PSYCHOMETRIKA—VOL. 31, NO. 2
JUNE, 1966

TESTING A SIMPLE STRUCTURE HYPOTHESIS IN FACTOR ANALYSIS*

K. G. JÖRESKOG†

UNIVERSITY OF UPPSALA, SWEDEN

It is assumed that the investigator has set up a simple structure hypothesis in the sense that he has specified the zero loadings of the factor matrix. The maximum-likelihood method is used to estimate the factor matrix and the factor correlation matrix directly without the use of rotation methods, and the likelihood-ratio technique is used to test the simple structure hypothesis. Numerical examples are presented.

1. *Introduction and Summary*

We shall be concerned with factor analysis of the confirmatory type. It will be assumed that the experimenter has set up a particular simple structure hypothesis in the sense that he has specified loadings in the factor matrix that are to be zero or close to zero. In such a case the factor analysis may be regarded as a problem of statistical estimation and hypothesis testing. The simple structure factor matrix can be estimated directly from the observed correlations, avoiding completely the rotational step, and the hypothesis can be tested by a statistical test. This approach has been considered by Anderson and Rubin [2], Howe [6], and Lawley [10], all using the maximum-likelihood method and all arriving at essentially the same solution. These works have not been given much attention. In fact, as far as the writer knows, there has been no application of their method reported in the literature. The number of existing numerical illustrations of the method is two, viz., those given by Howe and by Lawley.

In this paper we shall consider this problem more thoroughly. We shall show that the numerical method for obtaining the maximum-likelihood solution involves considerable difficulties. The procedures proposed by Howe and Lawley either fail to converge or converge at an extremely slow rate. We shall discuss various other methods to obtain the solution. These methods have the advantage that they always converge. One of them is

*The work was supported by a grant (NSF-GB 1985) from the National Science Foundation to Educational Testing Service. Reproduction in whole or in part for any purpose of the United States Government is permitted.

†The work was carried out when the author was Visiting Research Statistician at Educational Testing Service. The author wishes to thank Dr. Frederic M. Lord for many helpful suggestions throughout the course of this study.

superior to all the others and is strongly recommended. This is a generalization of the method of resultant descents proposed by Finkel [5] as a numerical method for minimization of a function of many variables. This method appears to be very powerful for solving our problem. A computer program, based on the method, will be briefly described and some results, based on empirical and artificial data, will be presented.

2. *Factor Analysis Model and Simple Structure Hypothesis*

The fundamental equation of the factor analysis model is

$$y = \Lambda x + z, \tag{1}$$

where y is a vector of p components (the test scores), x is a vector of $k < p$ components (the common factor scores), z a vector of p components (the unique part of the test scores) and $\Lambda = (\lambda_{im})$ is a $p \times k$ matrix (of factor loadings). It is assumed that x and z are independent random vectors with $\mathcal{E}(x) = \mathcal{E}(z) = 0$, $\mathcal{E}(zz') = \Psi$, a diagonal matrix and $\mathcal{E}(xx') = \Phi$, a matrix having unities in the diagonal. From these assumptions one deduces that the dispersion matrix Σ of y is given by

$$\Sigma = \Lambda\Phi\Lambda' + \Psi. \tag{2}$$

The elements of Λ, Φ, and Ψ are parameters of the model. The main statistical problem of factor analysis is to estimate these parameters on the basis of an estimate S of Σ. If Λ, Φ, Ψ satisfy (2) for a certain value of k and T is any nonsingular $k \times k$ matrix, such that $T\Phi T'$ has unities in the diagonal, then also Λ^*, Φ^*, Ψ satisfy the equation for the same value of k, where

$$\Lambda^* = \Lambda T^{-1} \tag{3}$$

and

$$\Phi^* = T\Phi T'. \tag{4}$$

Though Λ^* and Λ are equivalent from the mathematical point of view, they may not be so from the psychological point of view. The problem of choosing one particular psychologically meaningful Λ out of the infinite set $\{\Lambda T^{-1}\}$ has been called the *problem of rotation*, although the *problem of transformation* would be a better term, since it includes also the transformation to oblique factors, in which case the transformation matrix T is not orthogonal and hence does not represent only a rotation. The problem of rotation has largely rested with the psychologists, perhaps rightly so. But once the psychological meaning is translated to mean some precisely stated restrictions on Λ, it is quite conceivable that its choice will turn out to be a problem of statistical estimation.

To resolve the problem of rotation Thurstone [14] proposed the concept of *simple structure* as a principle for finding psychologically meaningful factors. If $\lambda_{im} = 0$ the mth factor does not enter into the ith test. The general idea of simple structure is that if the factors have real psychological meaning, then many tests will not depend on all the factors. The factor matrix should have as many zero coefficients as possible. Such a matrix can then be considered as giving the simplest structure and presumably the one with most meaningful psychological interpretation. More precisely, ([14], p. 335) simple structure requires that the factor matrix Λ shall have the following properties. (1) Each row of Λ shall have at least one zero. (2) Each column of Λ shall have at least k zeros. (3) For every pair of columns of Λ there shall be several rows in which one loading is zero and one is nonzero. (4) For every pair of columns of Λ a large proportion of rows shall have two zero loadings (if $k \geq 4$). (5) For every pair of columns of Λ there shall preferably be only a small number of rows with two nonzero coefficients.

These conditions are not stated in a mathematically precise form. It is therefore extremely difficult to tell whether, for a given Σ, such a simple structure exists or not or, if it exists, to tell whether it is unique or not. Reiersøl [12] investigated these conditions and modified them so as to effect identification. He assumes that there are at least k zeros in each column of Λ. Let Λ_m $(m = 1, 2, \cdots, k)$ be the submatrix of Λ that has zero elements in the mth column. Reiersøl further assumes that (i) the rank of Λ_m is $k - 1$, (ii) the rank of each submatrix obtained by deleting a row of Λ_m is $k - 1$, (iii) the addition to Λ_m of any row of Λ not contained in Λ_m increases the rank to k. Then if Ψ is identified, a necessary and sufficient condition for the identification of Λ is that Λ does not contain any other submatrices satisfying (i), (ii), and (iii) ([12], Theorem 9.2.).

In this paper we shall use the term simple structure for any factor matrix having zeros in specified positions. We shall assume that the experimenter has set up a particular simple structure hypothesis in the sense that he has specified the zero positions in Λ. If the experimenter is not able to specify such a hypothesis, it is suggested that he divide the sample randomly into two halves and use one half to generate such a hypothesis and the other to test this hypothesis. We shall only assume that the following two conditions are satisfied: (i) each column of Λ has at least $k - 1$ zeros and (ii) the rank of Λ_m is $k - 1$, $m = 1, 2, \cdots, k$. It should be noted that there are k normalizations in Φ and a minimum of $k(k - 1)$ zero conditions. This is equal to the number of elements of the transformation matrix T. If, for a given Σ, the matrices Λ, Φ, Ψ satisfy (2), a nonsingular transformation matrix T exists such that $\Lambda^* = \Lambda T^{-1}$ has $k - 1$ zeros in each column and also satisfies the condition (ii) and such that $\Phi^* = T\Phi T'$ has unities in the diagonal. If there are more than $k - 1$ zeros specified in one or more columns of Λ, then there may be more conditions than are required to take out the indeterminacy in Λ. In this case the conditions may restrict Ψ.

3. *Maximum-Likelihood Solution*

In this section we shall consider the statistical problems of parameter estimation and hypothesis testing associated with the model presented in the previous section. We shall use the method of maximum likelihood for parameter estimation and the likelihood-ratio technique for testing the hypothesis. These methods have been previously applied to the same problem by Anderson and Rubin [2], Howe [6], and Lawley [10] (see also [11], ch. 6), all arriving at essentially the same solution.

We assume that we have drawn a random sample of N observations $y_1, y_2, \cdots, y_N$ from a multivariate normal population with dispersion matrix Σ. The mean vector of the sample is

$$\bar{y} = \frac{1}{N} \sum_{a=1}^{N} y_a ,$$

and the sample dispersion matrix is

$$S = \frac{1}{N} \sum_{a=1}^{N} (y_a - \bar{y})(y_a - \bar{y})' = \frac{1}{N} \sum_{a=1}^{N} y_a y_a' - \bar{y}\bar{y}'.$$

The distribution of S is the Wishart distribution (e.g., [1], ch. 7). The logarithm of the likelihood function for the sample, omitting a function of the observations, is given by

$$L = -\tfrac{1}{2}n[\log |\Sigma| + \operatorname{tr} (S\Sigma^{-1})], \tag{5}$$

where $n = N - 1$. We regard L as a function of the nonzero elements of Λ, the nondiagonal elements of Φ and the diagonal elements of Ψ, defined through the relation (2). We consider the function in the region where all the diagonal elements of Ψ are positive and where Λ, Φ, and Ψ are such that Σ is positive definite. The maximum-likelihood estimates $\hat{\Lambda}$, $\hat{\Phi}$, and $\hat{\Psi}$ are the values of Λ, Φ, and Ψ that maximize the likelihood function. It can be shown that the maximum-likelihood method is independent of the units of measurements in the test scores in the sense that if a test score is multiplied by a constant, the factor loadings for that test will be multiplied by the same constant and the unique variance will be multiplied by the square of the constant. We are thus free to use correlations as well as covariances.

Let us for the moment assume that we have obtained the maximum-likelihood estimates $\hat{\Lambda}$, $\hat{\Phi}$, and $\hat{\Psi}$. We then proceed to construct a statistical test of our hypothesis. The hypothesis is that Σ is of the form (2) with a specified k, with Λ having zeros in specified positions and with Φ having unities in the diagonal. The maximum-likelihood estimate of Σ under this hypothesis is

$$\hat{\Sigma} = \hat{\Lambda}\hat{\Phi}\hat{\Lambda} + \hat{\Psi}.$$

The alternative of the hypothesis is that Σ is any positive definite matrix. Let Ω be the set of all positive definite matrices of order $p \times p$ and let ω be the set of all Σ of the form (2). Let L_Ω be the maximum of L for Σ in Ω and let L_ω be the maximum of L for Σ in ω. Then (see e.g., [1], ch. 3)

$$L_\Omega = -\tfrac{1}{2}n(\log |S| + p)$$

and

$$L_\omega = -\tfrac{1}{2}n[\log |\hat{\Sigma}| + \operatorname{tr}(S\hat{\Sigma}^{-1})]$$

so that (-2) times the logarithm of the likelihood ratio becomes

$$n[\log |\hat{\Sigma}| - \log |S| + \operatorname{tr}(S\hat{\Sigma}^{-1}) - p].$$

This is the same expression as obtained by Lawley ([10], eq. 12). The exact distribution of the criterion is not known. If n is large, however, the distribution is approximately a χ^2 distribution with degrees of freedom equal to the number of restrictions on the free parameters imposed by the hypothesis. This is a general property of likelihood-ratio tests [15]. Box [3] has shown that it is possible to improve on this approximation simply by multiplying the criterion by a scale factor which results in a criterion having the same moments as χ^2 ignoring only quantities of order n^{-2}. This requires the calculation of the moments of the likelihood-ratio statistic, a formidable task which has not been undertaken. The criterion will therefore be taken as

$$U = n[\log |\hat{\Sigma}| - \log |S| + \operatorname{tr}(S\hat{\Sigma}^{-1}) - p]. \tag{6}$$

If m is the number of nonzero loadings in Λ to be estimated, the number of degrees of freedom is

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

To test our hypothesis we thus compute the value of U and reject the hypothesis at the significance level α if this value exceeds the $(1 - \alpha)$th percentile of the χ^2 distribution with d degrees of freedom.

If the hypothesis is rejected either the number of factors or the pattern of zeros is not tenable. The number of factors can be tested by computing the unrestricted maximum-likelihood solution and the corresponding likelihood-ratio statistic (e.g., [11], ch. 2). If such a test indicates that the number of factors is tenable it is possible to test the pattern of zeros against the alternative that $\Sigma - \Psi$ is positive semidefinite and of rank k. This provides a test of the simple structure hypothesis given that k common factors are required.

In the preceding paragraphs it was assumed that the maximum-likelihood estimates had been obtained. These estimates cannot, however, be obtained easily. The reason for this is twofold. Firstly, in a medium-sized factor analysis, with 40 tests and 10 factors, say, the number of parameters is almost 400. Our problem is then equivalent to maximizing a function of 400 variables. This is clearly not an easy problem. Secondly, the function is very complicated. If, for example, the function had been a quadratic, the problem would have been much easier.

It will be convenient to minimize the function

$$F = n[\log |\Sigma| - \log |S| + \operatorname{tr}(S\Sigma^{-1}) - p]$$

instead of maximizing the likelihood function. The function F is nonnegative and is zero only when S is exactly of the form (2). This never happens with real data. The minimum value of F is the value of the test criterion U. The partial derivatives of F are given in [10].

$$\frac{\partial F}{\partial \Lambda} = 2n[\Sigma^{-1}(\Sigma - S)\Sigma^{-1}]\Lambda\Phi,$$

$$\frac{\partial F}{\partial \Phi} = 2n\Lambda'[\Sigma^{-1}(\Sigma - S)\Sigma^{-1}]\Lambda,$$

$$\frac{\partial F}{\partial \Psi} = n \operatorname{diag} [\Sigma^{-1}(\Sigma - S)\Sigma^{-1}].$$

The elements of $\partial F/\partial \Lambda$ that correspond to zero elements of Λ and the diagonal elements of $\partial F/\partial \Phi$ are zero.

The usual way to find the maximum-likelihood estimates is to set all partial derivatives equal to zero and solve the corresponding equations. In this case, however, these equations cannot be solved algebraically. Lawley [10], using the following expression for the inverse of Σ,

$$\Sigma^{-1} = \Psi^{-1} - \Psi^{-1}\Lambda(\Phi^{-1} + \Lambda'\Psi^{-1}\Lambda)^{-1}\Lambda'\Psi^{-1},$$

transformed the likelihood equations to an equivalent set of equations for computational purposes. This set of equations is

$$\begin{aligned}\Phi\Lambda' &\doteq (\Phi^{-1} + \Lambda'\Psi^{-1}\Lambda)^{-1}\Lambda'\Psi^{-1}S,\\ \Phi &= (\Lambda'\Psi^{-1}\Lambda)^{-1}(\Lambda'\Psi^{-1}S\Psi^{-1}\Lambda)(\Lambda'\Psi^{-1}\Lambda)^{-1} - (\Lambda'\Psi^{-1}\Lambda)^{-1},\\ \Psi &= \operatorname{diag}(S - \Lambda\Phi\Lambda'),\end{aligned}$$

where $\doteq$ in the first equation means that the corresponding elements are to be equated only when they correspond in position to nonzero elements of Λ. Lawley suggested an iterative procedure to solve these equations. This procedure involves (i) finding initial estimates Λ_1, Φ_1, and Ψ_1, (ii) using these estimates in the right-hand side of the equations to determine new estimates Λ_2, Φ_2, and Ψ_2, and (iii) repeating this until convergence is obtained. Howe [6] suggested a slightly different procedure with the object of increasing the rate of convergence. It is extremely difficult to establish exact conditions under which such iterative procedures converge. All we can do is to hope that the procedures converge if the initial estimates are close enough. In the last section of this paper we shall give an example where Lawley's procedure actually diverged. In cases when convergence occurs it is usually very slow. It is quite possible to stop at a point in the iteration where there is no change in the value at the decimal place to which accuracy is desired, and yet the correct estimate is a long way off.

Because of these difficulties we have investigated the possibilities of solving our problem by other kinds of methods. Efficient numerical methods for minimizing a function of several variables have been devised by numerical analysts. For a review of such methods we refer to Spang [13]. We shall describe three methods here.

To simplify the description we assume that the arguments of the function F have been arranged as a vector x of q components $x_1, x_2, \cdots, x_q$, where

$$q = m + \tfrac{1}{2}k(k - 1) + p,$$

and m, as before, is the number of nonzero loadings in Λ. We also assume that we have found initial estimates Λ_1, Φ_1, and Ψ_1, represented by the vector $x^{(1)}$. These estimates do not affect the final solution. They may, however, affect the number of iterations required to reach the solution. The closer the initial estimates are to the final solution the more rapidly will the procedure converge. In all methods to be described we shall construct a sequence of points $x^{(1)}, x^{(2)}, \cdots$ in q-dimensional space such that

$$F(x^{(i+1)}) < F(x^{(i)}).$$

The process will continue until no further decrease of the function is possible.

The first method is *the method of steepest descent* (see, e.g., [13]). In this method we determine $x^{(i+1)}$ by moving from $x^{(i)}$ along the direction in which the function is decreasing most rapidly. This direction is called the (negative) gradient and is defined by the vector

$$g^{(i)} = -\left[\frac{\partial F}{\partial x}\right]_{x=x^{(i)}} .$$

We thus determine the new point $x^{(i+1)}$ as

$$x^{(i+1)} = x^{(i)} + \alpha^{(i)} g^{(i)}, \tag{7}$$

where $\alpha^{(i)}$ is a suitable positive constant. If the gradient vector $g^{(i)}$ is not zero and if the step-size $\alpha^{(i)}$ is sufficiently small, the step will always result in a decrease of the function. The choice of the step-size $\alpha^{(i)}$ is of some importance for the rate of convergence. The optimal choice of $\alpha^{(i)}$ is such that the function F is minimized on the line

$$y = x^{(i)} + \gamma g^{(i)}, \qquad \gamma > 0.$$

The optimal value can be approximated by various interpolation procedures. Davidon [4] has suggested finding a trial point $z^{(i)}$ on the line and then interpolating cubically using the values of F and the slopes at $x^{(i)}$ and $z^{(i)}$. This involves the computation of the function and the gradient at two points in each iteration, but this is usually worthwhile. We have used

Davidon's interpolation technique not only in this but also in the other two methods which will be described and have found that it works quite satisfactorily.

After arriving at the new point $x^{(i+1)}$ we again determine the gradient and move along it. After many repetitions we arrive at a point from which no improvement is possible, in other words, at the minimum. We know when this happens, for at the minimum all the partial derivatives are zero, i.e., the gradient vector is zero.

The method of steepest descent has often very slow convergence. This is so because the path to the minimum is long and irregular. The step taken in each iteration is often very small and the gradient is changed considerably from one iteration to the next. It is possible to accelerate the convergence by, in every third iteration, moving along the direction which is the resultant of the two preceding gradients. We thus determine $x^{(i+1)}$ and $x^{(i+2)}$ as in the method of steepest descent but instead of determining $x^{(i+3)}$ from $x^{(i+2)}$ by the formula (7) we determine $x^{(i+3)}$ as

$$x^{(i+3)} = x^{(i+2)} + \alpha^{(i+2)}[x^{(i+2)} - x^{(i)}]. \tag{8}$$

This is *the method of resultant descents* proposed by Finkel [5]. If F were a quadratic function in two variables, the resultant direction would pass through the minimum and the function would thus be minimized in three iterations. For nonquadratic functions further iterations would be necessary. Finkel has reported that the method is fairly successful even for general functions of q variables. We have found that the number of iterations is often reduced by a factor five as compared to the method of steepest descent.

The type of cycle represented by (8) will be called a cycle of order 1. The procedure is easily generalized to cycles of higher order. Figs. 1 and 2 show examples of cycles of order 2 and 3. In these figures all directions except those which are continuations of broken lines are taken as the gradient direction. Each new point is taken so as to minimize the function along the line. This guarantees that the function is decreased in each step.

In a cycle of order 1 most of the computation time is devoted to the computation of the function and the gradient at the points $x^{(i)}$ and $x^{(i+1)}$.

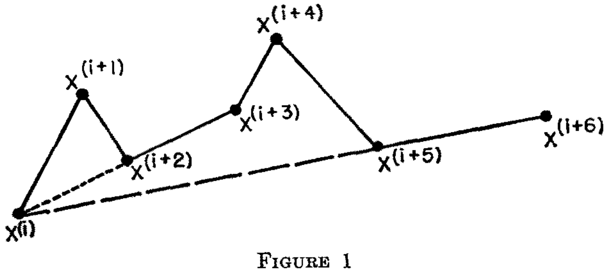

FIGURE 1
A Cycle of Order 2

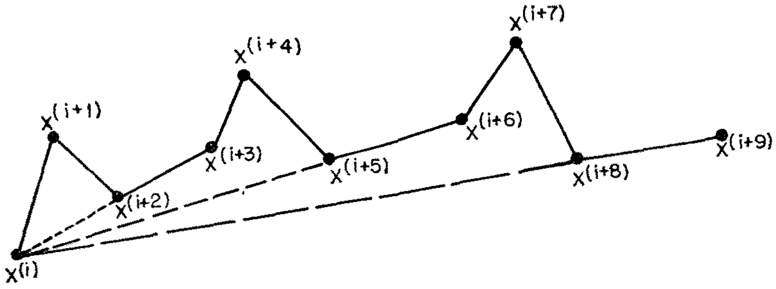

FIGURE 2
A Cycle of Order 3

The time to compute the resultant direction is small compared to this time. It is therefore clear that the time for a cycle of order 2 is approximately two times larger than for a cycle of order 1, the time for a cycle of order 3 is three times larger than for a cycle of order 1, etc. We have experimented with cycles of orders 1, 2, 3, and 4 and have found that high-order cycles are better far away from the minimum, whereas low-order cycles are better in the neighborhood of the minimum.

4. *Numerical Illustrations*

The author has constructed a computer program for the IBM 7094. This program performs the computations involved in the generalized method of resultant descents as described briefly in the preceding section.

The input to the program consists of the correlation matrix S and a specification of the zero positions of the hypothesized factor matrix. The initial estimates Λ_1 and Φ_1, necessary to start the iterative procedure, may either be determined by the program or specified by the user. In the former case these initial estimates are determined by an oblique rotation of an unrotated orthogonal factor matrix. The method of rotation is the least-squares procedure described in [9]. The unrotated factor matrix, necessary for this rotation, may either be determined by the program or specified by the user. In the former case the unrotated factor matrix is computed by the method of Jöreskog ([7], pp. 43–48). Thus there is great flexibility in the program in that data may be analyzed making use of various results from previous analyses.

The program is written in FORTRAN IV. The FORTRAN list and a brief description [8] of the program is available from the author.

The author has also programmed Lawley's method ([11], pp. 82–83) and the method of steepest descent. These methods cannot be recommended, however, since they are usually much slower than the generalized method of resultant descents.

In this section we shall give some results obtained by these computer programs. To illustrate the methods we shall use the same data as used by Lawley [10] and Howe [6]. Lawley's data consists of eight tests with the following correlation matrix:

$$\begin{bmatrix} 1. & & & & & & & \\ 0.691 & 1. & & & & & & \\ 0.679 & 0.791 & 1. & & & & & \\ 0.149 & 0.285 & 0.314 & 1. & & & & \\ 0.409 & 0.505 & 0.472 & 0.470 & 1. & & & \\ 0.382 & 0.443 & 0.372 & 0.385 & 0.680 & 1. & & \\ 0.346 & 0.471 & 0.426 & 0.434 & 0.639 & 0.576 & 1. & \\ 0.270 & 0.355 & 0.254 & 0.218 & 0.504 & 0.452 & 0.395 & 1. \end{bmatrix}.$$

The simple structure hypothesis is represented by the following pattern of zero and nonzero loadings:

$$\begin{bmatrix} \times & \times & 0 \\ \times & \times & \times \\ 0 & \times & \times \\ 0 & 0 & \times \\ \times & 0 & \times \\ \times & 0 & \times \\ \times & 0 & \times \\ \times & 0 & 0 \end{bmatrix}.$$

Using only this information the program determines the following initial estimates.

$$\Lambda_1 = \begin{bmatrix} 0.229 & 0.694 & 0. \\ 0.167 & 0.717 & 0.214 \\ 0. & 0.758 & 0.306 \\ 0. & 0. & 0.682 \\ 0.554 & 0. & 0.418 \\ 0.626 & 0. & 0.213 \\ 0.381 & 0. & 0.485 \\ 0.656 & 0. & 0. \end{bmatrix}; \quad \Phi_1 = \begin{bmatrix} 1. & & \\ 0.368 & 1. & \\ 0.485 & 0.238 & 1. \end{bmatrix}; \quad \Psi_1 = \begin{bmatrix} 0.350 \\ 0.215 \\ 0.222 \\ 0.535 \\ 0.293 \\ 0.356 \\ 0.440 \\ 0.569 \end{bmatrix}.$$

To save space we have here written Ψ_1 as a column vector though it is a diagonal matrix. The value of the function F at this initial point is

$$F_1 = 10.275.$$

Using these initial values we have performed 180 iterations with Lawley's procedure. The behavior of the function in these iterations was as follows.

The procedure diverged. In the first 79 iterations the function was decreased in a regular cyclical manner but then it started to increase rapidly. In the 180th iteration one of the elements of Ψ became negative. We had therefore come outside the region of definition for the function F.

The important question now arises whether actually the absolute minimum was reached at the 79th iteration. The answer is no. The results of the 79th iteration are

$$\Lambda_{79} = \begin{bmatrix} 0.286 & 0.658 & 0. \\ 0.232 & 0.699 & 0.203 \\ 0. & 0.780 & 0.367 \\ 0. & 0. & 0.677 \\ 0.616 & 0. & 0.375 \\ 0.655 & 0. & 0.214 \\ 0.456 & 0. & 0.404 \\ 0.607 & 0. & 0. \end{bmatrix}, \quad \Phi_{79} = \begin{bmatrix} 1. & & \\ 0.331 & 1. & \\ 0.526 & 0.169 & 1. \end{bmatrix}, \quad \Psi_{79} = \begin{bmatrix} 0.361 \\ 0.210 \\ 0.160 \\ 0.541 \\ 0.237 \\ 0.378 \\ 0.434 \\ 0.631 \end{bmatrix},$$

and the value of F at this point is

$$F_{79} = 5.468.$$

The maximum-likelihood estimates (correct to three decimals), obtained by the generalized method of resultant descents, are

$$\hat{\Lambda} = \begin{bmatrix} 0.353 & 0.644 & 0. \\ 0.268 & 0.676 & 0.241 \\ 0. & 0.769 & 0.451 \\ 0. & 0. & 0.651 \\ 0.601 & 0. & 0.369 \\ 0.662 & 0. & 0.187 \\ 0.440 & 0. & 0.400 \\ 0.600 & 0. & 0. \end{bmatrix}, \quad \hat{\Phi} = \begin{bmatrix} 1. & & \\ 0.237 & 1. & \\ 0.588 & 0.088 & 1. \end{bmatrix}, \quad \hat{\Psi} = \begin{bmatrix} 0.352 \\ 0.224 \\ 0.144 \\ 0.576 \\ 0.242 \\ 0.380 \\ 0.439 \\ 0.640 \end{bmatrix},$$

and the minimum value of F is

$$F_{\min} = 5.202.$$

Inspecting the differences between Λ_{79} and $\hat{\Lambda}$ we see that the largest difference is 0.084 for λ_{33} . For the factor correlations the largest difference is 0.094 for ϕ_{21} . Thus, with Lawley's procedure, the best possible result is not correct to one decimal. This is not considered satisfactory.

We proceed to report the analysis of the same data with the method of steepest descent and the generalized method of resultant descents. As mentioned already in the introduction, these methods converge by theory. The results can be obtained as accurately as desired. We have chosen to terminate the procedures when all the partial derivatives are less than a prescribed quantity ϵ. We have used several ϵ and have found that when $\epsilon = 0.0001$ the result is usually correct to three decimals.

We have performed 300 iterations with the method of steepest descent. This procedure is extremely slow. After 300 iterations the value of F was 5.213 (c.f., $F_{\min}$ above) and the partial derivatives were still of order 10^{-2}. This is a surprising result since the method of steepest descent is generally regarded as an efficient method for minimization. We conclude that the method of steepest descent cannot be used for our purposes.

TABLE 1

Number of Cycles Required To Reach the Solution with the Generalized Method of Resultant Descents

	Lawley's data		Howe's data	
Cycles of order	Number of cycles	Number of steepest descents	Number of cycles	Number of steepest descents
1	75	150	5	10
2	21	84	3	12
3	9	54	3	18
4	12	96	3	24

With the generalized method of resultant descents we have used cycles of order 1, 2, 3, and 4. The number of cycles required to reach the solution with $\epsilon = 0.0001$ is given in Table 1, column 2. Since a cycle of order 2 is equivalent to two cycles of order 1 etc., we have, for the sake of comparison, given in column 3 the number of times a step was taken in the direction of the steepest descents. We thus see that the method was most effective when we used cycles of order 3.

In his report Howe [6] has used a set of artificial data representing five tests and two factors. The following correlation matrix was used

$$\begin{bmatrix} 1. & & & & \\ 0.43 & 1. & & & \\ 0.50 & 0.56 & 1. & & \\ 0.35 & 0.40 & 0.44 & 1. & \\ 0.30 & 0.37 & 0.41 & 0.58 & 1. \end{bmatrix}$$

It was postulated that the first three tests should be loaded only in factor 1 and the last two tests only in factor 2. This constitutes an ideal simple structure and the method of Howe is restricted to such cases. Our method is not, however, restricted to such pure patterns only.

Using the initial estimates

$$\Lambda_1 = \begin{bmatrix} 0.6 & 0. \\ 0.7 & 0. \\ 0.8 & 0. \\ 0. & 0.8 \\ 0. & 0.7 \end{bmatrix}, \qquad \Phi_1 = \begin{bmatrix} 1. & \\ 0.0 & 1. \end{bmatrix}, \qquad \Psi_1 = \begin{bmatrix} 0.64 \\ 0.51 \\ 0.36 \\ 0.36 \\ 0.51 \end{bmatrix},$$

the following maximum-likelihood estimates were obtained

$$\hat{\Lambda} = \begin{bmatrix} 0.619 & 0. \\ 0.703 & 0. \\ 0.799 & 0. \\ 0. & 0.796 \\ 0. & 0.729 \end{bmatrix}, \qquad \hat{\Psi} = \begin{bmatrix} 1. & \\ 0.702 & 1. \end{bmatrix}, \qquad \hat{\Psi} = \begin{bmatrix} 0.617 \\ 0.505 \\ 0.362 \\ 0.366 \\ 0.469 \end{bmatrix}.$$

These results agree with those obtained by Howe. The number of cycles required to reach the solution is given in Table 1, column 4, and the number of steepest descents taken is given in column 5. In this case the method is most effective with cycles of order 1. This is probably due to the fact that the initial estimates were very good in this case.

We have deliberately chosen to illustrate the methods with small sets of data. It should be pointed out, however, that the method is quite feasible even with larger sets of data. We have also analyzed several other sets of real data. The time taken to analyze Lawley's data was approximately 10 seconds. Data with 17 tests and six factors took approximately one minute. Another set of data with 33 tests and 10 factors took approximately six minutes. This gives a rough idea of how fast the computation time increases with the number of tests and the number of factors.

REFERENCES

[1] Anderson, T. W. *An introduction to multivariate statistical analysis.* New York: Wiley, 1958.

[2] Anderson, T. W. and Rubin, H. Statistical inference in factor analysis. In J. Neyman (Ed.), *Proceedings of the third Berkeley symposium on mathematical statistics and probability.* Vol. V. Berkeley: Univ. Calif. Press, 1956, 111–150.

[3] Box, G. E. P. A general distribution theory for a class of likelihood criteria. *Biometrika,* 1949, **36**, 317–346.

[4] Davidon, W. C. Variable metric method for minimization. A.E.C. Research and Development Report, ANL 5990, Argonne National Laboratory, 1959.

[5] Finkel, R. W. The method of resultant descents for the minimization of an arbitrary function. Paper 71, Preprints of Papers Presented at 14th National Meeting of Association of Computing Machinery, 1959.

[6] Howe, W. G. Some contributions to factor analysis. Report No. ORNL-1919. Oak Ridge National Laboratory, Oak Ridge, Tenn., 1955.

[7] Jöreskog, K. G. *Statistical estimation in factor analysis.* Stockholm: Almqvist & Wiksell, 1963.

[8] Jöreskog, K. G. Computer program for estimating and testing a simple structure hypothesis in factor analysis. Res. Memo. 65-3. Princeton, N. J.: Educ. Test. Serv., 1965.

[9] Jöreskog, K. G. On rotation to a specified simple structure. Res. Bul. 65-13. Princeton, N. J.: Educ. Test. Serv., 1965.

[10] Lawley, D. N. Estimation in factor analysis under various initial assumptions. *Brit. J. statist. Psychol.,* 1958, **11**, 1–12.

[11] Lawley, D. N. and Maxwell, A. E. *Factor analysis as a statistical method.* London: Butterworths, 1963.

[12] Reiersøl, O. On the identifiability of parameters in Thurstone's multiple factor analysis. *Psychometrika,* 1950, **15**, 121–149.

[13] Spang, H. A. III. A review of minimization techniques for nonlinear functions. *SIAM Rev.,* 1962, **4**, 343–365.

[14] Thurstone, L. L. *Multiple-factor analysis.* Chicago: Univ. Chicago Press, 1947.

[15] Wilks, S. S. The large-sample distribution of the likelihood ratio for testing composite hypotheses. *Ann. math. Statist.,* 1938, **9**, 60–62.

Manuscript received 5/4/65
Revised manuscript received 9/14/65