STATISTICAL INFERENCE IN FACTOR ANALYSIS

BY

T. W. ANDERSON AND HERMAN RUBIN

TECHNICAL REPORT NO. 30
MAY 25, 1955

PREPARED UNDER CONTRACT N6onr-251 TASK ORDER XIII
(KR-042-993)
OFFICE OF NAVAL RESEARCH

DEPARTMENT OF STATISTICS
STANFORD UNIVERSITY
STANFORD, CALIFORNIA
I. Introduction.

In this paper we discuss some methods of factor analysis. The entire discussion is centered around one general probability model. We consider some mathematical problems of the model, such as whether certain kinds of observed data determine the model uniquely. We treat the statistical problems of estimation and tests of certain hypotheses. For these purposes the asymptotic distribution theory of some statistics is treated.

The primary aim of this paper is to give a unified exposition of this part of factor analysis from the viewpoint of the mathematical statistician. The literature on factor analysis is scattered; moreover, the many papers and books have been written from many different points of view. By confining ourselves to one model and by emphasizing statistical inferences for this model we hope to present a clear picture to the statistician.

The development given here is expected to point up features of model-building and statistical inference that occur in other areas where statistical theories are being developed. For example, nearly all of the problems met in factor analysis are met in latent structure analysis.

There are also some new results given in this paper. The proofs of these are mainly given in a technical Part II of the paper.

---

This work was started while the authors were research associates of the Cowles Commission for Research in Economics. It has been supported in part by the Office of Naval Research.
In confining ourselves to the mathematical and statistical aspects of one model, we are leaving out of consideration may important and interesting topics. We shall not consider how useful this model may be nor in what substantive areas one may expect to find data (and problems) that fit the model. We also do not consider methods based on other models. In doing this, we do not mean to imply that the model considered here is the most useful or important. It seems that this model has some usefulness and importance, it has been studied considerably, and one can give a fairly unified exposition of it.

Extensive discussion of the purposes and applications (as well as other developments) of factor analysis is given in books by psychologists (for example, Holzinger and Harman [10], Thomson [23], Thurstone [24]). Some general discussion of statistical inference has been given in papers by Bartlett [9] and Kendall [12].

Part 1. Expository

2. The Model.

The model we consider is

\[ X = \Lambda f + U + \mu \]

where \( X, U, \) and \( \mu \) are column vectors of \( p \) components, \( f \) is a column vector of \( m \leq p \) components, and \( \Lambda \) is a \( p \times m \) matrix. We assume that \( U \) is distributed independently of \( f \) and with mean \( \mathbb{E}U = 0 \) and covariance matrix \( \mathbb{E}UU^t = \Sigma \), which is diagonal. The vector \( f \) will in some cases be treated as a random vector, and in other cases will be treated as a
vector of parameters which varies from observation to observation. The vector \( X \) constitutes the observable quantities.

The most familiar interpretation of this model is in terms of mental tests. Each component of \( X \) is a score on a test or battery of tests. The corresponding component of \( \mu \) is the average score of this test in the population. The components of \( f \) are the mental factors; linear combinations of these enter into the test scores. The coefficients of these linear
combinations are the elements of \( \Lambda \), and these are called **factor loadings**. Sometimes the elements of \( f \) are called common factors because they are common to several different tests; in the first presentation of this kind of model (Spearman [20]) \( f \) consisted of one component and was termed the general factor. A component of \( U \) is the part of the test score not "explained" by the common factors. This is considered as made up of the error of measurement in the test plus a specific factor, this specific factor having to do only with this particular test. Since in our model (with one set of observations on each individual) we cannot distinguish between these two components of the coordinate of \( U \) we shall simply term the element of \( U \) as the error of measurement.

The specification of a given component of \( X \) is similar to that in regression theory (or analysis variance) in that it is a linear combination of other variables. Here, however, \( f \), which plays the role of the independent variable, is not observed.

We can distinguish between two kinds of models; In one we consider the vector \( f \) to be a random vector, and in the other we consider \( f \) to be a vector of non random quantities which varies from one individual to another. In the second case it would be more accurate to write \( X_{Q} = \Lambda f_{Q} + U + \mu \). In the former case one sample of size \( N \) is equivalent to any other sample of size \( N \). In the latter case, however, a set of observations \( x_{1}, \ldots, x_{N} \) is not equivalent to \( x_{N+1}, \ldots, x_{2N} \) because \( f_{1}, \ldots, f_{N} \) will not be the same as \( f_{N+1}, \ldots, f_{2N} \) and these enter as parameters. Another way of looking at this distinction is that in the latter case we have the conditional distribution of \( X \) given \( f \). The distinction we are making is the one made in analysis of variance models (components of variance and linear hypothesis models).
When \( f \) is taken as random we shall assume \( \Sigma f = 0 \). (Otherwise, \( \Sigma X = \Sigma \Sigma f + \mu \), and \( \mu \) can be redefined to absorb \( \Sigma \Sigma f \).) Let \( \Sigma ff' = V \).

Our analysis will be made entirely in terms of first and second moments. Usually, we shall consider \( f \) and \( V \) to have normal distributions. If \( f \) is not random, then \( f = f_{i} \) for the \( i^{th} \) individual. Then we shall assume usually \( \frac{1}{n} \sum f_{i} = 0 \) and \( \frac{1}{n} \sum f_{i} f_{i}' = V \).

There is a fundamental indeterminacy in this model. Let \( f = Af \) \((f^* = A^{-1} f)\) and \( \Sigma = AA' \), where \( A \) is a nonsingular \( (m \times m) \) matrix. Then (2.1) can be written as

\[
(2.2) \quad X = \Sigma f^* + U + \mu
\]

where here (when \( f \) is random)

\[
(2.3) \quad \Sigma ff^* = A^{-1} (A^{-1})' = I_{n}, \quad \text{say.}
\]

If \( f \) is normal or if we only consider second order moments, the model with \( A \) and \( f \) is equivalent to the model with \( A^* \) and \( f^* \); that is, by observing \( X \) we cannot distinguish between these two models.

Some of the indeterminacy in the model can be eliminated by requiring that \( \Sigma ff' = I \) (or \( \sum f_{i} f_{i}' = N \), if \( f \) is not random). In this case the factors are said to be **orthogonal**; if \( U \) is not diagonal, the factors are said to be **oblique**. When we assume \( V = I \), then (2.3) is \( A^{-1} (A^{-1})' = I \) \((I_{n} = AA')\). The indeterminacy is equivalent to multiplication by an orthogonal matrix; this is called the problem of rotation. Requiring that \( U \) be diagonal means that the components of \( f \) are independently distributed when \( f \) is assumed normal.

This has an appeal to psychologists because one idea of common mental factors is (by definition) that they are independent or uncorrelated quantities.
A crucial assumption is that the components of \( U \) are uncorrelated. Our viewpoint is that the errors of observation and the specific factors are by definition uncorrelated. That is, the interrelationships of the test scores is caused by the common factors, and that is what we want to investigate. There is another point of view on factor analysis that is fundamentally quite different; that is, that the common factors are supposed to explain or account for as much of the variance of the test scores as possible. To follow this point of view, we should use a different model.

At this point we perhaps should indicate another point of view which we do not treat. That is that mental factors are positive quantities; any individual has these to some degree; each test score depends on these in a positive way. This implies that all the coefficients of \( A \) are non-negative. This point of view leads to important and interesting considerations. However, in this paper we shall not consider this.

As in all problems of multivariate statistics, a geometric picture helps the intuition. We consider a \( p \)-dimensional space. The columns of \( A \) can be considered as vectors in this space. They span some \( q \)-dimensional subspace; in fact, they can be considered as coordinate axes in the \( q \)-dimensional space, and \( f \) can be considered as coordinates of a point in that space referred to this particular axis-system. This subspace is called the factor space. Multiplying \( A \) on the right by a matrix corresponds to taking a new set of coordinate axes in the factor space.

3. The Problems.

We now list the considerations which must be made for this model. We point out that exactly the same considerations enter into other models, for example, latent structure analysis. For the sake of outlining these problems we shall assume that \( f \) is random and is normally distributed with \( E f' = I \).
I. Existence of the Index. From (2.1) we deduce that $X$ is normally distributed with mean $\mu$ and covariance matrix

$$
(3.1) \quad \mathbb{E}(X-\mu)(X-\mu)^T = \mathbb{E}(\Lambda\Psi^\ast \Psi)(\Lambda\Psi^\ast \Psi)^T \\
= \mathbb{E}(\Lambda\Psi^\ast \Lambda^\ast \Psi \Psi^\ast \Lambda^\ast \Psi \Psi^\ast \Lambda^\ast \Psi) \\
= \Lambda \mathbb{E}(\Psi^\ast \Psi) \Lambda^\ast \mathbb{E} \Psi^\ast \Psi \\
= \Lambda \Lambda^\ast \Sigma \\
= \Psi
$$

say. Suppose we have some normal population with mean $\mu^\ast$ and covariance matrix $\Psi^\ast$, is there a factor analysis model that can generate this population? Essentially, this is a question whether the equation $\Psi^\ast = \Lambda \Lambda^\ast \Sigma$ can be solved, or rather, what conditions must $\Psi^\ast$ satisfy so that $\Psi^\ast = \Lambda \Lambda^\ast \Sigma$ can be solved. Another way of looking at this problem leads to the formulation of what is the minimum $m$ for which the equation can be solved.

II. Identification. Suppose there is some $\Lambda$ and $\Sigma$ such that $\Psi^\ast = \Lambda \Lambda^\ast + \Sigma$. Does this equation then have a unique solution? From the previous discussion it is clear that the above equation is also satisfied by $\Lambda \Theta$, where $\Theta$ is an orthogonal matrix. We can consider (i) if other restrictions are placed on $\Lambda$, is the solution unique, or (ii) are $\Lambda$ and $\Sigma$ determined uniquely except for multiplication of $\Lambda$ on the right by an orthogonal matrix?

III. Determination of the Structure. Given $\Psi$ (and suppose that (3.1) can be solved uniquely), how do we determine $\Lambda$ and $\Sigma$?

We now turn to the statistical problems.

IV. Estimation of Parameters. A sample of $N$ individuals is drawn, and from these observations we wish to estimate $\mu$, $\Sigma$, and $\Lambda$. It is
assumed that (31) can be solved uniquely and that \( a \) is known. One would like to know the properties of various estimation methods.

V. Test of the Hypothesis that the Model Fits. Here we suppose that \( m \) is given. We test the hypothesis that \( \mathcal{E}(\mathbf{X} - \mu)(\mathbf{X} - \mu)' \) can be of the form \( \Lambda \Lambda' + \Sigma \).

VI. Determination of the Number of Factors. In many situations the number of factors \( m \) cannot be specified in advance of the statistical investigation. In these cases, the investigator wants to use as few factors as possible to "explain" the population. On what basis should he decide that he has the right number of factors?

VII. Other tests of Hypothesis. There are various hypotheses about the parameters, particularly about \( \Lambda \), that are of interest.

VIII. Estimation of Factor Scores. We want to make statements about the \( f \)'s of our observed \( X \)'s.


If \( f \) and \( U \) are normally distributed, the model postulates that the vector of \( p \) test scores \( X \) has a multivariate normal distribution with a vector of means \( \mu \) and a covariance matrix \( \Sigma \) which has the form

\[
\mathcal{E}(\mathbf{X} - \mu)(\mathbf{X} - \mu)' = \mathcal{E}(\Lambda \mathbf{f} + U)(\Lambda \mathbf{f} + U) = \Lambda \Lambda' + \Sigma
\]

where \( \Sigma \) is diagonal and positive definite, \( \Lambda \) is a \( p \times m \) matrix with \( m \) specified, and \( U \) is an arbitrary positive definite matrix of order \( m \). In this case the problem of existence of the structure is the problem whether the distribution of a vector \( X \) has the above form. The question of normality will not be considered here; the vector of means \( \mu \) is unrestricted and hence is of no question. The essential question is whether the covariance matrix of \( X \)
has the form of (4.1); that is, given the pxp positive definite matrix $\Psi$, can it be expressed as $\sum \Lambda \Lambda^T$ ($\sum$ diagonal and $\Lambda$ of size p x m). If $\Psi$ is not normal, we restrict our considerations to second order moments, and the essential problem is the same.

As far as our present problem goes, we can assume that $M = I$, for if we are given a matrix $\Lambda \mu \Lambda^T$, we can write it as $\Lambda^* \Lambda$ by letting $\Lambda^* = \Lambda \Lambda^T$, where $\Lambda$ is a matrix such that $\Lambda \Lambda^T = M$. Thus we ask if there is a $\Lambda$ and $\sum$ such that

$$
\Psi = \sum \Lambda \Lambda^T ,
$$

(4.2)

One way of determining whether $\Psi$ can be expressed in the desired form is to set about solving the equations $\Psi_{ii} = \sum_{k=1}^{m} \Lambda_{ik}^2$ and $\Psi_{ij} = \sum_{k=1}^{m} \Lambda_{ik} \Lambda_{jk}$, $i \neq j$. These are polynomial equations, and there are well known methods for solving them. If there is an algebraic solution, one must ascertain that $\Lambda_{ik}$ are real and $\Psi_{ii}$ are real and non negative.

The algebraic solution is laborious and gives little insight. What we want are conditions on $\Psi$ that can be applied more directly.

A good deal of insight can be obtained by comparing the number of equations with the number of unknowns. In $\Psi$ there are $p(p+1)/2$ elements, and this is the number of equations involving the unknowns $\sum_{ii}$ and $\Lambda_{ik}$. There are $p$ elements of the diagonal $\sum$, and there are $pa$ elements of $\Lambda$. However, in any solution $\Lambda$ can be replaced by $\Lambda \Theta$, where $\Theta$ is an orthogonal (m x m) matrix, and $\Theta$ has $m(m-1)/2$ independent elements; that is, in any solution, $\Lambda$ can be made to satisfy $m(m-1)/2$ additional conditions. Thus the number of equations and conditions minus the number of unknowns to be determined is
\[(4.3) \quad q = \frac{2(p+1)}{2} \quad \frac{m(m+1)}{2} - p - q\]

\[= \left(\frac{3p}{2}\right)^2 - \frac{3p}{2}\]

It can be expected that if \( C \geq 0 \), then an algebraic solution to the equations is possible. If \( C > 0 \), one can expect that no solution is possible; in this case it appears that \( \Psi \) must satisfy some conditions for a solution to be possible. The inequality \( C \geq 0 \) can also be written

\[(4.4) \quad m \geq \frac{2p^2 + 3p - 1}{2} = \rho = \frac{4p^2 - 1}{2}\]

Some values of \( \rho \) and \( \left[ 2p^2 - 1 \right] / 2 \) are

<table>
<thead>
<tr>
<th>( p )</th>
<th>( 2p^2 - 1 ) / 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>2.3</td>
</tr>
<tr>
<td>4</td>
<td>3</td>
</tr>
<tr>
<td>5</td>
<td>4.5</td>
</tr>
<tr>
<td>6</td>
<td>5.2</td>
</tr>
<tr>
<td>7</td>
<td>6</td>
</tr>
<tr>
<td>8</td>
<td>7.3</td>
</tr>
<tr>
<td>9</td>
<td>8.6</td>
</tr>
<tr>
<td>10</td>
<td>9.2</td>
</tr>
<tr>
<td>11</td>
<td>10</td>
</tr>
</tbody>
</table>

This counting of equations and unknowns gives us a rough criterion of solvability; it does not, of course, lead to precise necessary and sufficient conditions for solvability. For one thing we cannot be sure that the equations are independently; another difficulty is that the solution may not be real or yield non-negative \( C_{ij} \).

It is well known that a necessary and sufficient condition that a \( p \times p \) matrix \( A \) can be expressed as \( BB^T \), where \( B \) is \( p \times m \) is that \( A \) be positive semi-definite of rank \( m \). Thus we can state the following:
Theorem 4.1. A necessary and sufficient condition that \( \mathbf{\Sigma} \) be a covariance matrix of a factor analysis model with \( m \) factors is that there exist a diagonal matrix \( \mathbf{\Sigma}^* \) with non-negative elements such that \( \mathbf{\Phi}^{-1} \mathbf{\Sigma}^* \) is positive semi-definite of rank \( m \).

Now the question is how we can tell whether there exists such a diagonal matrix \( \mathbf{\Sigma}^* \). It is instructive to consider the case \( m = 1 \). Then we can expect that \( \mathbf{\Phi} \) has to satisfy \( 6 = (p(p-1))/2 + p \) conditions of equality as well as some inequalities. In this case \( \mathbf{\Lambda} \) is a column vector and \( \mathbf{\Lambda} \mathbf{\Lambda}^t \) is a positive semi-definite matrix of rank one. The question is whether we can subtract non-negative numbers from the diagonal elements of \( \mathbf{\Phi} \) to give a positive definite matrix of rank 1. \( \mathbf{\Phi} - \mathbf{\Sigma} \) will be of rank one if and only if \( \mathbf{\Sigma} \) can be chosen so that all second-order-minors are zero. A second-order minor which does not include a diagonal element is known as a tetrad and has the form

\[
\begin{vmatrix}
\psi_{hi} & \psi_{hj} \\
\psi_{ki} & \psi_{kj}
\end{vmatrix} = \psi_{hi} \psi_{kj} - \psi_{hj} \psi_{ki} \quad (h, i, j, k \text{ different})
\]

These must all be zero. A second-order minor which includes one diagonal element has the form

\[
\begin{vmatrix}
\psi_{ii} & \psi_{ij} \\
\psi_{ki} & \psi_{kj}
\end{vmatrix} = (\psi_{ii} - \psi_{ki}) \psi_{ij} - \psi_{ij} \psi_{ki} \quad (i, j, k \text{ different}).
\]

Setting this equal to zero, shows \( \psi_{ii} \) must be chosen so

\[
\psi_{ii} = \psi_{ii} - \frac{\psi_{ii} \psi_{ki}}{\psi_{kj}} \quad (\psi_{kj} \neq 0).
\]

The conditions that the solution be consistent (i.e., independent of the pair \( j, k \)) are the tetrad conditions. Moreover, these conditions insure that second-order minors containing two diagonal elements are zero.
It can be shown that \( p(p-1)/2-p \) of the tetrad conditions imply \( \Psi_{ij} = a_i a_j \) (\( i \neq j \)), and this in turn implies the tetrad conditions for all \( i, j, k, h \) (all different).

If the tetrad conditions are satisfied, then \( \Phi - \Sigma \) will have rank 1. If this matrix is to be positive semi-definite, the diagonal elements must be non-negative; that is, \( \psi_{ki} \psi_{ij}/\psi_{kj} \geq 0 \). If \( \Sigma \) is to be positive semi-definite, \( \Sigma_{ii} \geq 0 \).

**Theorem 4.2** A necessary and sufficient condition that \( \Phi \) be a covariance matrix of a factor analysis model with one factor is that \( p(p-1)/2-p \) independent tetrad conditions are satisfied and

\[
0 \leq \frac{\psi_{ki} \psi_{ij}}{\psi_{kj}} \leq \psi_{ii}
\]

for one pair \( (j \neq k) \) for each \( i \).

Another way of expressing the condition \( \psi_{ki} \psi_{ij}/\psi_{kj} \geq 0 \) is to ask whether one can multiply some rows and corresponding columns by -1 to obtain a matrix with all non-negative elements.

The case of one factor is of particular interest. In fact, the original theory of Spearman was given for one "general" factor.

A similar analysis can be made for the case \( m = 2 \). However, the conditions become more complicated (see 26).

In Section 8 we shall consider the question of determining from the sample whether a factor analysis model with a given number of factors is adequate to "explain" the situation. The study of the problem of solvability in the population is of importance for the insight it gives into the model and for suggestions of how to use the sample to ascertain whether the model is suitable.
Albert [1] has given a theorem that leads to a direct procedure for determining whether $\Psi - \Sigma$ is of rank $m$. (The procedure does not verify whether $\Psi - \Sigma$ is positive definite.) Suppose that $m$ is the maximum rank of the submatrices of $\Psi$ that do not include diagonal elements. Then the rows and columns of $\Psi$ can be numbered so

$$\Psi = \begin{pmatrix}
\Psi_{11} & \Psi_{12} & \Psi_{13} \\
\Psi_{21} & \Psi_{22} & \Psi_{23} \\
\Psi_{31} & \Psi_{32} & \Psi_{33}
\end{pmatrix}$$

where $\Psi_{11}, \Psi_{12} = \Psi_{21},$ and $\Psi_{22}$ are square submatrices of order $m$ and $\Psi_{12}$ is nonsingular. Then $\Psi - \Sigma$ is of rank $m$ if

$$\Psi_{12} = (\Psi_{11} - \Sigma_1)^{-1} \Psi_{21}, \quad \Psi_{13} = (\Psi_{11} - \Sigma_1)^{-1} \Psi_{21} \Psi_{23}$$

$$\Psi_{32} = \Psi_{31}^{(-1)} (\Psi_{22} - \Sigma_2), \quad \Psi_{33} = \Sigma_3 = \Psi_{31}^{(-1)} \Psi_{21} \Psi_{23}.$$

Albert [2] has further shown that if $\Psi_{31}$ and $\Psi_{32}$ are also of rank $m$, then there is a uniquely determined $\Sigma$ such that $\Psi - \Sigma$ is of rank $m$. 
5. **Problems of the Population: Identification (II).**

Here we assume that there is at least one solution to \( \Psi = \sum \Lambda \Lambda' \), and we ask whether there is more than one solution. More precisely, we assume that there is at least one solution satisfying some conditions and ask whether there is more than one solution satisfying these conditions. Since any solution \( \sum \Lambda \) can be replaced by \( \sum \Lambda \theta \), where \( \theta \) is orthogonal, it is clear that if we are to have a unique solution, some additional conditions must be put on \( \Lambda \) and \( \sum \).

We can distinguish between two kinds of sets of restrictions. A set of one kind will not affect \( \Lambda \Lambda' \), while a set of the other kind may limit \( \Lambda \Lambda' \). A set of restrictions of the first kind is essentially a mathematical convenience, for any solution \( \sum \Lambda \) gives a whole class of solutions \( \sum \Lambda \theta \) and a set of restrictions of the first kind simply picks out of \( \Lambda \theta \) a representative solution. It is fairly clear how we can go from the class of solutions to the representative one and how we can generate the class from the representative solution.

In Section 4 we noted that there are \( p(p+1)/2 \) elements of \( \Psi \), \( p \) elements of \( \sum \), \( pm \) elements of \( \Lambda \) and \( m(m-1)/2 \) independent elements of \( \theta \). We can expect that \( m(m-1)/2 \) restrictions will be needed to eliminate the indeterminacy due to \( \theta \). If \( C = \frac{1}{2}[(p-m)^2-p-m] \) is non-negative we can then expect identification. If \( C \) is negative, we can expect that \(-C\) additional restrictions are necessary for identification; in this case there should be in all \(-C+m(m-1)/2 = p+pm-p(p+1)/2\).

This counting of equations is, of course, inadequate for making precise statements about identification. We shall now investigate the problem more adequately. It is possible to consider conditions on \( \Psi \) that imply
identification (i.e., unique solvability) just as in the previous section we considered conditions on \( \mathbf{Q} \) for solvability. However, it is more convenient to consider conditions involving \( \Sigma \) and \( \Lambda \) for the one assumed solution. We shall first consider conditions assuming that \( \Sigma \) and \( \Lambda \Lambda' \) are determined uniquely.

**Lemma 5.1.** If

\[
\Lambda L' = \Lambda \Lambda',
\]

where \( \Lambda \) and \( L \) are \( pxm \) and \( \Lambda \) is of rank \( m \), then \( L = \Lambda \Theta \), where \( \Theta \) is orthogonal.

**Proof:** The lemma is well known, but we give a proof for the sake of completeness; methods for finding \( L \) subject to certain restrictions are given in Section 6. Since \( \Lambda \) is of rank \( n \), \( \Lambda \Lambda' \) is of rank \( m \) and \( L \) must be of rank \( m \). Multiply (5.1) on the right by \( L(L'LL)^{-1} \) to obtain \( L = \Lambda B \), where \( B = \Lambda (L'L)^{-1} \). Multiplication on the left by \( (L'L)^{-1}L' \) shows \( L = B'B \). Q.E.D.

**Theorem 5.1.** A sufficient condition for identification of \( \Sigma \) and \( \Lambda \) up to multiplication on the right by an orthogonal matrix is that if any row of \( \Lambda \) is deleted there remains two disjoint submatrices of rank \( m \).

**Proof:** Let \( \mathbf{Q} = \Sigma + \Lambda \Lambda' \). To prove the theorem we shall now show that if \( \mathbf{Q} = S + LL' \), where \( S \) is diagonal and \( L \) is \( pxm \), then \( S = \Sigma \) and \( LL' = \Lambda \Lambda' \). Since the off-diagonal elements of \( \Lambda \Lambda' \) and of \( LL' \) are the corresponding off-diagonal elements of \( \mathbf{Q} \), we only have to show that the diagonal elements of \( LL' \) are equal to the diagonal elements of \( \Lambda \Lambda' \).

The condition implies that \( 2m+1 \leq p \). Let

\[
\Lambda = \begin{pmatrix}
\Lambda_1 \\
\Lambda_{m+1} \\
\Lambda_2 \\
\Lambda_3 
\end{pmatrix}, \quad L = \begin{pmatrix}
L_1 \\
L_{m+1} \\
L_2 \\
L_3 
\end{pmatrix}
\]

where \( \Lambda_1 \) and \( \Lambda_2 \) are non-singular, and \( \Lambda_{m+1} \) is the \((m+1)\)th row; \( L \) is partitioned...
in submatrices of the same number of rows. Then

\[
\Lambda' = \begin{pmatrix}
\lambda_1 & \lambda_1' & \lambda_1'_{m+1} & \lambda_1'_{m+1} & \lambda_1'_{m+1}
\lambda_2' & \lambda_2' & \lambda_2'_{m+1} & \lambda_2'_{m+1} & \lambda_2'_{m+1}
\lambda_3' & \lambda_3' & \lambda_3'_{m+1} & \lambda_3'_{m+1} & \lambda_3'_{m+1}
\end{pmatrix}
\]

and \( \Lambda' \) has the same form. Since \( \lambda_1'_{m+1} \lambda_2'_{m+1} = \lambda_{m+1} \lambda_2' \) and \( \lambda_1'_{m+1} \lambda_2' \) are

off-diagonal, \( \lambda_1' \lambda_{m+1} \lambda_{m+1} \lambda_2' = \lambda_{m+1} \lambda_2' \lambda_{m+1} \) and \( \lambda_1' = \lambda_{m+1} \lambda_2' \), which

is non-singular (since \( \lambda_1' \) and \( \lambda_2' \) are non-singular). Since \( \Lambda' \) is of rank \( m \)

\[
0 = \begin{vmatrix}
I_{m+1} & I_{m+1} \\
\ell_{m+1} & \ell_{m+1} \\
\end{vmatrix}
= \begin{vmatrix}
\lambda_1' \lambda_{m+1} & \lambda_1' \\
\lambda_{m+1} \lambda_{m+1} & \lambda_{m+1} \\
\end{vmatrix}
\]

\[
= \frac{(-1)^m}{\ell_{m+1} \ell_{m+1}'} \begin{vmatrix}
\lambda_1' & \lambda_1' \\
\lambda_1' & \lambda_1' \\
\end{vmatrix}
+ f(\Lambda)
\]

Similarly, \( 0 = \frac{(-1)^m}{\lambda_{m+1} \lambda_{m+1}'} \begin{vmatrix}
\lambda_1' & \lambda_1' \\
\lambda_1' & \lambda_1' \\
\end{vmatrix}
+ f(\Lambda) \). Since \( \begin{vmatrix}
\lambda_1' & \lambda_1' \\
\lambda_1' & \lambda_1' \\
\end{vmatrix} \neq 0 \), \( \ell_{m+1} \ell_{m+1}' = \lambda_{m+1} \lambda_{m+1}' \). In the same fashion, we show that the other diagonal elements

of \( \Lambda' \) are equal to those of \( \Lambda \Lambda' \). This proof is patterned after Albert 1.

We can give a geometric interpretation of this condition. The columns

of \( \Lambda \) are vectors in \( p \)-space; the columns of \( \Lambda \) after a row is deleted are

the projections of the vectors on the space of \( p-1 \) coordinate axes. We

require that the projection of these vectors on two different spaces of \( p \)

coordinate axes span the two spaces.

It is fairly clear that the condition is unnecessarily strong in general.

After one communality, i.e., diagonal element of \( \Lambda' \), is determined, it can be
used in determining another. Moreover, the condition that $2n + 1 \leq p$ is much stronger than that $C > 0$. Wilson and Worcester [27] have given an example of $p = 6$ and $m = 3$ where one and only one solution exists.
we now give some theorems that include necessary conditions for identification. It will be assumed now that Σ is positive definite.

Theorem 5.2.

Let $C_m(\Lambda)$ be a condition on $\Lambda$ that is necessary for identification. Then $C_m(\Lambda \Theta)$ for any orthogonal $\Theta$ is also a necessary condition for identification.

Proof.

If $C_m(\Lambda)$ is not true, there is an $S$ and an $L$ such that

$$\Lambda L^2 + \Sigma = L L^1 + S$$

and $\Lambda L^1 \neq L L^1$. If $C_m(\Lambda \Theta)$ is not true, then there is an $S^*$ and $L^*$ such that

$$\Lambda \Theta L^2 \Theta^2 + \Sigma = L^* L^* + S^*$$

and $\Lambda \Theta L^1 \neq L^* L^*$. But the equation implies $\Lambda L^1 + \Sigma = L^* L^* + S^*$ and $\Lambda L^1 \neq L^* L^*$.

Theorem 5.3.

Let $C_m(\Lambda)$ be a condition on $\Lambda$ that is necessary for identification. Let $\Lambda^*$ be a submatrix formed by taking $m^*$ columns of $\Lambda$. Then $C_m^*(\Lambda^*)$ is a necessary condition for identification.

Proof.

Let the columns of $\Lambda$ be arranged so that $\Lambda = (\Lambda^* \Lambda^{**})$. If $C_m^*(\Lambda^*)$ is not true, there is an $S$ and an $L^*$ such that

$$\Lambda^* L^2 + \Sigma = L^* L^* + S$$

and $\Lambda^* L^1 \neq L^* L^*$. Then (5.5) is satisfied for $L = (L^* \Lambda^{**})$ and $\Lambda L^1 = \Lambda^* L^* + \Lambda^{**} L^{**} \neq L L^* + \Lambda^{**} L^{**} = LL^*$. 
Theorem 5.4.

Let $C_{m,p}(\Lambda)$ be a condition on $\Lambda$ that is necessary for identification. Let $\Lambda^*$ be the matrix derived from $\Lambda$ by deleting the rows that have only zero elements. Then $C_{m,p}(\Lambda^*)$ is a necessary condition for identification.

Proof.

Let the rows be numbered so

$$\Lambda = \begin{pmatrix} \Lambda^* \\ 0 \end{pmatrix}, \Sigma = \begin{pmatrix} \Sigma^* & 0 \\ 0 & \Sigma^{**} \end{pmatrix}, \Xi = \begin{pmatrix} \Xi^* & 0 \\ 0 & \Xi^{**} \end{pmatrix}$$

Then $\Xi = \Lambda \Lambda^* + \Sigma$ becomes

$$\Xi^* = \Lambda^* \Lambda^{**} + \Sigma^{**},$$

$$\Xi^{**} = \Sigma^{**},$$

and only the first involves $\Lambda^*$ and $\Sigma^*$.

Lemma 5.2.

If $p = 2$ and $m = 1$, $\Lambda \Lambda^*$ and $\Sigma$ are not identified.

Proof.

In this case

$$\Xi = \begin{pmatrix} \psi_{11} & \psi_{12} \\ \psi_{21} & \psi_{22} \end{pmatrix} = \begin{pmatrix} \sigma_{11} + \lambda_{11}^2 & \lambda_{11} \lambda_{21} \\ \lambda_{21} & \lambda_{11} & \sigma_{22} + \lambda_{21}^2 \end{pmatrix}.$$  

If one component of $\Lambda$ is 0, say $\lambda_{21}$, then $\psi_{12} = \psi_{21} = 0$ and $\psi_{22} = \sigma_{22}$. Then we can take $s_{22} = \sigma_{22}$, $l_{21} = 0$, and $s_{11}$ and $l_{11}$ as any numbers satisfying $s_{11} + l_{11}^2 = \psi_{11} > 0$. If $\lambda_{11} \neq 0$, $\lambda_{21} \neq 0$, then $\psi_{12} \neq 0$. Let $l_{21}$ be any number so that $\psi_{12}^2/l_{11} < l_{21}^2 < \psi_{22}^2$. Then we take $s_{22} = \psi_{22} - l_{21}^2$, $l_{11} = \psi_{12}/l_{21}$, and $s_{11} = \psi_{11} - l_{11}^2 = \psi_{11} - \psi_{21}^2/l_{21}^2$. 
Theorem 5.5.

A necessary and sufficient condition for identification if \( m = 1 \) is that at least three factor loadings be nonzero.

Proof.

The necessity follows from Lemma 5.2 and Theorem 5.4; the sufficiency is a special case of Theorem 5.1.

Theorem 5.6.

A necessary condition for identification is that each column of \( \Lambda A \) has at least three nonzero elements for every nonsingular \( A \).

Proof.

For \( A = I \), the result follows from Theorems 5.3 and 5.5. Then Theorem 5.2 implies the result for \( A \) being orthogonal. If \( A \) is not orthogonal, suppose \( \Lambda A \) has less than three nonzero elements in the \( v \)th column. Then the same will be true for an orthogonal matrix with \( v \)th column proportional to the \( v \)th column of \( A \).

Lemma 5.3.

If \( p = 4 \) and \( m = 2 \), \( \Lambda \Lambda \) and \( \Sigma \) are not identified.

Proof.

Let the rows of \( \Lambda \) be numbered so there is a nonzero element in the first row (Theorem 5.6). We can multiply \( \Lambda \) on the right by an orthogonal matrix so \( \Lambda \) has the form

\[
\begin{pmatrix}
\lambda_{11} & 0 \\
\lambda_{1}^* & \lambda_2^*
\end{pmatrix}
\]

where \( \lambda_{11} \neq 0 \). All components of \( \lambda_2^* \) are nonzero by Theorem 5.6. We shall now find \( \Sigma \) of the form of \( \Lambda \) so
\[
(5.13) \begin{pmatrix}
\sigma_{11} + \chi_{11}^2 \\
\lambda_{11} \lambda_1^{*t} \\
\lambda_{11} \lambda_1^* \\
\end{pmatrix} = \begin{pmatrix}
\lambda_{11} \lambda_1^{*t} \\
\chi_1^{*t} + \lambda_1^{*t} + \chi_2^{*t} \lambda_2^* \\
\end{pmatrix} = \begin{pmatrix}
\chi_{11} + \chi_{11}^2 \\
\chi_1^{*t} + \chi_1^{*t} + \chi_2^{*t} \\
\end{pmatrix}.
\]

Let \( \chi_{11} = k \lambda_{11} \), where \( k \geq 1 \) and \( \sigma_{11} = \sigma_{11} + \chi_{11}^2 = \chi_{11} + \lambda_{11}^2 = \chi_{11} + (1 - k^2) \lambda_{11}^2 > 0 \). Let \( \lambda_1^{*t} = (1/k) \lambda_1^* \). Then

\[
(5.14) \quad \Sigma^{*t} + \lambda_2^{*t} \lambda_2^* = \Sigma^{*t} + \lambda_1^{*t} \lambda_1^* + \lambda_2^{*t} \lambda_2^* = \lambda_1^{*t} \lambda_1^*
\]

is positive definite. If \( 1 - 1/k^2 \) is taken small enough, the non-diagonal elements of the right hand side of (5.14) have the same signs as the corresponding elements of \( \lambda_2^{*t} \lambda_2^* \). By Theorem 4.2 there is a solution of (5.14) for \( S \) and \( \Sigma^{*t} \).

**Theorem 5.7.**

A necessary and sufficient for identification if \( n = 2 \) is that if any row of \( A \) is deleted, the remaining rows of \( A \) can be arranged to form two disjoint matrices of rank 2.

**Proof**

The sufficiency is a special case of Theorem 5.1. To prove the necessity suppose that if we delete the first row of \( A \) there are not two remaining disjoint matrices of rank 2. Let the row of \( A \) be arranged so \( A \) can be partitioned as

\[
(5.15) \quad A = \begin{pmatrix}
\lambda_1 \\
\lambda_2 \\
\lambda_3
\end{pmatrix}
\]

where \( \lambda_2 \) is 2 \times 2 and of rank at most 2, and \( \lambda_3 \) is of rank 1. Since \( \lambda_3 \) is of rank 1, there is an orthogonal matrix \( \Theta \) such that

\[
\lambda_3 \Theta = (\nu, 0), \quad \text{where } \nu \text{ and } 0 \text{ are vectors of } p-3 \text{ components.}
\]

Let
\((5.16) \quad \Lambda \theta = \Lambda^* = \begin{pmatrix} \lambda_{11}^* & \lambda_{12}^* \\ \lambda_{21}^* & \lambda_{22}^* \\ \lambda_{31}^* & \lambda_{32}^* \\ \nu & 0 \end{pmatrix}\)

By Theorem 5.6, \(\lambda_{22}^* \neq 0 \neq \lambda_{32}^*\). After deleting the first row of \(\Lambda^*\), we can get two submatrices of rank 2 only in the form

\[(5.17) \quad \begin{pmatrix} \lambda_{21}^* & \lambda_{22}^* \\ \nu_2 & 0 \end{pmatrix}, \quad \begin{pmatrix} \lambda_{31}^* & \lambda_{32}^* \\ \nu_3 & 0 \end{pmatrix}\]

The assumption that there are not two such matrices of rank 2 implies that \(\nu_i = 0\) except for at most one index \(i\). Then Theorem 5.4 and Lemma 5.2 imply \(\Lambda^*\) (and \(\Lambda\)) is not identified.

**Theorem 5.8.**

A necessary condition for identification is that for each pair of columns of \(\Lambda\) and for every nonsingular \(\Lambda\) when a row is deleted, the remaining rows of this two-column matrix can be arranged to form two disjoint submatrices of rank 2.

**Proof**

This follows from Theorems 5.2, 5.3 and 5.7.

Now let us consider restrictions that eliminate the indeterminacy of rotation. We might note in passing that we consider \(\Lambda\) and \(\Lambda^*\) as equivalent if each column of \(\Lambda^*\) is obtained by multiplying the column of \(\Lambda\) by \(+1\), for replacing a column of \(\Lambda\) by its negative is only equivalent to replacing a factor score by its negative. Each of the following set of restrictions is convenient for a particular method of solving \(C = \Lambda^1\) for \(\Lambda\) (Section 6) and for a method of estimation.
a) Triangular matrix of 0's. This condition is that
\[
\begin{pmatrix}
\lambda_{11} & 0 & 0 & \cdots & 0 \\
\lambda_{21} & \lambda_{22} & 0 & \cdots & 0 \\
\vdots & \vdots & \ddots & \ddots & \vdots \\
\lambda_{m1} & \lambda_{m2} & \lambda_{m3} & \cdots & \lambda_{mn} \\
\lambda_{p1} & \lambda_{p2} & \lambda_{p3} & \cdots & \lambda_{pn}
\end{pmatrix}
\]
(5.18) \( A = \)
that is, that the upper square matrix is triangular. If we think of a row of \( A \) as a vector in \( m \)-space, the condition is that the first row coincide with the first coordinate axis, the second row lie in the plane determined by the first two coordinate axes, etc.

b) General triangularity condition. Let \( E \) be a given \( p \times m \) matrix (of rank \( m \)). Here we require that
(5.17) \[ B' \Lambda = \begin{pmatrix} x & 0 & 0 & \ldots & 0 \\ x & x & 0 & \ldots & 0 \\ x & x & x & \ldots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ x & x & x & \ldots & x \end{pmatrix}, \]

where \( x \) indicates an element not specified zero. It is seen that if \( B' = (10) \), then we obtain condition (a).

(c) **Diagonality of \( \Lambda' \Lambda \).** Here we require that \( \Lambda' \Lambda \) be diagonal and that the diagonal elements of \( \Lambda' \Lambda \) be different and arranged in descending order. Given a positive definite matrix \( A \), there is a uniquely determined orthogonal matrix \( \Theta \) (except for multiplication of columns by \(-1\)) such that \( \Theta' A \Theta \) is diagonal with diagonal elements arranged in descending order assuming that the diagonal elements (which are the characteristic roots of \( A \)) are different. If \( A \) is already in this diagonal form, \( \Theta = I \).

(d) **Diagonality of \( \Lambda' \Sigma^{-1} \Lambda \).** Here we require that \( \Lambda' \Sigma^{-1} \Lambda \) be diagonal and that the diagonal elements be different and arranged in descending order. Rao [17] has related this condition to canonical correlation analysis.

The conditions given above are more or less arbitrary ways of determining the factor loadings uniquely. They do not correspond to any theoretical considerations of psychology: there is no inherent meaning on them. We shall now consider two types of restriction on \( \Lambda \) which may have intrinsic meaning; these conditions may also restrict \( \Lambda' \Lambda \).

**Simple Structure.** These are conditions proposed by Thurstone for choosing a matrix out of the class \( \Lambda \Theta \) that will have particular psychological meaning. If \( \lambda_{ij} = 0 \), then the \( q^{th} \) factor does not enter into the \( i^{th} \) test.
The general idea of "simple structure" is that many tests should not depend on all the factors when the factors have real psychological meaning. This suggests that given a $\Lambda$, one should consider all rotations, that is, all matrices $\Lambda \Theta$, where $\Theta$ is orthogonal, and choose the one giving most $0$ coefficients. This matrix can be considered as giving the simplest structure and presumably the one with most meaningful psychological interpretation.

It should be remembered that the psychologist can construct his tests so that they depend on the factors in different ways.

If we do not require $Eff' = I$, then

$$g(X^{-1})(X^{-1}) = \sum \lambda_0 \lambda'$$

where $M Eff'$. Then $\lambda^M = \lambda \lambda$ and $M^0 = \lambda^O M^{-1} (A^{-})^O$ also satisfies (5.13).

The indeterminacy here is indicated by the non singular matrix $\lambda$. Thurstone has suggested simple structure as a means of identification in this case also. Of course, one needs to add a normalization on each component of $f$ or on each column of $\lambda$ (as well as an ordering of the columns of $\lambda$).

Thurstone (p. 325 of [24]) suggests that the matrix $\lambda$ should be chosen so that there is a submatrix of $\lambda$ (obtained by deleting rows of $\lambda$) say $\tilde{\lambda}$ with the following properties: (1) Each row of $\tilde{\lambda}$ should have at least one non-zero element. (2) Each column of $\tilde{\lambda}$ should have zero elements in at least $n$ rows and these rows should be linearly independent. (It should be pointed out that linear independence is impossible because these rows have zero elements in a given column out of $n$ columns and hence the submatrix of these rows can have maximum rank of $n-1$.) (3) For every pair of columns of $\tilde{\lambda}$ there should be several rows in which one coefficient is zero and one is non-zero. (4) For every pair of columns in $\tilde{\lambda}$, a large proportion of rows should have two nonzero coefficients (if $n > 4$). (5) For every pair of columns
of \( \Lambda \) there should preferably be only a small number of rows with two non zero coefficients.

It is extremely difficult to study the adequacy of these conditions to affect identification. Reiersøl [19] has investigated these conditions, modified a bit. He assumes that there are at least \( m \) zero elements in each column of \( \Lambda \). Let \( \Lambda^{(q')}(q' = 1, \ldots, m) \) be the submatrix of \( \Lambda \) that has zero elements in the \( \alpha \)th column. Reiersøl further assumes that (i) the rank of \( \Lambda^{(q')} \) is \( m-1 \), (ii) the rank of each submatrix obtained by deleting a row of \( \Lambda^{(q')} \) is \( m-1 \), and (iii) the addition to \( \Lambda^{(q')} \) of any row of \( \Lambda \) not contained in \( \Lambda^{(q')} \) increases the rank to \( m \). Then if \( \Lambda \Lambda' \) is identified, a necessary and sufficient condition for the identification of \( \Lambda \) is that \( \Lambda \) does not contain any other submatrices satisfying (i), (ii), and (iii).
Zero elements in specified positions. Here we consider a set of conditions that require the investigator more a priori information. He must know that some tests do not depend on some factors. In this case the conditions are that $\lambda_{i\alpha} = 0$ for certain pairs $(i, \alpha)$; that is, that the $\alpha$th factor does not affect the $i$th test score. In this case we do not assume that $\text{Eff}' = I$. These conditions are similar to some used in econometric models. The coefficients of the $\alpha$th column are identified except for multiplication by a scale factor if (A) there are at least $m-1$ zero elements and (B) the rank of $\Lambda^{(\alpha)}$ is $m-1$ (see [13]).

It will be seen that there are $m$ normalizations and a minimum of $m(m-1)$ zero conditions. This is equal to the number of elements of $\Lambda$. If there are more than $m-1$ zero elements specified in one or more columns of $\Lambda$, then there may be more conditions than are required to take out the indeterminacy in $\Lambda \Lambda'$; in this case the conditions may restrict $\Lambda \Lambda' \Lambda'$.

Local identification. We can ask the question, when we suppose there is a $\Sigma$ and a $\Lambda$ satisfying $\Sigma' = \Sigma^+ \Lambda \Lambda'$ and some other conditions such as $\Lambda' \Sigma^{-1} \Lambda$ being diagonal, is there another pair of such matrices in the neighborhood of $\Sigma$, $\Lambda$. In other words, if we change $\Sigma$ and $\Lambda$ by small amounts, does $\Sigma^+ \Lambda \Lambda'$ necessarily change. If $\Sigma^+ \Lambda \Lambda'$ does change, then we say that $\Sigma$ and $\Lambda$ are locally identified. We can give a sufficient condition for this.

Theorem 7. Let $\Sigma' = \Sigma - \Lambda (\Lambda' \Sigma^{-1} \Lambda)^{-1} \Lambda'$. If $|\theta_{ij}^2| \neq 0$, then $\Sigma$ and $\Lambda$ are locally identified under the restriction that $\Lambda' \Sigma^{-1} \Lambda$ is diagonal and the non diagonal elements are different and arranged in descending order of size.
Proof. Let \( \overline{\Sigma} = \sum + \Lambda \Lambda^* \). Then any pair of matrices \( \sum^*, \Lambda^* \) satisfying \( \overline{\Sigma} = \sum^* + \Lambda^* \Lambda^* \) and \( \Lambda^* \sum_i^* \Lambda_i^* \) diagonal must also satisfy

\[
\Lambda^*(I + \Gamma^*) = \overline{\Sigma} \sum_i^* - 1 \Lambda^*,
\]

\[
\text{diag } \sum_i^* = \text{diag } (\overline{\Sigma} - \Lambda^* \Lambda^*),
\]

\[
\Lambda^* \sum_i^* - 1 \Lambda^* = \Gamma^*,
\]

and the condition that \( \Gamma^* \) is diagonal. As will be seen later, the above equations are analogous to a set of equations defining some estimates. These equations define \( \Lambda^* \) and \( \sum_i^* \) implicitly. We shall show that from these equations one can find the set of partial derivatives \( \frac{\partial \tilde{\sigma}_{ii}^*}{\partial \tilde{\psi}_{jk}}, \frac{\partial \Lambda_{ij}^*}{\partial \tilde{\psi}_{jk}} \). Under the conditions of the theorem the matrix of partial derivatives is of maximum rank (equal to the number of elements in \( \sum_i, \Lambda \)); this is proved in Section 12. The Taylor's series expansion for \( \sum_i^* \) and \( \Lambda^* \) in terms of \( \overline{\Sigma} \) is

\[
(\sum_i^* - \sum_i, \Lambda^* - \Lambda) = L(\overline{\Sigma}^* - \overline{\Sigma})
\]

where \( L \) is a linear function. The right hand side is zero if and only if the left hand side is zero. Q.E.D.

In a sense the study of identifiability is of more relevance than the study of solvability, for identification requires that the investigator specify some features of the model and he wants to know how to do this. As far as solvability goes, in principle, he either has it or he does not, and there is nothing for him to do about it.
6. **Problems of the Population: Determination of the Structure (III).**

The study of solvability and identification implies methods of solving for the structure, given the population of the observables. If the conditions of Theorem 5.1 are satisfied, then the communalities can be determined as indicated in the proof of that theorem; this determines \(\Lambda = C\), say. Let \(\Lambda = (\lambda^{(1)}, \lambda^{(2)}, \ldots, \lambda^{(m)})\), where \(\lambda^{(\alpha)}\) is the \(\alpha\)-th column of \(\Lambda\). Then

\[
C = \lambda^{(1)}(1)^t + \lambda^{(2)}(2)^t + \ldots + \lambda^{(m)}(m)^t.
\]

(6.1)

In many cases one determines the \(\lambda^{(\alpha)}\)'s successively. After \(\lambda^{(1)}\) is found, we define \(C^{(1)} = C - \lambda^{(1)}(1)^t = \lambda^{(2)}(2)^t + \ldots + \lambda^{(m)}(m)^t\), and proceed to find \(\lambda^{(2)}\). In turn we define \(C^{(\alpha)} = C^{(\alpha-1)} - \lambda^{(\alpha)}(\alpha)^t\) and find \(\lambda^{(\alpha+1)}\). The methods depend on the identification conditions.

(a) **Triangularity conditions.** Since the first components of \(\lambda^{(2)}, \ldots, \lambda^{(m)}\) are zero, the first column of \(C\) is \(\lambda_{11} \lambda^{(1)}\); \(\lambda_{11}\) is determined from \(c_{11} = \lambda_{11}^2\) and the rest of \(\lambda^{(1)}\) is found from the first column of \(C\). The matrix \(C^{(1)} = C - \lambda^{(1)}(1)^t\) has only 0's in the first row and column; since the first two components of \(\lambda^{(3)}, \ldots, \lambda^{(m)}\) are zero, the second column of \(C^{(1)}\) is \(\lambda_{22} \lambda^{(2)}\); this determines \(\lambda^{(2)}\). In turn \(\lambda^{(3)}, \ldots, \lambda^{(m)}\) are found similarly.

(b) **General Triangularity Conditions.** Let

\[
(6.2) \quad F = B^t \Lambda = \begin{pmatrix}
\ell_{11} & 0 & \ldots & 0 \\
\ell_{21} & \ell_{22} & \ldots & 0 \\
\vdots & \vdots & \ddots & \vdots \\
\ell_{m1} & \ell_{m2} & \ldots & \ell_{mm}
\end{pmatrix} = (f^{(1)} \ldots f^{(m)}),
\]

(6.3) \quad B = (b^{(1)} \ldots b^{(m)}).
Then \( CB = \Lambda F = \lambda^{(1)} f^{(1)} + \ldots + \lambda^{(m)} f^{(m)} \) and \( B^0 CB = FF = f^{(1)} f^{(1)^t} + \ldots + f^{(m)} f^{(m)^t} \). These two matrix equations can be written

\[
(6.4) \quad Cb^{(\alpha)} = \lambda^{(1)} f^{(1)\alpha} + \ldots + \lambda^{(n)} f^{(n)\alpha}, \quad \alpha = 1, \ldots, m.
\]

\[
(6.5) \quad \beta^{(\beta)} c^{(\beta)} = \beta^{(1)} f^{(1)\beta} + \ldots + \beta^{(n)} f^{(n)\beta}, \quad \beta \leq \alpha = 1, \ldots, m.
\]

The first column of \( CB \) is \( Cb^{(1)} = \lambda^{(1)} f^{(1)} \), and the first element of \( B^0 CB \) is \( b^{(1)} \). We determine \( f^{11} \) and \( \lambda^{(1)} \) from these, which only involve \( b^{(1)} \). The second column of \( CB \) is \( Cb^{(2)} = \lambda^{(1)} f^{(2)1} + \lambda^{(2)} f^{(2)} \), and two more elements of \( B^0 CB \) are \( b^{(1)} \) \( \cdot \) \( Cb^{(1)} = f^{(1)1} f^{(1)} \) and \( b^{(2)} \) \( \cdot \) \( Cb^{(2)} = f^{(1)2} f^{(2)} \). We find \( f^{21}, f^{22}, \) and \( \lambda^{(2)} \) from these which involve only the first two columns of \( B \). In turn we find each column of \( \Lambda \); the \( \alpha \)-th column only requires use of the first \( \alpha \)-columns of \( B \).

There is an alternative method for finding \( \lambda^{(2)} \) after \( \lambda^{(1)} \) is found. Let \( \phi^{(1)} = C - \lambda^{(1)} \phi^{(1)} = \lambda^{(2)} \phi^{(2)} + \ldots + \lambda^{(n)} \phi^{(n)} \) then \( \phi^{(1)} b^{(2)} = \lambda^{(2)} f^{22} \) and \( b^{(2)} \phi^{(1)} = f^{22} \). In turn we define \( \phi^{(\alpha)} = C^{(\alpha)} = \lambda^{(\alpha)} \phi^{(\alpha)} \) and find \( \lambda^{(\alpha+1)} \).

(a) Diagonality of \( \Lambda \). Let \( d_1, \ldots, d_m \) be the nonzero roots of \( |C-dI| = 0 \), ordered in descending order, and let \( \phi^{(\alpha)} \) be the corresponding vectors satisfying \( (C-d\phi^{(\alpha)}) \phi^{(\alpha)} = \phi^{(\alpha)} \) and \( \phi^{(\alpha)} \phi^{(\alpha)^t} = d^{\alpha} \). If \( \phi = (\phi^{(1)}, \ldots, \phi^{(m)}) \) and \( D = (d_1, \ldots, d_m) \) then the equations can be written

\[
(6.6) \quad \phi^{(\alpha)} = d^{\alpha} D,
\]

\[
(6.7) \quad \phi^{(\alpha)} \phi^{(\alpha))} = D.
\]
These equations (and the fact that $D$ is diagonal with ordered elements) determine $\Lambda$ and $D$ uniquely. Since $\Lambda^g \Lambda$ satisfies the equations ($C^g \Lambda = \Lambda^g \Lambda = \Lambda (\Lambda^g \Lambda)$), it is the unique solution.

(d) Diagonality of $\Lambda^g \sum_{1}^{m-1} \Lambda$. Let $d_{1}, \ldots, d_{k}$ be the nonzero roots of

$$ |C \sum_{1}^{m-1} d^j - d^j| = 0 \quad \text{for} \quad d \in \mathbb{C}, \quad |C - d \sum_{1}^{m-1} \Lambda| = 0,$$

ordered in descending order, and let $\Lambda^g(\alpha)$ be the corresponding vector satisfying ($C \sum_{1}^{m-1} d_{\alpha} I \Lambda^g(\alpha) = 0$ and $\Lambda^g(\alpha) \sum_{1}^{m-1} \Lambda^g(\alpha) = d_{\alpha}$). These equations can be summarized as

$$ (6.8) \quad C \sum_{1}^{m-1} \Lambda = \Lambda D, $$

$$ (6.9) \quad \Lambda^g \sum_{1}^{m-1} \Lambda = D.$$

These equations (for $\Lambda^g$ and $D$) have the unique solution $\Lambda = \Lambda$ and $D = \Lambda^g \sum_{1}^{m-1} \Lambda$.

It will be seen later that there is a relation between a method of estimation and a method of determining the structure from the population. However, several methods of estimation can be derived without the motivation of finding an analogue to a method for the population.

7. Problems of Statistical Inference: Methods of Estimation (IV).

7.1 Preliminary remarks. We now consider drawing a sample of $N$ observations on $X$, where $X = \mu + U + \varepsilon$, where $\varepsilon$ has the distribution $N(0, \sum_{1})$ and $U$ has the distribution $N(0, \sum_{1})$; that is, $N$ observations from $N(\mu, \sum_{1} \Lambda \Lambda^t)$. Let the observations be $x_{1}, \ldots, x_{N}$. In all methods of estimation $\mu$ is estimated by

$$ (7.1) \quad \bar{x} = \frac{1}{N} \sum_{1}^{N} x_{\alpha}, $$

which is the maximum likelihood estimate of $\mu$. The estimation of $\sum_{1}$ and $\Lambda$ is based upon
(7.2) \[ A = \frac{1}{N} \sum (x_{\alpha'} - \bar{x}) (x_{\alpha'} - \bar{x})' = \frac{1}{N} \left[ \sum_{\alpha'} x_{\alpha'} x_{\alpha'}' - N \bar{x} \bar{x}' \right]. \]

As is well known, \( \frac{N}{N-1} A \) is an unbiased estimate of the covariance matrix of \( X \).

We shall now consider a number of estimation methods for \( \Lambda \) and \( \Sigma \).

Later we shall consider estimation methods when \( f \) is not considered random, but \( f_{\alpha'} \) is a vector of parameters for the \( \alpha' \)-th individual.

7.2 Maximum Likelihood Estimates for Random Factor Scores, then \( \Lambda^* \) is Unrestricted. Maximum likelihood estimates were derived by Lawley [14] for the case of random factor scores when the restriction on the parameters is that \( \Lambda^* \sum^{-1} \Lambda \) is diagonal (and the diagonal elements are ordered in descending order of size). As was seen earlier this restriction merely takes out the indeterminacy of the rotation in \( \Lambda \). The logarithm of the likelihood function for the sample is

\[
(7.3) \quad -\frac{1}{2} p \mu \log(2 \pi) - \frac{1}{2} N \log \left| \sum^{*} + \Lambda^* \Lambda^{*'} \right| - \frac{1}{2} \sum_{\alpha'=1}^{N} (x_{\alpha'} - \mu^*)' \left( \sum^{*} + \Lambda^* \Lambda^{*'} \right)^{-1} (x_{\alpha'} - \mu^*)
\]

\[
= -\frac{1}{2} p \mu \log(2 \pi) - \frac{1}{2} N \log \left| \sum^{*} + \Lambda^* \Lambda^{*'} \right| - \frac{1}{2} \text{tr} \left[ \left( \sum^{*} + \Lambda^* \Lambda^{*'} \right)^{-1} \right]
\]

\[
- \frac{1}{2} N (\bar{x} - \mu^*)' \left( \sum^{*} + \Lambda^* \Lambda^{*'} \right)^{-1} (\bar{x} - \mu^*)
\]

where we write \( \mu^* \), \( \sum^* \), and \( \Lambda^* \) to denote that these are mathematical variables. It will be noticed that replacing \( \Lambda^* \) by \( \Lambda^* \sigma \), where \( \sigma \) is orthogonal, does not change the likelihood function. Thus if we find \( \mu^* \), \( \sum^* \), and \( \Lambda^* \) to maximize the likelihood function, then \( \mu^* \), \( \sum^* \) and \( \Lambda^* \sigma \) will also maximize it. The restriction that \( \Lambda^* \sum^{-1} \Lambda^* \) be diagonal is a convenience here to make the maximizing variables unique (for almost all samples).
When \( \Lambda^* \) is set equal to \( \bar{x} \), the last term on the right of (7.5) vanishes. It is easy to verify that (for almost all samples) the likelihood function is maximized when the derivatives (subject to \( \Lambda^* \sum^{-1} \Lambda^* \) being diagonal) are set equal to zero. The resulting equations (after considerable algebraic manipulation) are

\[
\begin{equation}
(7.4)
\hat{\Lambda}(1+\hat{\Lambda}) = \Lambda \sum^{-1} \Lambda,
\end{equation}
\[
(7.5)
\text{diag} \sum\hat{\Lambda} = \text{diag}(\Lambda - \hat{\Lambda} \hat{\Lambda}),
\]
\[
(7.6)
\hat{\Gamma} = \hat{\Lambda} \sum^{-1} \Lambda,
\]
\[
(7.7)
\text{non diag} \hat{\Gamma} = \text{non diag} 0,
\]

where \( \text{diag B} \) indicates the diagonal matrix formed from the diagonal elements of \( B \) and \( \text{non diag B} = B - \text{diag B} \). Equation (7.4) can also be written

\[
(7.8)
\hat{\Lambda}\hat{\Gamma} = (\Lambda - \sum) \sum^{-1} \Lambda.
\]

These equations may be compared with (6.8) and (6.9). It is seen that (7.8), (7.6) and (7.7) are similar to equations defining the characteristic vectors and roots of \( \Lambda \) in the metric of \( \sum \). \( \Lambda - \sum \) is the sample analogue of \( \Sigma \). It is assumed that the \( m \) largest roots are positive.

The above equations are practically impossible to solve algebraically. Leahey [14] suggests an iterative procedure which involves approximating \( \sum \), then solving for \( \hat{\Lambda} \), then using this in (7.5) to get a new approximation for \( \sum \), etc. In this paper we shall not discuss in detail computational procedures for any estimates; we hope to consider these in a later paper.

7.5 Maximum Likelihood Estimates for Random Factor Scores when \( \Lambda \Lambda^* \) is Unrestricted and \( \sum = \sigma^2 \text{I}; \text{Principal Components.} \) The assumption that \( \sum = \sigma^2 \text{I} \), that is, that \( \sum \) is a diagonal matrix with all diagonal elements equal is not an assumption that would ordinarily be suitable, but the assumption leads to an estimate of \( \Lambda \) that is closely related to other methods we discuss. Here
(7.9) \[ \Gamma = \Lambda^0 \sum_{\gamma} \Lambda^\gamma = \frac{1}{\sigma^2} \Lambda^0 \Lambda . \]

The condition that \( \Gamma \) is diagonal is equivalent to the condition that \( \Lambda^0 \Lambda \) is diagonal. The equations defining the maximum likelihood estimates are

(7.10) \[ \hat{\Lambda}(\Gamma \ast \mathbf{1}) = \Lambda(-\frac{1}{\sigma^2} \mathbf{1}) \hat{\Lambda} . \]

(7.11) \[ p \hat{\sigma}^2 = \text{tr} (A \ast \hat{\Lambda} \Lambda^\gamma) . \]

(7.12) \[ \Gamma = \Lambda^0 \left( -\frac{1}{\sigma^2} \mathbf{1} \right) \hat{\Lambda} . \]

(7.13) \[ \text{non diag } \hat{\Lambda} = \text{non diag } 0 . \]

Comparison of these equations with (7.4) to (7.7) shows the effect of assuming \( \sum = \sigma^2 \mathbf{1} \). We can write the above equations by letting \( H = \sigma^2 (\Gamma \ast \mathbf{1}) \) as

(7.14) \[ \hat{\Lambda} H = \Lambda \hat{\Lambda} . \]

(7.15) \[ p \hat{\sigma}^2 = \text{tr} (A \ast \hat{\Lambda} \Lambda^\gamma) . \]

(7.16) \[ H = \Lambda^0 \hat{\Lambda} + \sigma^2 \mathbf{1} . \]

(7.17) \[ \text{non diag } \hat{\Lambda} = \text{non diag } 0 . \]

Since \( \text{tr}(A \ast \hat{\Lambda} \Lambda^\gamma) = \text{tr} A - \text{tr} \hat{\Lambda} \Lambda^\gamma = \text{tr} A - \text{tr} \Lambda^0 \hat{\Lambda} = \text{tr} A - \text{tr} (H - \sigma^2 \mathbf{1}) \)

\[ = \text{tr} A - \text{tr} H = -\sigma^2 \]

we have

(7.18) \[ \hat{\sigma}^2 = \frac{1}{p-1} (\text{tr} A - \text{tr} H) . \]

Now let us see the relation of the above equation to those defining the characteristic roots and vectors of \( A \). Let the solutions to \( \det(A - \lambda I) = 0 \) be \( \lambda_1 > \lambda_2 > \cdots > \lambda_p \) and let \( \lambda_1, \lambda_2, \ldots, \lambda_p \) be the corresponding characteristic vectors
(1. e., solutions to \((A-dI)\mathcal{L}_j = 0\) normalized by \(\mathcal{L}_j^\top \mathcal{L}_j = 1\). Let \(D\) be the \(m\times m\) diagonal matrix with \(d_1, \ldots, d_m\) as diagonal elements and let \(L = (\mathcal{L}_1, \ldots, \mathcal{L}_m)\). Then

\[(7.19) \quad AL = LD_d\]

\[(7.20) \quad L^\top L = I_d\]

These equations define \(D\) and \(L\) uniquely (with the condition that the elements of \(D\) are the largest possible). Thus \(D = \Lambda\) and \(L \Lambda = \hat{\Lambda}\), where \(\hat{\Lambda}\) is diagonal.

Then \((p-m) \hat{\sigma}^2 = \text{tr} A - \text{tr} H = \sum d_i = \sum d_i = \sum d_i^2 \). Also

\[(7.21) \quad H = \hat{\sigma}^2 I = \hat{\Lambda}^\top \hat{\Lambda} = \Delta^2 \quad L^\top L \Lambda = \Delta^2\]

Thus the \(\alpha\)-th diagonal element of \(\hat{\Lambda}\), say \(\hat{\lambda}_\alpha\), is \(\sqrt{\hat{\lambda}_{\alpha}} = \hat{\sigma} \hat{\alpha}\), and \(\hat{\Lambda}(\alpha)\) is the characteristic vector \(\hat{\mathcal{L}}_\alpha\). The characteristic vectors \(\hat{\mathcal{L}}_\alpha\) are known as the principal components of \(A\). We see here that these are proportional to the maximum likelihood estimates of \(\lambda\) in our model when \(\sum \hat{\lambda}_\alpha = \hat{\sigma}^2 I\). Hotelling [11] suggested this method when \(\sum \hat{\lambda}_\alpha = 0\), or rather when \(\sum \hat{\lambda}_\alpha\) is very small; his point of view was that \(X\) had an arbitrary normal distribution and \(\hat{\lambda}\) should account for most of the variability of \(X\). For our model we should consider his estimate of \(\lambda\) as \(\hat{\lambda}^2\).

7.4 Thompson's modification of the principal component method for random factor scores when \(\hat{\lambda}\hat{\lambda}^\top\) is unrestricted. For convenience here we require \(\hat{\lambda}\hat{\lambda}^\top\) to be diagonal. The equations are

\[(7.22) \quad \hat{\lambda}_j = (\hat{\Lambda}^\top \hat{\Lambda})\hat{\lambda}_j\]

\[(7.23) \quad \text{diag} \sum \hat{\lambda} = \text{diag} (\hat{\Lambda} \hat{\Lambda}^\top)\]

\[(7.24) \quad J = \hat{\Lambda}^\top \hat{\Lambda}\]
(7.25) \non{dng} J = \non{dng} 0.

Given \( \sum_{\lambda}' \), the characteristic vectors of \( \lambda = \sum_{\lambda} \) corresponding to the largest characteristic roots constitute the columns of \( \lambda \) (normalized according to the diagonal elements of (7.26), i.e., \( \lambda \), the corresponding characteristic roots). Thus the Thomson method [21] is essentially the method of principal components applied to \( \lambda = \sum_{\lambda} \).

This method can be compared to the maximum likelihood method by seeing that the maximum likelihood method involves the characteristic vectors and roots of \( \lambda = \sum_{\lambda} \) in the metric of \( \sum_{\lambda} \).

7.5 The centroid method. This method is based on the algebra used to find \( \lambda \) from \( \theta \) when \( \lambda \) is restricted by \( \lambda' = \theta' \) being triangular (see Section 6). Let \( \sum_{\lambda} \) be an initial approximation to \( \sum_{\lambda} \) and let \( \hat{\lambda} = \lambda' \sum_{\lambda} \). In applying the algebra described in Section 6 we choose the columns of \( \theta \), say \( \theta_o \), in a way that is apparently suitable for this \( \hat{\lambda} \). The first row of \( \theta_o' \) is \( b_o^{(1)} = (1, 1, 2, \ldots, n) \); then an element of \( \hat{\lambda} \) is the sum of the elements of that row of \( \theta_o \) and \( b_o^{(1)}, \) \( \theta_o' \) is the sum of all elements of \( \theta_o \).

To form \( \hat{\theta}_o^{(1)} = \theta_o - \lambda_o \hat{\theta}_o^{(1)} \), and now apply \( b_o^{(2)} \). The elements of this vector are 1 or -1. They are chosen so as to make \( b_o^{(2)} \) \( \hat{\theta}_o^{(1)} b_o^{(2)} \) as large as possible. The computation of \( \hat{\lambda} b_o^{(2)} \) is easy because only addition and subtraction of elements of \( \hat{\theta}_o^{(1)} \) are involved. In turn \( \hat{\theta}_o^{(1)} = \theta_o^{(1)} \hat{\theta}_o^{(1)} \lambda_o \) is computed, and then \( \hat{\theta}_o^{(1)} (\hat{\theta}_o^{(1)} = 2, 3, \ldots, n-1) \). Then \( \lambda_o = (\hat{\lambda}_o^{(1)}, \ldots, \hat{\lambda}_o^{(n)}) \) is a first approximation to the estimate of \( \hat{\lambda}_o \). Next \( \hat{\lambda}_o \) is computed, and the diagonal elements of this matrix (if nonnegative) are taken for \( \sum_{\lambda} \).

Then, the same procedure is followed to obtain \( \hat{\lambda}_{o1} \), another approximation to the estimate of \( \hat{\lambda} \). The matrix taken for \( \theta \), say \( \theta_o \), need not be the same as
B_{0} (except for the first column). In turn \( \sum_{i} \) and \( \Lambda_{i} \) are computed until 
\[ A - \hat{\Lambda}_{i} \hat{\sum}_{i} \] is a close enough approximation to \( \sum_{i} \).

In a sense the centroid method is an approximation to Thomson's modification of the principal components method. In that method the first column of \( \hat{\Lambda} \) is the characteristic vector of \( A - \sum \) corresponding to the largest characteristic root. This vector is proportional to the normalized vector \( y \) (i.e., \( y'y = 1 \)) that maximizes \( y'(A - \sum')y \), and \( y \) satisfies \( (A - \sum')y = \lambda_{1} y \), where \( \lambda_{1} \) is the largest characteristic root of \( A - \sum \). If the elements of \( y \) are about equal, then \( y \) is approximately proportional to \( b^{(1)} \), the first column of \( B \), and hence \( \lambda_{1} y \) is approximately proportional to the first vector of \( \hat{\Lambda} \) found by the centroid. Similarly if the second characteristic vector of \( A - \sum \) is approximately proportional to \( b^{(2)} \), then it is also approximately proportional to the second column of \( \hat{\Lambda} \) by the centroid method. We can say that the centroid method approximated the principal components method by trying to use vectors with elements \( \pm 1 \) as the characteristic vectors of \( A - \sum \).

The big advantage of the centroid method is the ease of computation. Accordingly, it is the most used method.

7.6 Maximum likelihood estimates for random factor scores when \( \Lambda \) is identified by specified zero elements. In this case we have \( C'f = v \), where \( M \) is not required to be diagonal. However, we require the diagonal elements to be unity. Certain coefficients of \( \Lambda \) are required to be zero, say

\[ \hat{\Lambda}_{i} = 0, \ \text{i} = 1, \ldots, p, \ \text{where} \ \hat{\Lambda}_{i} = 0, \ \text{for} \ \lambda = 1, \ldots, m. \]

In the \( \lambda \)-th column of \( \Lambda \), there are \( p_{\lambda} \) zero elements and these are in rows
numbered \( i(1, \varphi), \ldots, i(p, \varphi) \). We assume that these conditions effect identification. We can now apply the method of maximum likelihood. We write down the resulting equations, inserting another unknown matrix \( J \) (essentially Lagrange multipliers) which has zero elements where \( \Lambda \) does not; that is,

\[
(7.27) \quad j_{i \varphi} = 0, \quad i \neq i(1, \varphi), \ldots, i(p, \varphi), \varphi = 1, \ldots, m.
\]

The equations are

\[
(7.28) \quad \text{diag} \sum_\varphi \hat{\Lambda}_{\varphi} = \text{diag} (\Lambda - \hat{\Lambda} \hat{H} \hat{\Lambda}) ,
\]

\[
(7.29) \quad J^0 \Lambda = 0 ,
\]

\[
(7.30) \quad \hat{\Lambda}, \sum_\varphi \hat{\Lambda}^0 = \hat{\Lambda}, \sum_\varphi \hat{\Lambda} = (\hat{H}^{-1} + \hat{\Lambda}^0 \sum_\varphi \hat{\Lambda}^0) \hat{J}^0 \sum_\varphi \hat{\Lambda} .
\]

The derivation of these equations is given in Section 10. We also consider in more detail a special case when \( m = 2 \). The above equations cannot be solved algebraically, but iteration methods can be devised.

7.7 Estimates for non-random factor scores when \( \Lambda \Lambda^0 \) is unrestricted.

We now consider \( x_{\varphi}(\varphi = 1, \ldots, K) \) to be an observation on

\[
(7.31) \quad x_{\varphi} = \Lambda f_{\varphi} + U + \mu ,
\]

where \( f_{\varphi} \) is a fixed vector. Then the expected value of \( x_{\varphi} \) is

\[
(7.32) \quad E x_{\varphi} = \Lambda f_{\varphi} + \mu ,
\]

and the covariance matrix is

\[
(7.33) \quad E(x_{\varphi} = E x_{\varphi}) (x_{\varphi} = E x_{\varphi})' = \sum_\varphi .
\]

This model is similar to the usual model for least squares (or linear regression) except that here the "independent variates", the \( f_{\varphi} \), are unknown; the \( f_{\varphi} \) are also parameters.
In one terminology \( \Lambda, \mu \) and \( \Sigma \) are considered "structural parameters" because they affect all the random variables, and the \( f_{\alpha} \) are considered "incidental parameters" because each \( f_{\alpha} \) affects only one \( X_{\alpha} \). The problem of estimating \( \Lambda \) is essentially equivalent to estimating linear equations on the "systematic parts" of \( X_{\alpha} \). Let \( E X_{\alpha} = f_{\alpha} \). The hypothesis that \( f_{\alpha} \) is of the form \( f_{\alpha} = \Lambda f_{\alpha} + \mu \) is equivalent to the hypothesis that \( P f_{\alpha} = \gamma \) where \( P \) is a \((p-m) \times p\) matrix such that \( P \Lambda = 0 \) and \( P \mu = \gamma \)(that is, that \( f_{\alpha} \) satisfies \( p-m \) linear equations).

If we assume \( U \) has a normal distribution, the likelihood function is

\[
(7.34) \quad \frac{1}{(2\pi)^{\frac{1}{2}}pN \sqrt{\sum_{i=1}^{N} \frac{1}{2}}} \quad e^{-\frac{1}{2} \sum_{i=1}^{N} \left( x_{\alpha_i} - \Lambda f_{\alpha_i} - \mu \right)^2} \sum_{i=1}^{N} \left( x_{\alpha_i} - \Lambda f_{\alpha_i} - \mu \right)^{-\frac{1}{2}}
\]

\[
= \frac{1}{(2\pi)^{\frac{1}{2}}pN} \prod_{i=1}^{k} \frac{1}{\sigma_{i\alpha_i}^{\frac{1}{2}}} \quad e^{-\frac{1}{2} \sum_{i=1}^{N} \left( x_{i\alpha_i} - \gamma_{i\gamma} f_{i\gamma} - \mu_{i\gamma} \right)^2}.
\]

The likelihood function does not have a maximum. To show this, let \( \mu_{11} = 0, \Lambda_{11} = 1, \Lambda_{1\gamma} = 0, \gamma_{1\gamma} = 1, f_{1\alpha_i} = x_{1\alpha_i} \). Then the first term in the product in (7.34) is \( \sigma_{11}^{-\frac{1}{2}} \). As \( \sigma_{11} \to 0 \), this term is unbounded. Thus, the likelihood function has no maximum, and maximum likelihood estimates do not exist. It might be observed that in [15] Lawley obtained some estimation equations by setting equal to zero the derivatives of the likelihood function; it is not clear, however, whether these equations define even a relative maximum, and they obviously cannot define an absolute maximum. (Lawley applied an iterative
method for these equations to some data and found that $\sigma_{11}$ tended towards zero.)

While we cannot apply the method of maximum likelihood to the distribution of $x_1, \ldots, x_n$ to find estimates of all parameters, we can apply the method to the distribution of $A = (1/N) \sum_{\alpha} (x_{\alpha} - \bar{x})(x_{\alpha} - \bar{x})^T$ to find estimates of $\Lambda$ and $\Sigma$. The distribution of $A$ is the non-central Wishart distribution [3] and depends on $\Sigma$ and

$$\sum_{\alpha} \left( \sum_{\alpha} (E x^T_{\alpha} - \frac{1}{N} \sum_{\alpha} E x_{\alpha} \beta) (E x^T_{\alpha} - \frac{1}{N} \sum_{\alpha} E x_{\alpha} \beta)^T \right)$$

$$= \frac{1}{N} \Lambda \sum_{\alpha} f_{\alpha} f^T_{\alpha} \Lambda'^T$$

when $\sum_{\alpha} f_{\alpha} = 0$. If $\Lambda = (1/N) \sum_{\alpha} f_{\alpha} f^T_{\alpha}$, then the matrix is $\Lambda \Lambda'$. If we require $\Lambda = I$, then the matrix is $\Lambda \Lambda'$. With some restrictions on $\Lambda$ to take out the rotation, $\Sigma$ and $\Lambda$ are identified.

The application of the method of maximum likelihood to the distribution of $A$ is detailed in Section II. Of the resulting equations, one set of $m$ are extremely complicated and cannot be solved explicitly. The other equations are similar to the equations obtained by applying the method of maximum likelihood to the case of random factors.

The question arises whether the maximum likelihood estimates for the case of random factors are suitable for the case of non-random factors. In Section II we prove that the estimates based on maximizing the non-central Wishart likelihood function are asymptotically equivalent to the maximum likelihood estimates for random factors in the sense that $\sqrt{N}$ times the difference of the two respective estimates converges stochastically to zero. It would, therefore,
appear that for large samples in the case of non-random factors one can use the maximum likelihood estimates for random factors.

Another asymptotic result that is proved in Part II is that under certain suitable identification conditions the asymptotic distribution of the maximum likelihood estimate of \( \Lambda \) for random factors is the same whatever the assumption on the factors.
7.8 Units of Measurement

In the preceding sections we have considered factor analysis methods applied to covariance matrices. In many cases the unit of measurement of each component of $x$ is arbitrary. For instance, in psychological tests, the unit of scoring has no intrinsic meaning. We now consider how changes in the units of measurement affect the analysis.

Changing the units of measurement means multiplying each component of $x$ by a constant; we are interested in cases where not all of these constants are equal. It would be desirable that when a given test score is multiplied by a constant the factor loadings for the test are multiplied by the same constant and the error variance is multiplied by the square of the constant. Suppose $Dx = x^*$, where $D$ is a diagonal matrix and not all the diagonal elements are the same. Then $E x^* = D\mu = \mu^*$, say, and

$$(7.36) \quad E(x^* - \mu^*)'(x^* - \mu^*) = D \Sigma D = DA(D'\Lambda)' + D \Sigma^2 D = \Phi^*$$

say. Now represent this as

$$(7.37) \quad \Phi^* = \Lambda^* \Lambda'^* + \Sigma^*$$

Clearly $\Sigma^*$ can be taken as $D \Sigma D$ and $\Lambda^* \Lambda'^*$ can be taken as $D\Lambda(D\Lambda)'$ (and must be taken this way if $\Sigma$ and $\Lambda\Lambda'$ are identified), but whether $\Lambda^*$ can be taken as $D\Lambda$ depends on what kind of restrictions are imposed on $\Lambda$ and $\Lambda^*$ to make each unique. If $\Lambda$ (and $\Lambda^*$) is required to have an upper triangular matrix of 0's, then so does $D\Lambda$ and $DA = \Lambda^*$. If $B\Lambda'\Lambda$ (and $B^*\Lambda^*$) is required to have an upper triangle of 0's, then usually $B^*DA$ will not, and $DA \neq \Lambda^*$. If $\Lambda$ (and $\Lambda^*\Lambda'$) is required to be diagonal then usually $(DA)'DA \neq \Lambda'^*DA$ will not, and $DA \neq \Lambda^*$.

If $\Lambda' \Sigma^{-1}$ (and $\Lambda^2 \Sigma^{-1} \Lambda^*$) is required to be diagonal,
then \( (D^2)^* (D \sum D)^{-1} D^2 \lambda = \lambda^* \sum_{-1} \lambda^* \) and \( D^2 \lambda = \lambda^* \).

Now let us see how the estimation methods depend on the units of measurement. Let \( D x_{\lambda} = x_{\lambda^*} \). Then \( MA^* = \sum (x_{\lambda} - x_{\lambda^*})(x_{\lambda} - x_{\lambda^*})^2 = NDAD \). The equations for the maximum likelihood estimates of Section 7.2 are then

\[
\begin{align*}
(7.37) & \quad \lambda^* (I + \hat{\lambda}^*) = DAD \sum^*-1 \lambda^* , \\
(7.38) & \quad \text{diag} \sum^* = \text{diag} (DAD - \hat{\lambda}^* \hat{\lambda}^*), \\
(7.39) & \quad \hat{\lambda}^* = \lambda^* \sum^*-1 \lambda^* , \\
(7.40) & \quad \text{non diag} \hat{\lambda}^* = \text{non diag} 0.
\end{align*}
\]

Clearly \( \hat{\lambda}^* = D \lambda \), and \( \sum^* = D \sum D \) is the solution (when \( \lambda \) and \( \sum \) is a solution to (7.4) to (7.7)). Then the results of this method do not essentially depend on the units of measurement.

The second estimation procedure considered assumed \( \sum = \sigma^2 I \). In the new units \( \sum^* = D \sum D = \sigma^2 D^2 \) which is not proportional to \( I \) and therefore, if this method is applicable to \( \overrightarrow{\psi} \), it is not applicable to \( \overrightarrow{\psi}^* \).

In the third method the transformed equations are

\[
\begin{align*}
(7.41) & \quad \lambda^* J^* = (DAD - \sum^*) \lambda^* , \\
(7.42) & \quad \text{diag} \sum^* = \text{diag} (DAD - \hat{\lambda}^* \hat{\lambda}^*), \\
(7.43) & \quad J^* = \lambda^* \hat{\lambda}^* , \\
(7.44) & \quad \text{non diag} J^* = \text{non diag} 0.
\end{align*}
\]

We know that because of (7.43) and (7.44) \( \hat{\lambda}^* \neq D \lambda \), but we can ask the question whether \( \hat{\lambda}^* = D \lambda P \), where \( P \) is orthogonal; that is, whether \( \lambda^* \lambda^* = D \hat{\lambda}^* \hat{\lambda}^* D \) (whether \( \hat{\lambda}^* \) defines the same factor space
as $D \Lambda D$. If $\Lambda^* \Lambda^* = D \Lambda D$, then $\sum^* = D \sum D$ and (7.41) can be written

$$(7.45) \quad (D^{-1} \Lambda^*) J^* = (A - \sum) D^2 (D^{-1} \Lambda^*)$$

This indicates that the diagonal elements of $J^*$ are the $m$ largest roots of

$$(7.46) \quad |(A - \sum) - jD^2| = 0,$$

and the columns of $D^{-1} \Lambda^*$ are the corresponding vectors satisfying

$$(7.47) \quad [(A - \sum) D^2 - j I] \Lambda^*(a) = 0.$$ 

However, the roots of (7.46) will, in general, not be the roots of (7.46) and the vectors satisfying (7.47) will not span the same linear subspace as the first $m$ characteristic vectors of $A - \sum$.

Thus changing the units of measurement will change the estimated factor space in the Thomson method.

Now let us consider the centroid method. Since we know that if $B^* \Lambda$ has an upper triangle of $0's$, then $B^*D \Lambda$ in general will not, we ask whether the centroid method applied to $A^* = D \Lambda D$ will give $\Lambda^* \Lambda^* = D \Lambda \Lambda D$; that is, whether $\Lambda^* = D \Lambda P$, where $P$ is some orthogonal matrix. In the original metric, we have $\hat{\mathbf{C}}B = \Lambda P'$, where $\hat{\mathbf{C}} = A - \sum$ and $\text{diag} \sum = \text{diag} (A - \Lambda \Lambda)$. If $\Lambda^* = D \Lambda P$, then $\text{diag} \sum^* = \text{diag} (D \Lambda \Lambda D) = \text{diag} D \sum D$ and $\hat{\mathbf{C}}^* = \hat{D} \mathbf{C} \mathbf{D}$. Then $\Lambda^* P^* = \hat{C}B = D \hat{D} \mathbf{C} \mathbf{D}$. Let $\Lambda^* = D \Lambda$. Then $\Lambda P^* = \hat{D} \mathbf{C} \mathbf{D}$ and we ask whether $\Lambda = \Lambda P$. This can be true in general only if $DB(P^*)^{-1} = B(F)^{-1}P$; that is, only if $DBQ = B$ for some nonsingular $Q$. In general, this is not true (only if the $m$ columns of $B$ lie in an $m$-dimensional space spanned by some $m$ characteristic vectors of $D$). However, in the centroid method
the choice of \( B \) is left to the investigator, subject to the conditions that the first column is composed of \( 1 \)'s and the other columns have \( 1 \)'s and \(-1\)'s as elements. Thus, the \( B^* \) used for \( A^* \) would usually not be the \( B \) used for \( A \). Then we would need \( DB^*Q = B \). While it is hard to describe exactly how \( B \) is chosen by the investigator, we can say roughly that the columns of \( B \) are selected as characteristic vectors of \( C \), and thus the columns of \( \hat{\Lambda} \) are approximately proportional to the first \( m \) characteristic vectors of \( C \). But in the latter case we have shown that the transformation of \( A \) to \( DAD \) does not transform \( \hat{\Lambda} \hat{\Lambda}^t \) to \( D \hat{\Lambda} \hat{\Lambda}^t D \); hence, we can conclude that to the extent that the centroid method approximates the principal components method (applied to \( C \)), it does not transform properly with changes of scale of measurement.

In the case where \( \Lambda \) is identified by 0 coefficients in specified positions, \( D \Lambda \) satisfies the same conditions and hence \( \Lambda^* = D \Lambda \). In estimation \( J^* \) has 0's specified in the same positions as in \( J \). It is a straightforward matter to show that \( \hat{\Lambda}^* = D \hat{\Lambda} \), \( \hat{\Sigma}^* = D \sum D \), \( \hat{M}^* = \hat{M} \), and \( \hat{J}^* = D^{-1}J \) satisfy (7.28), (7.29) and (7.30) when \( A^* = DAD \).
In the case of non-random factor scores we suggest applying the method of maximum likelihood to the likelihood of $A$. It can be seen from results in Part II that $\hat{D}$ and $D \hat{\Sigma} D$ satisfy the condition for removing the rotation from $\hat{\Lambda}$ (and $\hat{\Lambda}$). The value of the likelihood function at $A \hat{\Lambda}$ is the same as at $D \hat{\Sigma} D$, $D \hat{\Lambda}$; hence, the maximum for $A^* = D \hat{\Sigma} D$, $\hat{\Lambda}^* = D \hat{\Lambda}$ when the maximum for $A$ is at $\hat{\Sigma} \hat{\Lambda}$.

As has been noted above, the estimation of $\Lambda$ by the centroid or principal components method depends essentially on the units of measurement of the test scores, even though these units may have no intrinsic meaning. A practical remedy to this undesirable indeterminacy is to prescribe a "statistical" unit of measurement. It is customary to let the sample determine the unit of measurement by requiring that each test score have sample variance 1. Thus $d_{ii}$ is taken to be $1/\sqrt{\hat{a}_{ii}}$. The new matrix is $R = (r_{ij})$, where $r_{ij} = a_{ij}/\sqrt{\hat{a}_{ii} \hat{a}_{jj}}$ are the sample correlation coefficients. Besides taking out the indeterminacy, this convention has some other advantages. From the practical point of view it is convenient to have the diagonal elements unity and the other numbers between -1 and +1; this makes it easier to find rules of thumb and convenient computational procedures.

The centroid method is an approximation to modified principal components method. If we compare the equations for the latter with those for the maximum likelihood solution, we see that when $\hat{\Sigma}$ is roughly proportional to $I$, then the principal component estimates are close to the maximum likelihood estimate, which have certain desirable properties (e.g., asymptotic efficiency). If the transformation to test scores with unit sample variance, tends to make the error variances ($\hat{\sigma}_{ii}^2$) approximately equal, then the
efficiency of these procedures is presumably improved.\(^2\)

It might be pointed out that the assumption of Section 7.3 that \(\sigma^2 = I\) is a little less restrictive than it seems. Suppose one knows the error variances except for a constant of proportionality \(\sigma^2\). Then \(\Sigma = \sigma^2 D^{-2}\), say, where \(D^{-2}\) is known. Then we can let \(Dx_A = x_A^*\) and apply the principal components method to \(A^* = DAD\). It might also be noted that Whittle [26] has treated the non-random factor case under the assumption that \(\Sigma = \sigma^2 I\) and has obtained a solution for \(A\) in terms of the principal components of \(A\).

7.9 Invariance of factor loadings under changes of factor score populations.

Now let us consider the model \(X = \Lambda f + \mu + U\), where \(E \mu = 0\) is not necessarily required to be the identity, and where \(f\) and \(U\) are considered random. Of the various ways of identifying \(\Lambda\) (where \(\Sigma\) is identified), consider (a) \(B^0 \Lambda\) has an upper triangle of 0's and \(M = I\), (b) \(\Lambda' \Lambda\) is diagonal and \(M = I\), (c) \(\Lambda' \Sigma^{-1} \Lambda\) is diagonal and \(M = I\), and (d) \(\Lambda\) has specified 0's. Only the last does not involve \(M\).

A mathematical factor analysis is supposed to be a representation of some real population of individuals from which we sample randomly. In defining such a representation it is desirable that at least

\(^2\) Whittle [\_] has suggested that if one assumes the variance of the measurement is proportional to the error variance, then it is reasonable to use the correlation matrix. In the case of non-random factor scores, he has assumed \(X_{iy} \Lambda_{iy}^2 = c \sigma_{ii}\), but finds he is led to principal components of \(R\) only in the case of \(m = 1\).
certain parts of the model do not change even though the population is changed. For example, consider a model for certain mental test scores of a certain population, say, boys of age 16 in New York State. Then consider a sub-population, say, boys of age 16 in eleventh grade in New York State. Can the same model apply to this sub-population? To put it another way, if one investigator factor analyses the first population and another analyses the second, what results of the analyses might be common to the two studies (see also [18]).

If the definition of the sub-population is independent of \( f \) and \( U \) (i.e., does not depend on the factor scores and "errors" including specific factors), then the sub-population is a miniature of the first and any model for the first furnishes the same model for the second. However, in the example above it would seem reasonable that the sub-population involves a selection based on the factor scores related to the set of tests considered (as well as other factors).

Let us consider what happens in the above model if \( f \) is replaced by \( g \), where \( \xi_g = \gamma \) and \( \xi(g - \gamma)(g - \gamma)^t = P \). Then in the sub-population

\[
(7.48) \quad x^* = \Lambda_g + \mu = U = \Lambda(g - \gamma) + (\mu + \Lambda\gamma) + U.
\]

The investigator is going to represent this as

\[
(7.49) \quad x^* = \Lambda^*f^* + \mu^* + U^*,
\]

where \( \xi U^* = 0, \xi f^* = 0, \xi U^* U^* = \sum^* \xi f^* f^* = 1^* \) and \( \Lambda^* \) and \( 1^* \) satisfy the identification conditions.

Let \( \mu^* = \mu + \Lambda \gamma, U^* = U, \sum^* = \sum, \) and \( f^* = Q(g - \gamma) \) and \( \Lambda^* = \Lambda Q^{-1} \) for some non-singular \( Q \). It is clear that the columns of \( \Lambda^* \) span the same space as the columns of \( \Lambda \). If (d) is used for identification \( Q \) must be diagonal, and each column of \( \Lambda^* \).
must be proportional to the corresponding column of $\Lambda$; also $Q_{ij} = 1/\sqrt{p_{ij}}$. If the normalization of a column of $\Lambda$ (and $\Lambda^*$) is done by a rule involving only that column (e.g., by means of making a specified element equal to one), then that column of $\Lambda$ is equal to that column of $\Lambda^*$. It can also be shown that if simple structure effects identification of $\Lambda$ in the original population, it will in the second and will lead to a $\Lambda^*$ with proportional columns. In the case of identification by methods (a), (b), or (c) $\Lambda^*$ is not related as simply to $\Lambda$. In each case $M^* = QPQ^{-1} = I$. In (a) $Q$ also satisfies $B^t \Lambda^* = B^t \Lambda Q^{-1} = P^*$ (with upper triangle of $0$'s); in (b) $\Lambda^*; \Lambda^* = (Q^{-1})^t \Lambda^t \Lambda Q^{-1}$ is diagonal; in (c) $\Lambda^*; \Sigma^{-1} \Lambda^* = (Q^{-1})^t \Lambda^t \Sigma^{-1} \Lambda Q^{-1}$ is diagonal. In each of these cases $\Lambda^*$ will in general, not be a rotation of $\Lambda$. Thus, only if identification does not essentially involve $\Lambda$, can one hope that the results of a factor analysis for one population will bear a simple relation to the results for another population if the two populations differ with respect to the factors involved.

There seems to have been a considerable discussion by psychologists of the requirement that $R = I$. Some claim that the orthogonality (i.e., lack of correlation) of the factor scores is essential if one is to consider the factor scores as more basic than the test scores. However, if the factor scores are orthogonal for some population, in general they will not be orthogonal for another population or for a subpopulation. Hence, this requirement would seem to lead to a less basic definition of factors.

We might also consider the effect of the use of correlations in factor analysis on the comparability of analyses of different populations. In the original population $\Phi = \Lambda M \Lambda^t + \Sigma$ and $R = D \Psi D = (D \Lambda) M (D \Lambda)^t + D \Sigma D$, where $d_{ij}^2 = \zeta_{ij} \lambda_{ii} \mu_{ji}^* + c_{ij}^*.$