance measurements at different wave lengths are independent and have nearly the same distribution.

According to Beer's law, the vector $\boldsymbol{X}$ of measurements at 20 different wave lengths can be written as

$$\boldsymbol{X} = c\boldsymbol{\beta}+\boldsymbol{\epsilon} \qquad \ldots \quad (15.1)$$

where $\boldsymbol{\beta}$ depends on the wave lengths, c is the concentration of the solution used and $\boldsymbol{\epsilon}$ represents the errors of absorbance readings. Thus for different values of c, the true locus of the points $\boldsymbol{X}$ is a straight line through the origin, with direction cosines proportional to $\boldsymbol{\beta}$. But if $(\boldsymbol{\beta}+\boldsymbol{\xi}_{ij})$ is the vector appropriate to the i-th spectrophotometer on the j-th day, where ξ_{ij} represents the systematic error, then for the measurements with the i-th spectrophotometer on the j-th day at concentration c

$$\boldsymbol{X} = (\boldsymbol{\beta}+\boldsymbol{\xi}_{ij})c+\boldsymbol{\epsilon} \qquad \ldots \quad (15.2)$$

so that Beer's law holds for a given instrument on a given day. In such a case we should expect the 54 points to cluster round a maximum of 18 different lines corresponding to the nine instruments and two days. The 18 lines may lie in a space of as many as 18 dimensions depending on the complexity of the systematic errors. Let us examine this problem by the principal component analysis by pooling the data from all the instruments. The 54 vectors provide a 20×20 dispersion matrix. (Actually in a problem with the model such as (15.2), the principal component analysis could be carried out on the uncorrected dispersion matrix. The computations referred to are on the corrected dispersion matrix as reported by Wernimont.) The first two eigen values of this matrix have been found to be (0.119590×10^7) and (0.356424×10^3) which explain 99.94 and 0.03 percent of total variation respectively. The rest of the eigen values are much smaller indicating that the configuration of the points can be examined by one or two dimensional representations.

To find the coordinates of the projected points on the best two dimensional plane we have to determine the first two eigen vectors $\boldsymbol{P}_1$ and $\boldsymbol{P}_2$ and compute the coordinates

$$\boldsymbol{P}_1'\boldsymbol{X},\ \boldsymbol{P}_2'\boldsymbol{X} \qquad \ldots \quad (15.3)$$

corresponding to each original point $\boldsymbol{X}$. Transferring the origin to $\boldsymbol{P}_1'\bar{\boldsymbol{X}}$, $\boldsymbol{P}_2'\bar{\boldsymbol{X}}$, the coordinates are

$$\boldsymbol{P}_1'(\boldsymbol{X}-\bar{\boldsymbol{X}}),\ \boldsymbol{P}_2'(\boldsymbol{X}-\bar{\boldsymbol{X}}). \qquad \ldots \quad (15.4)$$

For each point, it is also necessary to compute the length of the perpendicular on the best fitting plane to examine whether the representation is adequate with respect to that particular point. The square of the perpendicular from the point $\boldsymbol{X}$ is

$$R^2 = (\boldsymbol{X}-\bar{\boldsymbol{X}})'(\boldsymbol{X}-\bar{\boldsymbol{X}})-[\boldsymbol{P}_1'(\boldsymbol{X}-\bar{\boldsymbol{X}})]^2-[\boldsymbol{P}_2'(\boldsymbol{X}-\bar{\boldsymbol{X}})]^2. \qquad \ldots \quad (15.5)$$

The following table gives the values of (15.4) scaled by the reciprocal of the square root of the eigen values (which is not necessary for our purpose, but adopted from the computations provided by Wernimont) and R^2 for two sets of six absorbance curves of S. P. Meters 1 and 7. Similar values are available for all the 54 absorbance curves.

no.	S.P.M.	concentration	day	$P_1'(X-\bar{X})/\sqrt{\lambda_1}$	$P_2'(X-\bar{X})/\sqrt{\lambda_2}$	R^2
1.1	1	30	1	−1.220	0.614	54.17
1.2		60	1	−0.016	1.419	254.51
1.3		90	1	1.173	1.875	443.93
1.4		30	2	−1.200	0.574	122.15
1.5		60	2	0.001	1.518	372.83
1.6		90	2	1.200	1.877	334.46
7.1	7	30	1	−1.235	0.590	715.20
7.2		60	1	−0.065	0.140	1166.08
7.3		90	1	1.148	−0.266	714.84
7.4		30	2	−1.210	0.649	405.33
7.5		60	2	−0.061	0.261	610.58
7.6		90	2	1.134	0.074	655.72

Using the coordinates $(P_1'(X-\bar{X})/\sqrt{\lambda_1},\ P_2'(X-\bar{X}/\sqrt{\lambda_2})$, the 54 points are plotted on two charts (to avoid over crowding). In the charts (A.1, A.2, A.3) and (A.4, A.5, A.6) represent the absorbance curves at three levels of concentration obtained with S_A (the spectrphotometer A) on the first and second days respectively. The following conclusions emerge. (a) Generally, the points (A.1, A.2, A.3) are close to a line and so also the points (A.4, A.5, A.6) indicating the validity of Beer's law. (b) The points (A.1, A.2, A.3) and (A.4, A.5, A.6) lie on the same line except for S_6 and S_9, indicating errors specific to the instruments. (c) The lines for different instruments are differently oriented indicating systematic errors.

16. Determination of clusters of points

Let $X_1, \ldots, X_n$ be points in a p-dimensional space. In some problems it is of interest to examine whether the n points can be classified into groups or clusters such that the points within a cluster are close together, while the clusters themselves are far apart. Thus if the p coordinates of each point represent the production of p agricultural commodities in a region and different points represent different regions we may like to group together regions which are similar with respect to over all agricultural production (Kendall, 1939). Or if the p coordinates of each point represent the mean anthropological characters of a population and different points represent different populations we may like to examine whether the populations can be grouped into distinct clusters on the basis of similarity of the characters (Mahalanobis, 1936;

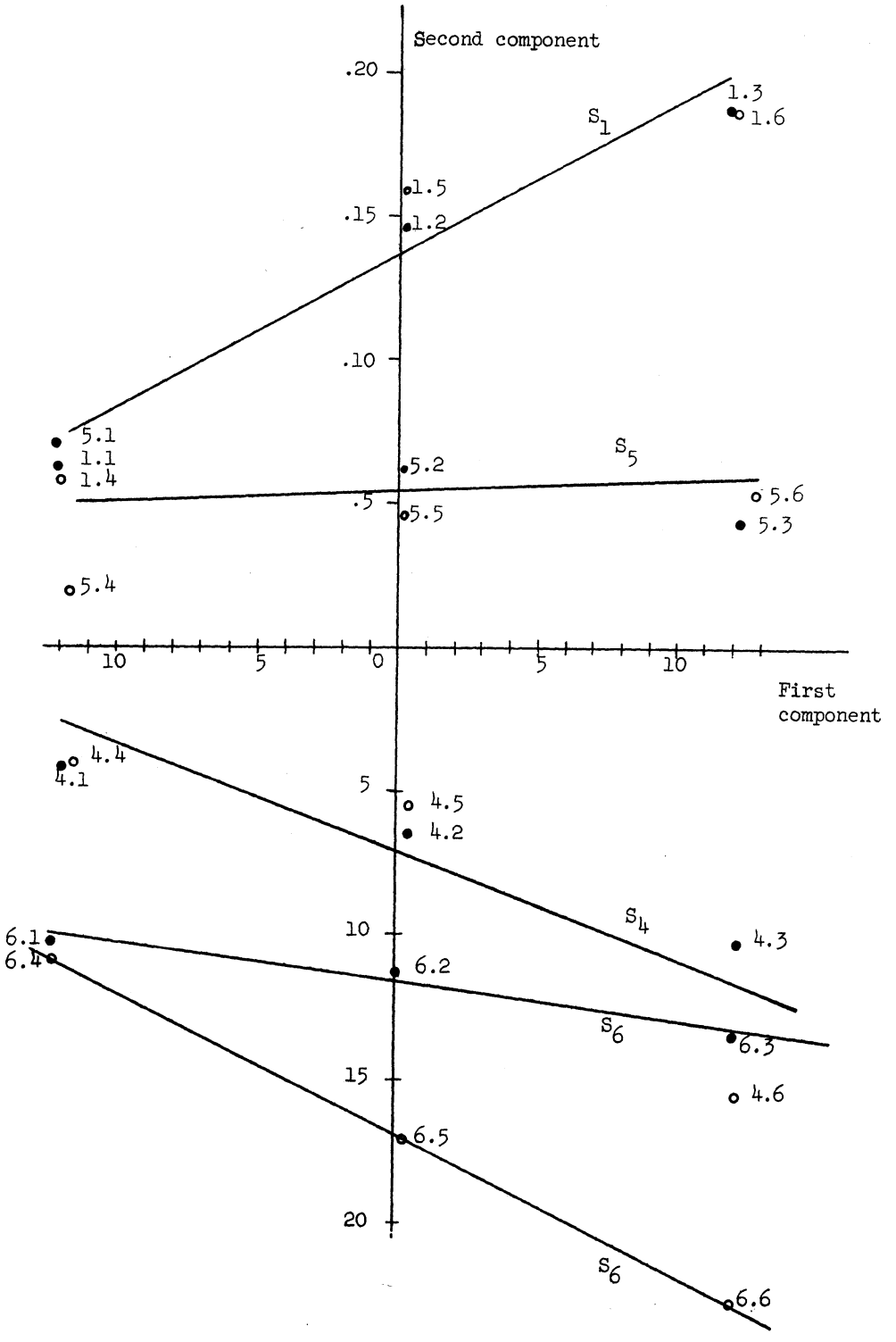

Chart 1. Configuration of the points with respect to the first two Principal Components. The points A.1, A.2, A.3 represent the measurements at three concentrations with spectrophotometer S_A and the points A.4, A.5, A.6 the corresponding measurements on a subsequent day.

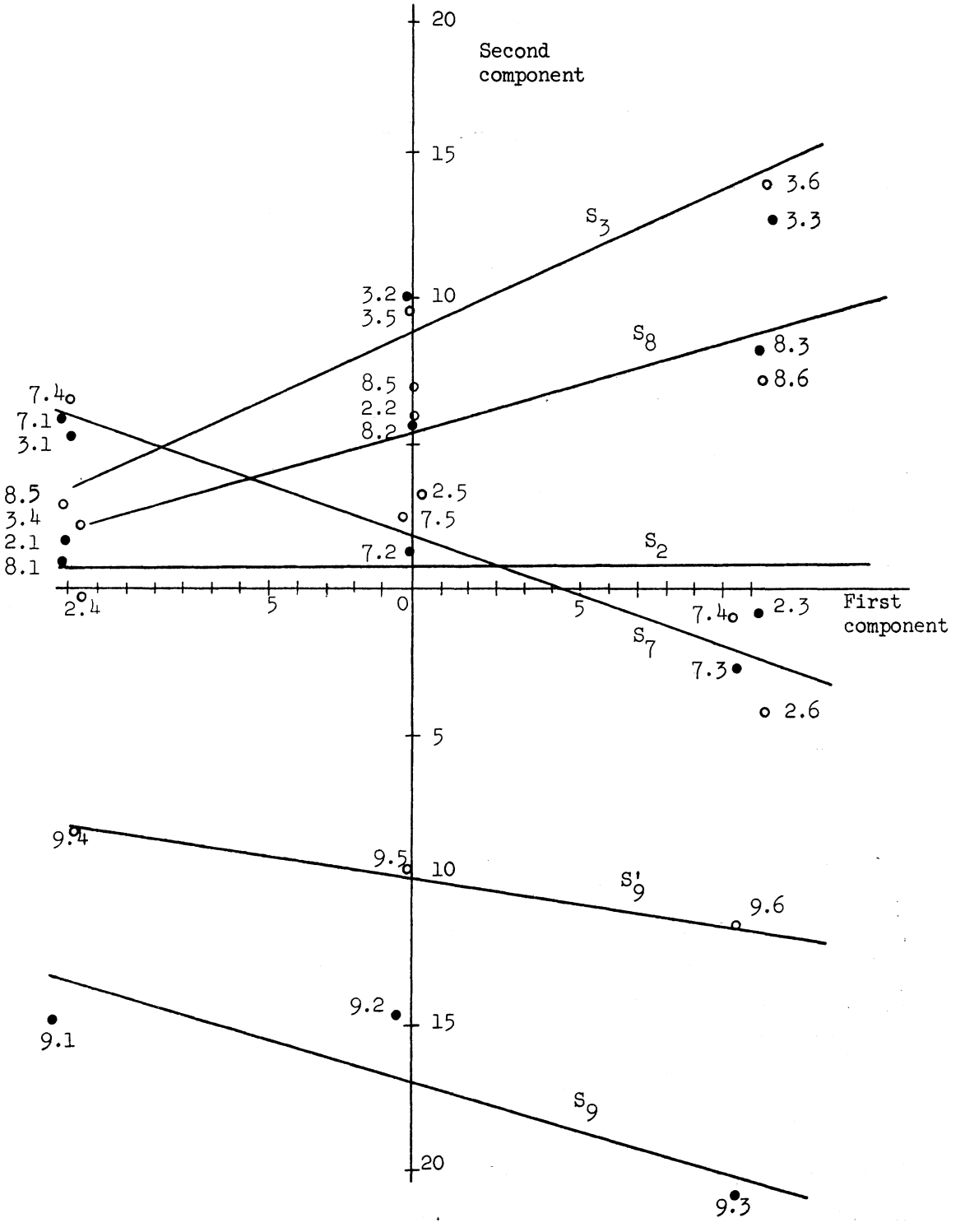

Chart 2. Configuration of the points with respect to the first two Principal Components etc.

Mahalanobis, Majumdar and Rao, 1949 etc.). In all such situations we need to have a measure of distance between two points. One may choose as the distance between $\boldsymbol{X}_i$ and $\boldsymbol{X}_j$

$$d_{ij} = (\boldsymbol{X}_i - \boldsymbol{X}_j)'(\boldsymbol{X}_i - \boldsymbol{X}_j) \quad \ldots \quad (16.1)$$

representing the points in a Euclidean space with orthogonal axes or

$$d_{ij} = (\boldsymbol{X}_i - \boldsymbol{X}_j)\boldsymbol{\Lambda}^{-1}(\boldsymbol{X}_i - \boldsymbol{X}_j) \quad \ldots \quad (16.2)$$

representing the points in a Euclidean space with oblique axes. When $\boldsymbol{\Lambda}$ is the covariance matrix of the characters within a population (or region) the distance (16.2) is known as Mahalanobis distance.

Whatever distance function is chosen, we have a $n \times n$ matrix $D = (d_{ij})$ of all possible distances. We, then, look for clusters of points, by fixing a few pairs of points which are close together, but the pairs themselves being more distant from each other, and building clusters around them by adjoining neighbouring points. In practice such a programme can be successfully carried out although considerable difficulties are involved (see Mahalanobis, Majumdar and Rao, 1949; Majumdar and Rao, 1958).

When $p = 1$, we have one dimensional ordering of points and no visual representation of points on a line is necessary to determine the clusters. When $p = 2$, the determination of clusters is also simple, since we can represent the points on a two dimensional chart (using orthogonal or oblique axes) and mark out the subsets of points close together. For $p = 3$, we may use a three dimensional model. But no such visual aid is available for larger values of p.

We may, then raise the question as to whether the points in a p-dimensional space can be reasonably represented in 2 or 3 dimensional spaces without distorting the configuration as specified by the mutual distances. We have seen in Sections 3 and 4, that such a representation is secured through the principal components. Let $\boldsymbol{P}_1, \boldsymbol{P}_2, \boldsymbol{P}_3$ be the first three eigen vectors of $\boldsymbol{\Sigma}$ (or $\boldsymbol{\Sigma}$ with respect to $\boldsymbol{\Lambda}$ in the oblique case) giving the coordinates $\boldsymbol{P}_1'\boldsymbol{X}_i$, $\boldsymbol{P}_2'\boldsymbol{X}_i$, $\boldsymbol{P}_3'\boldsymbol{X}_i$ for the representation of $\boldsymbol{X}_i$ in the best three dimensional space. We may choose, the first, or the first two or all the three coordinates to obtain a visual representation of the points and then determine the clusters of points.

The success of the method obviously depends on how well the configuration of the points is preserved in the reduced space. If $\lambda_1, \ldots, \lambda_p$ are all the eigen values an overall measure of the adequacy of the best q dimensional representation as introduced in Section 3 is $(\lambda_1 + \ldots + \lambda_q)/(\lambda_1 + \ldots + \lambda_p)$. We may examine the values of this measure for $q = 1, 2, 3$ and decide on a suitable value of q. But such a decision may be misleading in individual cases. So it is necessary to obtain a tabulation of

all the distances in the original and reduced space for a preliminary comparison. The distance between the i-th and j-th points in the reduced space of q dimensions is computed by the formula

$$d_{ij}^q = (\boldsymbol{P}_1'\boldsymbol{X}_i - \boldsymbol{P}_1'\boldsymbol{X}_j)^2 + \ldots + (\boldsymbol{P}_q'\boldsymbol{X}_i - \boldsymbol{P}_q'\boldsymbol{X}_j)^2 \quad \ldots \quad (16.3)$$

whether the original points are in an orthogonal space or not. The difference $d_{ij} - d_{ij}^q$ is always positive and measures the efficiency of reduction in the dimensions of the space with respect to the points i and j. If the differences $(d_{ij} - d_{ij}^q)$ are uniformly small, i.e., for all i and j, we are on safe grounds in determining the clusters on the basis of the first q principal components. If not we can still use the q principal components for the points where the distances are not distorted. The affinities of the points whose distances with the others are distorted may then be determined by considering their actual distances between themselves and with the others.

17. Tests of significance for principal components

In the previous sections no reference has been made to tests of significance in principal component analysis. The reason is that the problems have not been posed in terms of testing of well-defined hypotheses but in terms of estimating some features of the measurements on the individuals of a population. It is assumed that in any realistic problem, reduction in the number of measurements (by omission of some or by replacing the original measurements by a smaller number of linear functions) entails some loss of information. If so we can only consider null hypotheses which specify the amount of information lost. For instance, a null hypothesis may specify that the ratio of the sum of the last $p-q$ eigen values to the sum of all eigen values of a hypothetical dispersion matrix Σ is π. How can we test such a hypothesis when we have an estimate $\hat{\Sigma}$ of Σ. The relevant statistic appears to be

$$(\hat{\lambda}_{q+1} + \ldots + \hat{\lambda}_p)/(\hat{\lambda}_1 + \ldots + \hat{\lambda}_p) \quad \ldots \quad (17.1)$$

where $\hat{\lambda}_1, \ldots, \hat{\lambda}_p$ are the eigen values of $\hat{\Sigma}$. It is extremely difficult to derive the distribution of (17.1) even under the assumption that $\hat{\Sigma}$ has a Wishart distribution.

Another null hypothesis of interest is the equality of the last $p-q$ true eigen values. An appropriate statistic to test such a hypothesis is the ratio of the geometric mean to the arithmetic mean of the estimated roots $\hat{\lambda}_{q+1}, \ldots, \hat{\lambda}_p$ (see Bartlett, 1950, 1951a, 1951b; Rao, 1955). The large sample distribution of the logarithm of the ratio suitably standardized is chi-square when $\hat{\Sigma}$ has a Wishart distribution. However, the tests based on the roots $\hat{\lambda}_1, \ldots, \hat{\lambda}_p$ are likely to be sensitive to departure from Wishartness of the distribution of $\hat{\Sigma}$ and no appropriate robust test criteria are known.

18. USE OF PRINCIPAL COMPONENTS IN TESTING DIFFERENCES IN MEAN VALUES BETWEEN GROUPS

It is well known that when multiple measurments are available on samples of individuals from different populations, the total sum of products (S.P.) matrix can be analyzed as due to 'within' and 'between' groups. The analysis of dispersion (which is a generalization of analysis of variance) is indicated as follows (Rao, 1952).

	D.F.	S.P. matrix
Between	$k-1$	B
Within	$\frac{n}{n+k-1}$	$\frac{W}{T}$

Differences between the mean values of k populations and the configuration of the mean values when differences exist are examined by the roots of the determinantal equation

$$|B-\theta T| = 0. \qquad \ldots \quad (18.1)$$

For applications the reader is referred to Bartlett (1948); Fisher (1939); Rao (1948, 1952); Williams (1959) etc.

Recently, Roy (1958) proposed stepdown procedures wherby differences are examined in one observed variable first, and then in another observed variable eliminating the regression due to the first, and so on using each time a univariate procedure. There is, however, the problem of fixing the order in which the variables are considered. It is suggested that instead of the original variables, the principal components

$$\boldsymbol{P}_1'\boldsymbol{X}, \ldots, \boldsymbol{P}_q'\boldsymbol{X} \qquad \ldots \quad (18.2)$$

where $\boldsymbol{P}_1, \ldots, \boldsymbol{P}_q$ are the first q eigen vectors of the total S.P. matrix T may be considered in the given order. The performance of such a procedure (Dempster, 1963) is not fully studied except in a very pecial case arising in the comparison of growth curves (Rao, 1958). Although the method suggested is in line with the general philosophy of the principal component analysis it is clear that the first few principal components (18.2) are not designed to summarize as much information as possible on the differences between the mean values of the groups. In actual practice it may turn out that some of the principal components with smaller eigen values are better discriminators between the populations with respect to the differences in means than the first few principal components. This is not so only in special cases such as those considered by Rao (1958). However, the general drawback of the step-down procedures is that they do not enable us to study the configuration of the mean values of the different populations, which in practical problems is more important than merely establishing differences in the mean values.

References

Bartlett, M. S. (1947) : Multivariate analysis. *J. Roy. Stat. Soc.* (Supple.), **2**, 176–197.

——— (1950) : Tests of significance in factor analysis. *Brit. J. Psych.* (Stat. Sec.), **3**, 77–85.

——— (1951a) : The effect of standardization of a χ^2-approximation in factor analysis. *Biometrika*, **38**, 337–344.

——— (1951b) : A further note on tests of significance in factor analysis. *Brit. J. Psych*, (Stat. Sec.), **4**, 1–2.

Dempster, A. P. (1963) : Multivariate theory for general stepwise methods. *Ann. Math. Stat.*, **34**, 873–883.

Fisher, R. A. (1939) : The sampling distribution of some statistics obtained from non-linear equations. *Ann. Eugen.*, **9**, 238–249.

Frisch, R. (1929) : Correlation and scatter in statistical variables. *Nordic Stat. J.*, **8**, 36–102.

Hotelling, H. (1933) : Analysis of a complex statistical variable into principal components. *J. Educ. Psych.*, **26**, 417–441, 498–520.

——— (1935) : The most predictable criterion. *J. Educ. Psych.*, **26**, 139–142.

——— (1936a) : Simplified calculation of principal components. *Psychometrika*, **1**, 27–35.

——— (1936b) : Relations between two sets of variates. *Biometrika*, **28**, 321–377.

Jolicoeur, P. and Mosimann, J. E. (1960) : Size and shape variation in the painted turtle, a principal component analysis. *Growth*, **24**, 339–354.

Kendall, M. G. (1939) : The geographical distribution of crop productivity in England. *J. Roy. Stat. Soc.*, **102**, 21–

Mahalanobis, P. C. (1936) : On the generalized distance in statistics. *Proc. Nat. Inst. Sci. India*, **12**, 49–55.

Mahalanobis, P. C., Majumdar, D. N. and Rao, C. R. (1949) : Anthropometric survey of the United Provinces, 1941. A statistical survey. *Sankhyā*, **9**, 90–327.

Majumdar, D. N. and Rao, C. R. (1958) : Bengal anthropometric survey, 1945. A statistical study. *Sankhyā*, **19**, 201–408.

Mosimann, J. E. (1958) : An analysis of allometry in Chelonian shell. *Rev. Can. Biol.*, **17**, 137–228.

Pearson, K. (1901) : On lines and planes of closest fit to systems of points in space. *Phil. Mag.*, **2** (sixth series), 559–572.

Penrose, L. S. (1947) : Some notes on discrimination. *Ann. Eugen.*, **13**, 228–237.

Rao, C. R. and Shaw, D. C. (1948) : On a formula for the prediction of Cranial capacity. *Biometrika*, **4**, 247–253.

Rao, C. R. (1948) : Tests of significance in multivariate analysis. *Biometrika*, **35**, 58–79.

——— (1952) : *Advanced Statistical Methods in Biometric Research.* John Wiley and Sons.

——— (1955) : Estimation and tests of significance in factor analysis. *Psychometrika*, **20**, 93–111.

——— (1958) : Some statistical methods for comparison of growth curves. *Biometrics*, **14**, 1–17.

——— (1960) : Multivariate analysis : An indispensable aid in applied research. *Sankhyā*, **22**, 317–338.

——— (1961) : Some observations on multivariate statistical methods in anthropological research. *Bull. Int. Stat. Inst.*, **38**, 99–109.

——— (1962a) : Use of discriminant and applied function in multivariate analysis. *Sankhyā*, **22**, 317–338.

——— (1962b) : Problems of selection with restrictions. *J. Roy. Stat. Soc.*, (series B), **24**, 401–405.

——— (1964) : *Problems of Selection Involving Programming Techniques*, IBM Symposium on Statistics.

Roy, J. (1958) : Step-down procedure in multivariate analysis. *Ann. Math. Stat.*, **29**, 1177–1187.

SIMONDS, J. L. (1963) : Applications of characteristic vector analysis to photographic and optical response data. *J. Optical Society of America*, **53**, 968–974.

STONE, R. (1947) : An interdependence of blocks of transactions. *J. Roy. Stat. Soc.* (Supple.), **9**, 1–3.

TORGERSON, W. S. (1958) : *Theory and Methods of Scaling*, John Wiley and Sons.

WERNIMONT, G. (1963) : *Use of the Computer to Compare Spectrophotometric Curves*, Management Systems Development Department. Eastman Kodak Company.

WHITTLE, P. (1953) : On principal components and least square methods of factor analysis. *Skand. Aktuarictidskr.*, **36**, 223–239.

WILLIAMS, E. G. (1958) : *Regression Analysis*, John Wiley and Sons.

Paper received : November, 1964.