*Factor Analysis as a Statistical Method**

D. N. LAWLEY and A. E. MAXWELL

1. Introduction

Factor analysis is a branch of multivariate analysis that was developed initially by psychologists, the most prominent pioneers being Spearman, Thomson, Thurstone, and Burt. At first it was concerned primarily with hypotheses about the organization of mental ability suggested by the examination of matrices of correlations between cognitive test variates. The early work, in this difficult field, gave rise to many and protracted controversies of a psychological nature which tended to frighten away all but a few mathematicians and statisticians, so that adequate consideration of the statistical problems involved was long delayed. As a consequence a large number of approximate and sometimes confusing techniques for dealing with these problems appeared in psychological journals and factor analysis became the black sheep of statistical theory. But factor analysis (Bartlett, 1953, p. 23) must be regarded as a natural development in a field in which large sets of correlated variates arise, as a means of examining and describing the internal structure of the covariance and correlation matrices concerned. Today it is the most widely used branch of multivariate analysis in the psychological field, and helped by the advent of electronic computers, its use is quickly spreading to other disciplines, to economics, botany, biology as well as to the social sciences in general. In view of this growing interest it is desirable to get a clear notion of what the subject is about, and in particular to set out the statistical theory as it has been developed to date.

2. Factor Models

When analysing the structure of covariance (or correlation) matrices two approaches, which formally resemble each other to some extent but have rather different aims, are currently employed. The best known of these is principal component analysis following Pearson (1901) and Hotelling (1933); the other stems from the work of Spearman (1904, 1926) and is generally thought of as factor analysis proper. The former is a relatively straightforward method of

* This article contains extracts from a book by the authors with the same title, which is shortly to be published by Butterworths. The extracts are reproduced here by kind permission of the publishers.

"breaking down" a covariance or correlation matrix into a set of orthogonal components or axes equal in number to the number of variates concerned. These correspond to the latent roots and accompanying latent vectors of the matrix. The method has the property that the roots are extracted in descending order of magnitude, which is important if only a few of the components are to be used for summarizing the data. The vectors are mutually orthogonal, and the components derived from them are uncorrelated. Although a few components may extract a large percentage of the total variance of the variates, all components are required to reproduce the correlations between the variates exactly. When the principal component method is employed no hypothesis need be made about the variates. They need not even be random variates, though in practice their observed values are usually regarded as a sample from some population. In contrast to the principal component method the aim in factor analysis (with which this paper will be primarily concerned) is to account for, or "explain," the matrix of covariances by a much smaller number of hypothetical variates or "factors."

To make the distinction between the two methods clear we shall state both of them in terms of the observed variates and the derived components or factors. Let the p observed variates be denoted by $x_1, x_2, \ldots, x_p$. In component analysis an orthogonal transformation is applied to them to produce a new set of uncorrelated variates, $y_1, y_2, \ldots, y_p$. These are chosen such that y_1 has maximum variance, y_2 has maximum variance subject to being uncorrelated with y_1, and so on. No hypothesis need be made regarding the x's. The transformed variates, y_r, are then standardized to give a new set which we shall denote by z_r. The basic equations in a principal component analysis may thus be stated in the form

$$x_i = \sum_{r=1} w_{ir} z_r \quad (i, r = 1, 2, \ldots, p), \tag{1}$$

where z_r stands for the r-th component and w_{ir} is the weight of the r-th component in the i-th variate. In matrix notation the equations (1) may be written

$$\mathbf{x} = \mathbf{W}\mathbf{z}, \tag{2}$$

where $\mathbf{x} = \{x_1\ x_2 \ldots x_p\}$,

$\mathbf{z} = \{z_1\ z_2 \ldots z_p\}$,

and $\mathbf{W} = [w_{ir}]$.

Principal component analysis is most useful when the variates x_i are all measured in the same units. If they are not, the method is more difficult to justify. A change in the scales of measurement of some or all of the variates results in the covariance matrix being multiplied on both sides by a diagonal matrix. The effect of this on the latent roots and vectors is very complicated, and the components

are unfortunately not invariant under such changes of scale. In this respect principal component analysis contrasts unfavourably with the maximum likelihood method of factor analysis which we shall presently describe, though in the latter case the calculations which are required are considerably more onerous.

In contrast to principal component analysis the basic assumption in factor analysis is that

$$x_i = \sum_{r=1}^{k} l_{ir} f_r + e_i \qquad (i = 1, 2, \ldots, p), \tag{3}$$

where f_r is the r-th common factor, k is specified, and e_i is a residual representing sources of variation affecting only the variate x_i. The p random variates e_i are supposed to be independent of one another and also to be independent of the k variates f_r. We shall initially suppose that the latter are orthogonal, or uncorrelated, but later we shall consider the case where they are correlated. Without loss of generality we may take the variance of each f_r to be unity. The variance of e_i we shall denote by v_i. All the means are supposed to be zero. The coefficient l_{ir} is usually termed either the "loading" of the i-th variate on the r-th factor, or the loading of the r-th factor in the i-th variate. The quantities l_{ir}, and usually also the v_i, are taken to be unknown parameters which have to be estimated. Here it may be mentioned that attempts have been made to treat the individual values of the f's also as parameters, but this raises difficulties, a serious one being that when the sample size tends to infinity so also does the number of parameters (Whittle, 1953). This approach will be ignored here.

It is clear that equations (3) are not capable of direct verification since the p variates x_i are expressed in terms of $(p + k)$ other variates which are not observable. However, the equations imply a hypothesis, which can be tested, concerning the variances and covariances of the x's. Before considering this hypothesis we shall digress a little to mention a further contrast between principal component analysis and factor analysis to which Bartlett (1953, p. 32 *et seq.*) has drawn attention.

Principal component analysis is by definition linear and additive and no question of a hypothesis arises, but factor analysis includes what Bartlett calls a *hypothesis of linearity*, which, though it might be expected to work as a first approximation even if it were untrue, would lead us to reject the linear model postulated in equations (3) if the evidence demanded it. Since correlation is essentially concerned with linear relationships it is not capable of dealing with this point, and Bartlett briefly indicates how the basic factor equations would have to be amended to include, as a second approximation, second order and product terms of the postulated factors, to improve the

adequacy of the model. In the amended form the product terms especially would be of interest and the equation

$$x_i = l_{i1}f_1 + l_{i2}f_2 + e_i, \tag{4}$$

which involves just two factors, would then become

$$x_i = l_{i1}f_1 + l_{i2}f_2 + l_{i3}f_3 + e_i, \tag{5}$$

where $f_3 = f_1 f_2$. While the details of this more elaborate formulation have still to be worked out mention of it serves to remind us of the assumptions of linearity implied in equations (3), and to emphasize the contrast between factor analysis and the empirical nature of component analysis.

3. Estimating Factor Loadings by the Centroid Method

As mentioned earlier numerous approximate methods for estimating the loadings in equation (3) have been put forward, and before showing how maximum likelihood estimates may be obtained, one of these methods deserves mention. It is the *centroid* or *simple summation method* and is well described by numerous writers (cf. Burt, 1940; Thurstone, 1947; Thomson, 1951; Jowett, 1958). Although a statistical assessment of the method and an examination of its sampling properties is virtually impossible because of a certain arbitrariness in its procedure, it often gives estimates of the loadings which are fairly close to maximum likelihood estimates and which can be usefully employed when finding the latter.

The centroid method can be illustrated by means of a geometrical model (Thomson, 1954), and in this way the procedures involved are easy to grasp. Let the variates $x_1, x_2, \ldots, x_p$, be represented by vectors radiating from an origin in a space of p dimensions, the cosines of the angles between pairs of them being equal to the corresponding correlations. Let the lengths of the vectors, measured from the origin, be taken equal to the standard deviations of the variates they represent. Now if the directions in which the variates are scored are chosen, by temporarily reversing their signs if necessary, so that as many as possible of the correlations are positive, then the vectors will tend to cluster in a sheaf or pencil. Under these conditions the first centroid of the system is defined as the resultant of the vectors, and will pass somewhere through the middle of the sheaf.

The effect of this centroid can now be removed and, by a further reflection of signs, a new sheaf of vectors formed. A second centroid can now be removed, and so on, until the variance of the variates is completely accounted for. In general we shall have p vectors in a space of p dimensions; but the essential features of the model can be demonstrated by considering just the two-dimensional case.

Suppose that there are two variates x_1 and x_2, with variances s_1^2 and s_2^2 and correlation coefficient r. Then their covariance matrix is

$$\mathbf{A} = \begin{bmatrix} s_1^2 & rs_1s_2 \\ rs_1s_2 & s_2^2 \end{bmatrix}.$$

Represent the variates x_1, x_2 by the vectors OX_1, OX_2 at an angle θ, where

$$OX_1 = s_1, \quad OX_2 = s_2, \quad \cos\theta = r.$$

Then the resultant of OX_1 and OX_2 is represented by the vector OF, where OX_1FX_2 is a parallelogram. This vector, after standardization to unit length, represents the first centroid, or factor, f.

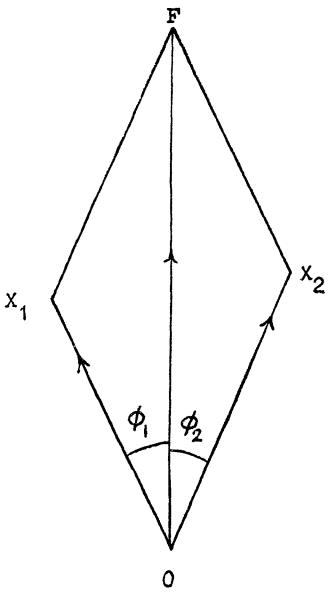

FIG. 1
Vectors OX_1 and OX_2 and the Resultant OF

Let ϕ_1 and ϕ_2 be the angles which OX_1 and OX_2 make with OF, so that $\phi_1 + \phi_2 = \theta$. The loadings of x_1 and x_2 on f are given by

$$\begin{aligned} l_{1f} &= s_1 \cos\phi_1, \\ l_{2f} &= s_2 \cos\phi_2. \end{aligned} \tag{6}$$

If the column vector of loadings is denoted by **l**, then the partial variances and covariances of x_1 and x_2 after eliminating the effect of the first centroid are the elements of the matrix

$$\mathbf{A} - \mathbf{ll}'.$$

This residual covariance matrix, when written out, is

$$\begin{bmatrix} s_1^2 - (s_1 \cos\phi_1)^2 & rs_1s_2 - s_1s_2 \cos\phi_1 \cos\phi_2 \\ rs_1s_2 - s_1s_2 \cos\phi_1 \cos\phi_2 & s_2^2 - (s_2 \cos\phi_2)^2 \end{bmatrix}$$
$$= \begin{bmatrix} s_1^2 \sin^2\phi_1 & -s_1s_2 \sin\phi_1 \sin\phi_2 \\ -s_1s_2 \sin\phi_1 \sin\phi_2 & s_2^2 \sin^2\phi_2 \end{bmatrix},$$

since
$$\begin{aligned} r &= \cos(\phi_1 + \phi_2) \\ &= \cos\phi_1 \cos\phi_2 - \sin\phi_1 \sin\phi_2. \end{aligned}$$

As $s_1 \sin\phi_1 = s_2 \sin\phi_2$, the column and row totals of the above matrix vanish.

By expressing $\cos\phi_1$ and $\cos\phi_2$ in equations (6) in terms of s_1, s_2 and r a simple method of calculating the loadings of the variates on the factor, direct from the covariance matrix **A**, is obtained. For instance

$$\begin{aligned} l_{1f} &= s_1 \cos\phi_1 \\ &= \frac{s_1(s_1 + rs_2)}{\sqrt{(s_1^2 + s_2^2 + 2rs_1s_2)}}. \end{aligned} \tag{7}$$

The numerator in the latter expression is the sum of the elements in the first column of **A**, while the denominator is the square root of the sum of all the elements of **A**. A similar expression can be obtained for l_{2f}.

Next it is necessary to obtain the loadings of the variates on the second centroid. These cannot be obtained directly from the residual matrix since its column and row sums vanish. To proceed further the signs of one or other of the variates must be reversed. This is equivalent to reversing the signs in one row and column of the residual matrix. When this has been done the loadings of the variates on the second centroid are obtained by summing the columns and by dividing each by the square root of the sum of all the entries in the residual covariance matrix. Finally the loading of the variate which had its sign reversed must have its original sign restored.

The arbitrariness in the centroid procedure, which was mentioned earlier, lies in the reversal of signs necessary to break the equilibrium which results after each centroid is extracted, and which arises from the fact that the sums of the columns in each residual matrix are zero. When the number of variates is large more than one variate will usually need to be reversed in sign at each stage, and these reversals can often be done in many different ways. Though numerous rules

for sign reversal have been given, whose aim is to extract maximum variance with each successive centroid, in practice a unique solution is not always possible.

Another difficulty in using the centroid method is that before applying the process described above the diagonal elements of **A**, i.e. the total variances, are replaced by smaller quantities known as "communalities." A communality represents the portion of the variance of a variate which is due to the common factors. In other words it is the total variance minus the residual variance. As a result of this replacement the covariance matrix is reduced to a matrix of rank k, apart from sampling errors. Since the communalities (and often the value of k) are initially unknown they are estimated by trial values; these lead to sets of loadings from which new estimates are obtained, and the final solution is got by an iterative process. Convergence is, however, relatively fast.

A disadvantage of the centroid method is that it is not independent of the metric used. The loadings obtained depend, as in the principal component model, upon the scales in which the variates are measured. The usual practice is to standardize the variates, so that the covariance matrix **A** becomes the correlation matrix. This, however, leads to difficulties when tests of significance are considered. But despite its weaknesses the centroid method is a quick and ready method of getting initial estimates of the factor loadings, and when greater accuracy is desired these estimates can be employed as starting values in the maximum likelihood method of factor analysis, which provides efficient estimates.

4. Estimating Factor Loadings by the Method of Maximum Likelihood

The discussion begins with equations (3) and the assumption is made that the x_i follow a multivariate normal distribution. Their variances and covariances form a matrix $\mathbf{C} = [c_{ij}]$, of order p. The common factors f_r are assumed to be orthogonal and uncorrelated. It follows from equations (3) that the variances and covariances are given in terms of the loadings and the residual variances by

$$\begin{aligned} c_{ii} &= \sum_{r=1}^{k} l_{ir}^2 + v_i, \\ c_{ij} &= \sum_{r=1}^{k} l_{ir} l_{jr} \quad (i \neq j). \end{aligned} \tag{8}$$

In terms of matrix algebra these equations may be written as

$$\mathbf{C} = \mathbf{L}\mathbf{L}' + \mathbf{V}, \tag{9}$$

where $\mathbf{L} = [l_{ir}]$ is the $p \times k$ matrix of loadings and $\mathbf{V}$ is the diagonal matrix with elements v_i.

The basic model thus implies a hypothesis H_0 regarding the covariance matrix $\mathbf{C}$, namely that it can be expressed as the sum of a diagonal matrix with positive elements and a matrix of rank k with positive latent roots. The value postulated for k must not be too large, otherwise this hypothesis would be trivially true. If the v_i were known we should merely require $k < p$, but in the more usual case where they are unknown the condition becomes

$$(p + k) < (p - k)^2,$$

see page 10.

Let $\mathbf{A} = [a_{ij}]$ be a sample covariance matrix whose elements are the usual sample estimates of the variances and covariances of the x_i with n degrees of freedom (corresponding as a rule to a sample of size $n + 1$). Our object is to use the information provided by $\mathbf{A}$ to obtain a set of consistent and efficient estimates of the parameters l_{ir} and v_i, all supposed unknown. Since the x_i are normally distributed, the a_{ij} follow a Wishart distribution and the log-likelihood function is, omitting a function of the observations, given by

$$L = -\tfrac{1}{2}n \log_e |\mathbf{C}| - \tfrac{1}{2}n \sum_{i,j} a_{ij}c^{ij}, \tag{10}$$

where c^{ij} is the element in the i-th row and j-th column of $\mathbf{C}^{-1}$. The sum may alternatively be written as $\mathrm{tr}(\mathbf{AC}^{-1})$, where "tr" stands for the trace or sum of the diagonal elements of a matrix.

Expression (10) is now maximized with respect to the l_{ir} and the v_i. A difficulty arises, however, when $k > 1$ because there are then too many l-parameters in the basic model for them to be specified uniquely. In equations (3) the factors may be replaced by any orthogonal transformation of them. The effect on the loadings is that $\mathbf{L}$ is post-multiplied by a $k \times k$ orthogonal matrix. But any such post-multiplication leaves $\mathbf{LL}'$ and hence also $\mathbf{C}$ unaltered. This means that the maximum likelihood method, though it provides a unique set of estimates of the c_{ij}, leads to equations for estimating the l_{ir} which are satisfied by an infinity of solutions, all equally good from a statistical point of view. In a sense it determines the k-dimensional space in which the f_r lie, but it cannot determine their directions in that space.

In this situation all the statistician can do is to select a particular solution, one which is convenient to find, and leave the experimenter to apply whatever rotation he thinks desirable. We shall in fact choose $\mathbf{L}$ in such a way that the $k \times k$ matrix

$$\mathbf{J} = \mathbf{L}'\mathbf{V}^{-1}\mathbf{L}$$

is diagonal. We shall ignore the possibility that any of the diagonal elements of $\mathbf{J}$ are equal, which is unlikely in practice, and suppose that they are arranged in order of magnitude. This fixes $\mathbf{L}$ except that any column may have its elements reversed in sign.

To maximize L we equate to zero its partial derivatives with respect to the l's and the v's. With a certain amount of algebra, the details of which are given in our forthcoming book, it is found that $\partial L/\partial l_{ir}$ is equal to $-n$ times

$$\sum_j l_{jr}c^{ji} - \sum_{j,u,w} l_{jr}c^{ju}a_{uw}c^{wi},$$

which is the element in the r-th row and i-th column of the matrix

$$\mathbf{L}'\mathbf{C}^{-1} - \mathbf{L}'\mathbf{C}^{-1}\mathbf{A}\mathbf{C}^{-1}.$$

We also find that $\partial L/\partial v_i$ is $-\frac{1}{2}n$ times

$$c^{ii} - \sum_{u,w} c^{iu}a_{uw}c^{wi},$$

which is the i-th diagonal element of the matrix

$$\mathbf{C}^{-1} - \mathbf{C}^{-1}\mathbf{A}\mathbf{C}^{-1}.$$

The equations thus obtained are not in forms which are suitable for direct solution. By use of matrix algebra they can, however, be considerably simplified. Denoting estimates by circumflex accents, we find that

$$\hat{c}_{ii} = a_{ii},$$

or

$$\hat{v}_i = a_{ii} - \sum_{r=1}^{k} \hat{l}_{ir}^2 \quad (i = 1, 2, \ldots, p). \tag{11}$$

We also obtain the equation

$$\hat{\mathbf{L}}' = \hat{\mathbf{J}}^{-1}\hat{\mathbf{L}}'\hat{\mathbf{V}}^{-1}(\mathbf{A} - \hat{\mathbf{V}}), \tag{12}$$

where $\hat{\mathbf{J}}$ is restricted to being diagonal and satisfies

$$\hat{\mathbf{J}} = \hat{\mathbf{L}}'\hat{\mathbf{V}}^{-1}\hat{\mathbf{L}},$$
$$\hat{\mathbf{J}}^2 = \hat{\mathbf{L}}'\hat{\mathbf{V}}^{-1}(\mathbf{A} - \hat{\mathbf{V}})\hat{\mathbf{V}}^{-1}\hat{\mathbf{L}}.$$

Equations (11) and (12) can usually be solved by iteration. The process is well illustrated in an article by Emmett (1949) and will receive further discussion in our book. In general the iterative process is slow, but it can be hastened if fairly good initial estimates of the loadings, such as are provided by a centroid analysis, are available.

A satisfactory property of the maximum likelihood method of estimation, as compared with the principal component and centroid methods, is that it is independent of the metric used. A change of scale of any variate x_i merely induces proportional changes in its loadings l_{ir}. This may easily be verified by examination of the equations of estimation.

5. Testing Hypotheses about the Number of Factors

A satisfactory test of the hypothesis H_0 that there are precisely k common factors is possible only if n is moderately large, in which case a criterion of the large-sample χ^2 type, found by the likelihood ratio method of Neyman and Pearson, can be constructed. The criterion is found to be

$$n\{\log_e(|\hat{\mathbf{C}}|/|\mathbf{A}|) + \mathrm{tr}(\mathbf{A}\hat{\mathbf{C}}^{-1}) - p\}. \tag{13}$$

The number of degrees of freedom for χ^2 is equal to the number, $\frac{1}{2}p(p+1)$, of variances and covariances minus the effective number of unknown parameters which, under H_0, have to be estimated. On account of the indeterminacy arising from possible rotations of the factors the effective number of unknown parameters is

$$p + pk - \tfrac{1}{2}k(k-1) = p + \tfrac{1}{2}k + \tfrac{1}{2}p^2 - \tfrac{1}{2}(p-k)^2.$$

Hence the number of degrees of freedom for χ^2 is

$$\tfrac{1}{2}\{(p-k)^2 - (p+k)\}. \tag{14}$$

This is positive provided that H_0 is non-trivial.

It has been pointed out by Bartlett (1951) that the distribution of the criterion (13) approximates more closely to that of χ^2 if n is replaced by the multiplying factor

$$n' = n - \tfrac{1}{6}(2p+5) - \tfrac{2}{3}k. \tag{15}$$

(Strictly speaking this is slightly conjectural. The multiplying factor in the special case where $k = 0$ is known to be $n - \frac{1}{6}(2p+5)$. For $k > 0$ it seems reasonable to replace n by $n - k$ and p by $p - k$.)

If the equations of estimation have been solved exactly, then we have $\mathrm{tr}(\mathbf{A}\hat{\mathbf{C}}^{-1}) = \mathrm{tr}(\mathbf{I}_p) = p$. Hence expression (13) may be replaced by

$$n' \log_e(|\hat{\mathbf{C}}|/|\mathbf{A}|). \tag{16}$$

On the other hand, if insufficient iterations have been performed to obtain an exact solution, use of (16) may well give an entirely false result. In fact a negative value for χ^2 may even be obtained!

Even with simplifications and with moderate values of p the evaluation of either (13) or (16) is laborious since the determinant of $\mathbf{A}$ is required. Fortunately an approximation can be found which is adequate in most practical cases. This depends on the fact that, for large n, the values of the differences $a_{ij} - \hat{c}_{ij}$ tend to be small, so that terms of degree higher than the second in these differences may be neglected. The approximate χ^2 criterion is

$$n' \sum_{i,j} (a_{ij} - \hat{c}_{ij})^2/(\hat{v}_i\hat{v}_j). \tag{17}$$

This expression is very easy to calculate. It has been found as a rule to provide a good approximation even when comparatively few iterations have been performed and when the exact maximum likelihood estimates have not been attained.

If the value of χ^2 is found to exceed the chosen level of significance, and H_0 is rejected, the conclusion is that in the basic model at least $k + 1$ common factors are required.

6. Factor Interpretation

Once a set of factor loadings or component weights, corresponding to a set of hypothetical variates, has been found the next step is to try to interpret them in a way which will give a reasonable summary of the data. While interpretation is not strictly the statisticians' job the problem has statistical aspects which require consideration. To illustrate this let us take the loadings on three factors of a set of ten cognitive variates. These appear in table I*a*, and the names of the variates may be taken as a fair description of their psychological content.

TABLE I

Loadings on Three Factors for Ten Cognitive Variates

	Unrotated			*Rotated*		
		a			*b*	
Variates	I_0	II_0	III_0	I_1	II_1	III_2
1. Comprehension	0·788	−0·152	−0·352	0·863	−0·106	0·109
2. Arithmetic	0·874	0·381	0·041	0·781	0·548	0·000
3. Similarities	0·814	−0·043	−0·213	0·830	0·069	0·127
4. Vocabulary	0·798	−0·170	−0·204	0·812	−0·018	0·219
5. Digit Span	0·641	0·070	−0·042	0·602	0·209	0·105
6. Picture Completion	0·755	−0·298	0·067	0·662	0·058	0·471
7. Picture Arrangement	0·782	−0·221	0·028	0·703	0·094	0·399
8. Block Design	0·767	−0·091	0·358	0·554	0·397	0·510
9. Object Assembly	0·733	−0·384	0·229	0·576	0·098	0·629
10. Coding	0·771	−0·101	0·071	0·675	0·202	0·336

All the correlations between the variates are positive. This indicates the family relationship between them and accounts for the fact that all the loadings on the first factor are positive. This factor could then be thought of as one of overall general intelligence. When its effect is removed from the correlation matrix and a second factor found some of the loadings on it are positive and some are negative. Thus the second factor divides the variates into two sets which, once the effect of the first factor has been removed, contrast with each other. In this instance the contrast would appear to be in particular between *Arithmetic* and *Object Assembly*, or in psychological language between "numerical" and "spatial" ability. A further

contrast between the variates, after the effects of the first two factors have been removed, is seen on the third factor, and so the problem of interpretation could be carried further.

Another common approach to the interpretation of factors is to apply a linear transformation to the set of loadings given by the analysis and so produce an equivalent set which is in some sense

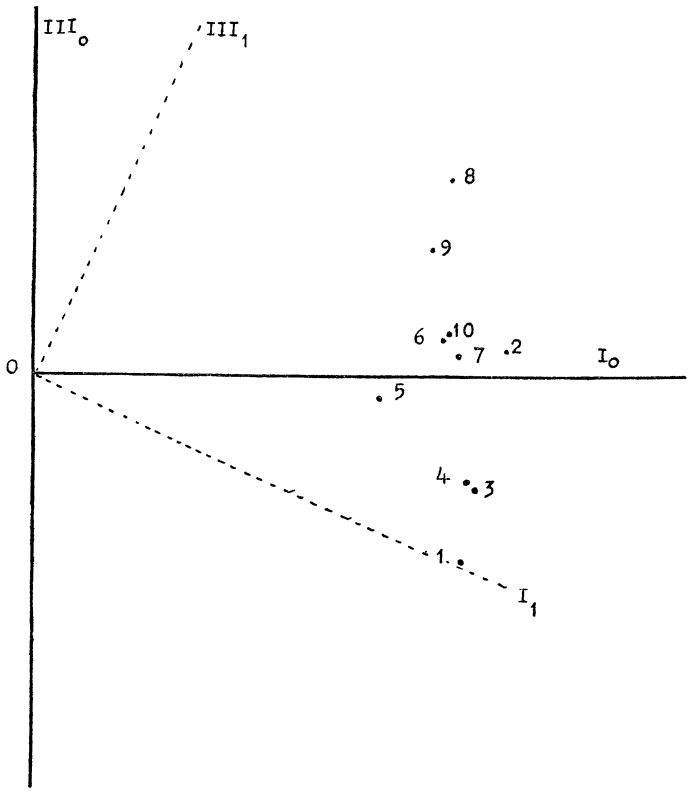

FIG. 2

Plot of Factors I_0 and III_0

simpler than the original. This process is generally known as factor rotation. When rotation is undertaken the experimenter generally has three aims in mind. One is to reduce the number of negative loadings to a minimum, for negative loadings can be awkward to interpret. Another is to reduce to zero, or near zero, as many loadings as possible, for this reduces the number of parameters necessary to describe the data. The third aim is to concentrate on different factors the loadings of variates which contrast with each other, for this may facilitate the interpretation of the factors.

To help us in clarifying possible rotation procedures let us look again at the loadings in table I*a*. An ideal way of proceeding would be to make a three-dimensional model in which three orthogonal axes were taken to represent the factors. The variates could then be represented in the space by points whose co-ordinates were equal to the loadings on the factors. Once this model had been made the actual axes employed could temporarily be forgotten and attention concentrated on the real point of interest, namely, the positions of the points representing the variates relative to each other. Though the axes may be rotated about their origin, or may even be allowed to become oblique, the distribution of the points will remain invariant. When it is examined the points may be found to lie in clusters, or perhaps be concentrated in one or two octants of the space. If this were so it would then be reasonable to fix the axes in a way which would allow the positions of the points to be described as simply as possible.

If the number of factors exceeded three a model would not be possible. The customary procedure thus is to plot the factors two at a time and in that way decide on a set of rotations which simplify the picture. For the data in table I*a* a rotation of factors I_0 and III_0 through about 24° in a clockwise direction, and a rotation of factor II_0 and the new factor III_1 (where subscripts indicate the number of rotations a factor has had) anticlockwise through about 46° led to the loadings given in table I*b*. A further slight rotation of factors I_1 and II_1 would eliminate all negative signs and thus indicate that the axes can be so placed as to include all the points in one octant of the factor space. This procedure incidentally leads to a set of loadings which can be readily interpreted though we need not go into the details here. What is obvious is that other rotations slightly different from ours would also achieve more or less the same result. This subjective element in factor rotation has led to a great deal of disquiet about the interpretation of factors, and to overcome the difficulty various empirical techniques for rotating factors, which in some predetermined sense lead to a unique set of loadings, have been proposed (cf. Thurstone, 1954; Kaiser, 1958). We shall not consider them, but shall instead describe other methods, which seem promising, and allow the investigator to test specific hypotheses about the factor content of his data without being involved in rotation techniques.

7. Estimation when Certain Loadings are Zero

In pilot studies it is natural for an investigator to look at his results from different points of view and to see how they can best be summarized. After this preliminary work has been done he may be in a position to postulate in advance not only the number of factors

to be expected in the analysis of further data but also where the zero, or near zero, loadings on them should occur. If the zero loadings are such as to determine the factors uniquely it is possible to obtain directly a set of loadings which does not require rotation. Approximate methods of doing this have been given elsewhere by one of the authors (Lawley, 1960). The estimates obtained may be used as initial values in iterative procedures based on maximum likelihood. These procedures have been discussed by Anderson and Rubin (1955), by Howe (1955), and by Lawley (1958). They will be fully described in our book and illustrated with numerical data. Here they are briefly summarized.

We shall begin by making the same assumptions as before when we used the maximum likelihood method of estimation. The only difference is that the hypothesis set up is now more specific. We suppose not merely that there are k factors but that certain loadings are *a priori* zero. These will in future be referred to simply as "zero loadings." It is assumed that the number and positions of the zero loadings are such that the equations of estimation have a unique solution.

As an example suppose that with seven variates we make the hypothesis that there are three factors and that the pattern of loadings (using x's to denote those that are non-zero) is as follows:—

$$\begin{array}{ccccccc} x & x & x & x & x & x & x \\ x & x & x & x & x & 0 & 0 \\ 0 & 0 & 0 & x & x & x & x \end{array}$$

In general such a pattern would determine the factors uniquely since any non-trivial orthogonal transformation of the factors would alter the pattern.

The log-likelihood function L is, as before, given by (10), with $\mathbf{C} = [c_{ij}]$ as in (8) and (9). The partial derivatives of L with respect to the l_{ir} and the v_i are also exactly as given previously. Now, however, we maximize L only with respect to the *non-zero* l_{ir} and the v_i. Thus $\partial L/\partial l_{ir}$ is equated to zero only for values of i and r for which l_{ir} is not a zero loading. The resulting equations of estimation, even after simplification, are more complicated in form than before and are less easily solved numerically (unless an electronic computer is available). This is largely due to the fact that the matrix $\hat{\mathbf{J}} = \hat{\mathbf{L}}'\hat{\mathbf{V}}^{-1}\hat{\mathbf{L}}$ is no longer necessarily diagonal. Each iteration therefore requires the inversion of a matrix of order k.

So far we have assumed that (for $k > 1$) the factors are uncorrelated. In some investigations, however, the data have been interpreted in terms of correlated factors. In such cases estimation by maximum likelihood becomes rather more laborious since the correlation coefficients between the factors also need to be estimated.

Let us denote the correlation coefficient between f_r and f_s by ρ_{rs} (with $\rho_{rr} = 1$). As before we may take the variance of f_r to be unity. The basic equations (3) are unchanged, but equations (8) and (9) no longer hold and the variances and covariances of the x_i are now given by

$$c_{ii} = \sum_{r,s} l_{ir} l_{is} \rho_{rs} + v_i,$$
$$c_{ij} = \sum_{r,s} l_{ir} l_{js} \rho_{rs} \quad (i \neq j). \tag{18}$$

In matrix notation this becomes

$$\mathbf{C} = \mathbf{LPL}' + \mathbf{V}, \tag{19}$$

where $\mathbf{P}$ is the matrix of order k with elements ρ_{rs}.

With correlated factors any non-singular linear transformation may be applied to the f_r in equations (3). This must be kept in mind when setting up a hypothesis if the factors are to be determined uniquely. The pattern of zero and non-zero loadings previously given (with $p = 7$ and $k = 3$) is not satisfactory for correlated factors, since the first factor may be replaced by any linear combination of all three factors without altering the pattern. We might instead postulate a pattern such as the following:—

x	x	0	0	x	x	x
x	x	x	x	x	0	0
0	0	0	x	x	x	x

This would in general lead to a unique solution since any non-trivial linear transformation of the factors would alter the arrangement of zeros.

The log-likelihood function given by (10) has now to be maximized with respect to the non-zero l_{ir}, the v_i and the ρ_{rs} (for $r \neq s$). The equations of estimation become even more complicated in form, but after some simplification they can be solved by an iterative procedure. In each iteration two matrices of order k need inversion. It is clear that for $k > 3$ an electronic computer becomes almost a necessity if an exact solution of the equations of estimation is required.

8. Testing Hypotheses involving Zero Loadings

The hypotheses discussed above may be tested by χ^2 criteria of the large-sample type, and these are still given by expressions (13) or (16). The approximation of (17) is, however, no longer valid except in special cases. It has to be replaced by a less simple function of the observations.

If the number of non-zero loadings is denoted by m, then for uncorrelated factors, assumed to be uniquely determined, the number of degrees of freedom for χ^2 is

$$\tfrac{1}{2}p(p+1) - p - m = \tfrac{1}{2}p(p-1) - m.$$

If, on the other hand, the factors are correlated, the number of degrees of freedom must be reduced by $\frac{1}{2}k(k-1)$ to allow for the estimation of the ρ_{rs}, and therefore becomes

$$\tfrac{1}{2}(p-k)(p+k-1) - m.$$

9. The Estimation of Factor Scores

In previous sections we have concerned ourselves mainly with methods of estimating factor loadings. While these problems of estimation constitute the main interest of factor analysis, it is sometimes desirable in practice to go a step further and find equations by which scores on the hypothetical factors may be estimated from a set of observations of the variates. One method of doing this, due to Thomson (1951), is generally known as the "regression method." To illustrate it let us suppose that the factors are uncorrelated and that, by some method, loadings and residual variances have been estimated, forming matrices $\mathbf{L} = [l_{ir}]$ and $\mathbf{V} = [v_i]$ (circumflex accents are here omitted). As before, the estimated covariance matrix for the x_i is $\mathbf{C} = \mathbf{LL}' + \mathbf{V}$.

For a given set of observations of the p variates x_i we clearly cannot estimate, in the usual statistical sense, the values of the k factors f_r and of the p residuals e_i since the number of hypothetical variates exceeds the number of observed variates. We can, however, find linear functions of the x_i which in some sense provide reasonable estimates of the f_r.

If the true values of the f_r were known we could choose their estimates $\hat{f}_r$ in such a way as to minimize, for each value of r, the sum of squares $\Sigma(f_r - \hat{f}_r)^2$, where Σ denotes summation over the sample. Since $\hat{f}_r$ is a linear function of the x_i, this "Least Squares" method would be the same as finding the linear regression of f_r on the x_i. The coefficients of linear regression would be given in terms of the covariances between f_r and the x_i and of the variances and covariances of the x_i. The former cannot be calculated in the usual way, but it seems reasonable to use as estimates of them the elements of the vector

$$\mathbf{l}'_r = [l_{1r}\ l_{2r} \ldots l_{pr}],$$

which is the r-th row of $\mathbf{L}'$. For the variances and covariances of the x_i we use the matrix $\mathbf{C}$.

The estimates of the f_r are then given by

$$\hat{f}_r = \mathbf{l}'_r \mathbf{C}^{-1}\mathbf{x} \quad (r = 1, 2, \ldots, k),$$

where $\mathbf{x}$ is the column vector $\{x_1\ x_2 \ldots x_p\}$. These equations may be written as

$$\hat{\mathbf{f}} = \mathbf{L}'\mathbf{C}^{-1}\mathbf{x}, \tag{20}$$

where $\hat{\mathbf{f}} = \{\hat{f}_1 \hat{f}_2 \ldots \hat{f}_k\}$.

With a little algebra this may be put in the form

$$\hat{\mathbf{f}} = (\mathbf{I} + \mathbf{J})^{-1}\mathbf{L}'\mathbf{V}^{-1}\mathbf{x}, \tag{21}$$

where, as before, $\mathbf{J} = \mathbf{L}'\mathbf{V}^{-1}\mathbf{L}$.

If we neglect sampling errors in $\mathbf{L}$ and $\mathbf{V}$, the covariance matrix for the errors of estimation $\hat{f}_r - f_r$ is

$$(\mathbf{I} + \mathbf{J})^{-1}.$$

If the factors are correlated and their estimated correlation matrix is denoted by $\mathbf{P}$, then the only changes required are that the covariances between the f_r and the x_i should be estimated by the matrix $\mathbf{PL}'$ and that $\mathbf{C} = \mathbf{LPL}' + \mathbf{V}$. We thus have the equation

$$\hat{\mathbf{f}} = \mathbf{PL}'\mathbf{C}^{-1}\mathbf{x}, \tag{22}$$

which may be put in the form

$$\hat{\mathbf{f}} = (\mathbf{P}^{-1} + \mathbf{J})^{-1}\mathbf{L}'\mathbf{V}^{-1}\mathbf{x}. \tag{23}$$

The covariance matrix for the errors $\hat{f}_r - f_r$ is now $(\mathbf{P}^{-1} + \mathbf{J})^{-1}$.

An alternative method of procedure for estimating factor scores is due to Bartlett (1938). Here the principle adopted is the minimization, for a given set of observations, of $\sum_i e_i^2/v_i$, which is the sum of squares of standardized residuals. The summation in this case is over variates. The above sum of squares may be written as

$$\sum_{i=1}^{p} (x_i - \sum_r l_{ir} f_r)^2/v_i,$$

which must be minimized with respect to $f_1, f_2, \ldots, f_k$. This leads to the equations

$$\sum_{i,s} (l_{ir} l_{is}/v_i) f_s = \sum_i (l_{ir} x_i/v_i) \quad (r = 1, 2, \ldots, k),$$

where the estimate of f_r is now denoted by $\check{f}_r$. These equations may be written as

$$(\mathbf{L}'\mathbf{V}^{-1}\mathbf{L})\check{\mathbf{f}} = \mathbf{L}'\mathbf{V}^{-1}\mathbf{x},$$

or

$$\check{\mathbf{f}} = \mathbf{J}^{-1}\mathbf{L}'\mathbf{V}^{-1}\mathbf{x}. \tag{24}$$

Though the sets of estimates obtained by the two methods have been reached by entirely different approaches, a comparison of (21) and (24) shows that, for uncorrelated factors, they are simply related by the equation

$$\check{\mathbf{f}} = (\mathbf{I} + \mathbf{J}^{-1})\hat{\mathbf{f}}. \tag{25}$$

Similarly, for correlated factors, we have

$$\check{\mathbf{f}} = (\mathbf{I} + \mathbf{J}^{-1}\mathbf{P}^{-1})\hat{\mathbf{f}}. \tag{26}$$

In either case one set is a linear transformation of the other.

10. Identifying Factors in Different Populations

A problem which sometimes arises in factor analysis is that of comparing and combining results obtained from different sources. Suppose that random samples have been drawn from each of two multivariate normal populations with different covariance matrices. The question which then naturally arises is whether the same factors are operating in each case. This problem and others related to it have been discussed by earlier writers, notably Thomson and Thurstone in their respective textbooks (Thomson, 1951; Thurstone, 1947), and elsewhere.

Thomson was particularly interested in the effect produced on factors by selection of one or more of the variates. By employing Karl Pearson's selection formulae (Pearson, 1912; Aitken, 1934) he was able to show that such selection may cause factors which were originally orthogonal to become correlated. A further complication is that additional factors may be introduced having loadings in the variates directly selected. It follows that for purposes of comparison it is as a rule useless to perform separate factorial analyses on covariance or correlation matrices obtained by sampling two different populations, especially if the factors are restricted to being orthogonal in each case. For when more than one common factor is involved it becomes difficult to discover the relation between the factors derived from one analysis and those derived from another.

As a result of these findings Thomson came to somewhat pessimistic conclusions regarding the permanence of factors, even for a given set of variates. A more hopeful attitude was adopted by Thurstone, who went some way towards overcoming the difficulties by employing "oblique," or correlated, factors and by imposing his idea of "simple structure." To make further progress, however, a fresh approach seemed to be necessary, and this has led us to propose a new model. In this model the basic assumption is that any selection process operates directly on the factors and only indirectly on the variates. Any two populations are assumed to

differ only as regards the variances and covariances of the factor scores. The value of the model depends ultimately, of course, upon how well it works in practice.

11. A Model for Two Populations

Let us suppose that for each of two p-variate normal populations the same k common factors are in operation. The covariance matrices in the two populations will be denoted by $\mathbf{\Gamma}_1$ and $\mathbf{\Gamma}_2$. The relationship between the variates x_i and the factors f_r is, as before, given by equations (3). The coefficients l_{ir} in these equations are invariant under changes of population and therefore the loading matrix $\mathbf{L}$ is the same for both populations. We shall make what seems a reasonable assumption, that the residual variances are the same for both populations and that they form a matrix $\mathbf{V}$. (The model can be generalized to some extent by allowing the populations to have different residual variance matrices $\mathbf{V}_1$ and $\mathbf{V}_2$, but this complicates the subsequent estimation procedures and we shall not discuss it here.) The population covariance matrices for the x_i are thus given respectively by

$$\begin{aligned} \mathbf{C}_1 &= \mathbf{L}\mathbf{\Gamma}_1\mathbf{L}' + \mathbf{V}, \\ \mathbf{C}_2 &= \mathbf{L}\mathbf{\Gamma}_2\mathbf{L}' + \mathbf{V}. \end{aligned} \tag{27}$$

We suppose, as before, that certain loadings are *a priori* zero, and that the number and positions of these are such as to determine the factors uniquely. The scales of the factors are arbitrary and, for computational convenience, we choose them in such a way that (with n_1 and n_2 as in the next paragraph) the matrix

$$\mathbf{\Gamma} = (n_1\mathbf{\Gamma}_1 + n_2\mathbf{\Gamma}_2)/(n_1 + n_2)$$

has unit diagonal elements. Thus the hypothesis set up is that there are precisely k factors, that certain specified elements of the loading matrix $\mathbf{L}$ are zero and that the population covariance matrices satisfy the relations (27).

Let $\mathbf{A}_1$ and $\mathbf{A}_2$ be the usual sample covariance matrices with respectively n_1 and n_2 degrees of freedom, obtained by taking a random sample from each population. Then the log-likelihood function is, omitting a function of the observations,

$$-\tfrac{1}{2}n_1\{\log_e |\mathbf{C}_1| + \operatorname{tr}(\mathbf{A}_1\mathbf{C}_1^{-1})\} - \tfrac{1}{2}n_2\{\log_e |\mathbf{C}_2| + \operatorname{tr}(\mathbf{A}_2\mathbf{C}_2^{-1})\}. \tag{28}$$

To estimate the unknown parameters we maximize this with respect to the non-zero elements of $\mathbf{L}$, the elements of $\mathbf{V}$, and the elements of $\mathbf{\Gamma}_1$ and $\mathbf{\Gamma}_2$ subject to the restriction that $\mathbf{\Gamma}$ has unit diagonal elements. As in previous cases the resulting equations of estimation

may be simplified and solved iteratively. As yet the estimation procedure has been tried out only on a simple numerical example in which there were six variates and two factors.

The above hypothesis may be tested by means of the criterion

$$n_1 \log_e (|\hat{\mathbf{C}}_1|/|\mathbf{A}_1|) + n_2 \log_e (|\hat{\mathbf{C}}_2|/|\mathbf{A}_2|), \tag{29}$$

which for large samples is distributed approximately as χ^2 with $(p^2 - k^2 - m)$ degrees of freedom, where m is the number of non-zero loadings. The approximation can be improved by the use of a multiplying factor in (29). A large-sample χ^2 criterion may also be derived for testing the hypothesis that $\mathbf{\Gamma}_1 = \mathbf{\Gamma}_2$.

REFERENCES

AITKEN, A. C. (1934). "Note on selection from a multivariate normal population." *Proc. Edin. Math. Soc.*, **4,** 106–110.

ANDERSON, T. W. and RUBIN, H. (1955). "Statistical inference in factor analysis." *Proc. Third Berkeley Symposium*, **5,** 111–150.

BARTLETT, M. S. (1953) . "Factor analysis in psychology as a statistician sees it," in *Uppsala Symposium on Psychological Factor Analysis.* Nordisk Psykologi's Monograph Series No. 3, 23–34.

BARTLETT, M. S. (1938). "Methods of estimating mental factors." *Nature*, **141,** 609–611.

BARTLETT, M. S. (1951). "The effect of standardization on an approximation in factor analysis." *Biometrika*, **38,** 337–344.

BURT, C. (1940). *The Factors of the Mind.* London Univ. Press.

EMMETT, W. G. (1949). "Factor analysis by Lawley's method of maximum likelihood." *Brit. J. Psychol. (Stat. Sect.)*, **2,** 90–97.

HOTELLING, H. (1933). "Analysis of a complex of statistical variables into principal components." *J. Educ. Psychol.*, **24,** 417–441; 498–520.

HOWE, W. G. (1955). *Some Contributions to Factor Analysis*, Report No. ORNL–1919, Oak Ridge National Laboratory.

JOWETT, G. H. (1958). "Factor analysis." *Appl. Statist.*, **7,** 114–125.

KAISER, H. F. (1958). "The varimax criterion for analytic rotation in factor analysis." *Psychometrika*, **23,** 187–200.

LAWLEY, D. N. (1958). "Estimation in factor analysis under various initial assumptions." *Brit. J. Statist. Psychol.*, **11,** 1–12.

LAWLEY, D. N. (1960). "Approximate methods in factor analysis." *Brit. J. Statist. Psychol.*, **13,** 11–17.

PEARSON, K. (1901). "On lines and planes of closest fit to a system of points in space." *Phil. Mag. II*, 6-th series, 557–572.

PEARSON, K. (1912). "On the general theory of the influence of selection on correlation and variation." *Biometrika*, **8,** 437–443.

SPEARMAN, C. (1904). "General intelligence objectively determined and measured." *Amer. J. Psychol.*, **15**, 201–293.

SPEARMAN, C. (1926). *The Abilities of Man.* London.

THOMSON, G. H. (1951). *The Factorial Analysis of Human Ability*, 5-th ed. London Univ. Press.

THOMSON, G. H. (1954). *The Geometry of Mental Measurement.* London Univ. Press.

THURSTONE, L. L. (1947). *Multiple Factor Analysis.* Univ. of Chicago Press.

THURSTONE, L. L. (1954). "Analytical method for simple structure." *Psychometrika*, **19**, 173–182.

WHITTLE, P. (1953). "On principal components and least square methods of factor analysis." *Skandinavisk Aktuarietidskrift*, **35**, 223–239.