PSYCHOMETRIKA—VOL. 31, NO. 3
SEPTEMBER, 1966

# FACTOR ANALYSIS BY MINIMIZING RESIDUALS (MINRES)*

HARRY H. HARMAN AND WAYNE H. JONES†

SYSTEM DEVELOPMENT CORPORATION

This paper is addressed to the classical problem of estimating factor loadings under the condition that the sum of squares of off-diagonal residuals be minimized. Communalities consistent with this criterion are produced as a by-product. The experimental work included several alternative algorithms before a highly efficient method was developed. The final procedure is illustrated with a numerical example. Some relationships of minres to principal-factor analysis and maximum-likelihood factor estimates are discussed, and several unresolved problems are pointed out.

The research work reported in this paper was started in early 1964 and was initially directed "toward an objective definition of communality." It has led to much more general results, including a basic new method of factor analysis that bears a strong resemblance to a maximum-likelihood solution.

For a clear understanding of the point of departure of the present paper, we distinguish two objectives as the basis for estimating certain parameters in factor analysis, namely:

1. to extract the maximum variance;
2. to "best" reproduce the original correlations.

The first of these readily leads to the principal components of a correlation matrix (arrived at by different routes, some of which are mentioned below), while the second goes back to Thurstone's quest for "best fit" to the off-diagonal elements of a matrix: "The object of a factor problem is to account for the tests, or their intercorrelations, in terms of a small number of derived variables, the smallest possible number that is consistent with acceptable residual errors" [17, p. 61]. In this paper we solve the factor analysis problem as posed by Thurstone by maximally (in the least-squares sense) reproducing the off-diagonal elements of the correlation matrix, and, as a by-product, obtaining communalities consistent with this criterion. The method is designated "minres" for "minimum residuals."

## 1. *Brief History*

Conceptually, the idea of getting a factor solution by minimizing the residual correlations is an obviously direct approach; the wonder is that it

*The authors wish to thank the Factor Analysis Work Group (supported, in part, by ONR) for valuable criticisms and suggestions made in the course of a discussion of the present work in April, 1965.

†Now with the Department of Defense.

has not been done before. The idea certainly is not new—its accomplishment, however, was dependent on the high-speed computer. No doubt the idea must have crossed the minds of many workers in factor analysis over the last three decades. The first theoretical treatment appeared in 1936, when Eckart and Young noted that "if the least-squares criterion of approximation [of one matrix by another of lower rank] be adopted, this problem has a general solution which is relatively simple in a theoretical sense, though the amount of numerical work involved in applications may be prohibitive" ([4], p. 211). This was followed in the next couple of years by additional theoretical work by Householder and Young [9] and Horst [8].

More recently, several papers have appeared that seem to bear some relationship to the present work. Whittle [20] specifically considered the residual sum of squares, but in relation to the principal-component solution. Howe, in his doctoral dissertation, seeks an alternative approach to Lawley's maximum-likelihood equations and finds that his method—maximizing the determinant of partial correlations—is approximately equivalent to minimizing the sum of squares of the partial correlations" ([11], p. 22). Even more germane is the 1962 paper by Keller [13], which is a generalized mathematical treatment skirting the precise problem to which our minres solution is addressed, but which, in a brief concluding section, actually suggests the form of solution contained in our method I (see Section 4).

It should be noted that none (with the possible exception of Keller [13]) of the foregoing papers, found after diligent search, considers the minimization of off-diagonal residuals—the minimization of the total residual matrix (including diagonal terms) leads to the conventional principal-factor solution. The exclusion of the diagonal elements, although appearing trivial, is of paramount importance. More specifically, as will be amplified below, the diagonal elements of the sample correlation matrix are not fixed but are parameters to be determined along with the factor loadings.

Probably the first attempt to obtain a practical factor solution by minimizing off-diagonal residuals was suggested by Thurstone in 1954 and carried out by Rolf Bargmann and also by Sten Henrysson (see [18], p. 61). More recently, Comrey [3] independently developed a computing procedure for such a solution. However, these investigators do not tackle the complete problem of determining a factor solution with the property that the sum of squares of residuals between observed and reproduced correlations is a minimum. Instead they consider what might be termed a "stepwise" minimum residual method, obtaining one factor and a residual matrix, which is then the starting point for the factor in the next step; this process is continued until a desired number of factors are extracted. In general, of course, such a solution is different from one obtained under the least-squares criterion for the entire set of factors. It may be of interest to note similar "stepwise" approximations to standard statistical procedures, namely: the determination

of the coefficients in multiple regression *successively* rather than *simultaneously* [6]; and an approximation to a maximum-likelihood factor solution [1].

Finally, work being done by Boldt [2] certainly is related to the present paper. He poses the problem in essentially the same form as is done here (see Section 2) and considers solutions by procedures that are very close to our Method I and the basic gradient method at the beginning of our Method II (see Section 4). While it is interesting to note some similarities in the independent attacks on the problem, the present paper only refers to those methods as background for the development of the more efficient techniques.

Interestingly enough, in the course of discussion of the Thurstone paper, John W. Tukey raised the question: "Would we not be better off, in principle, . . . to accept this formula and minimize with respect to [all the factors postulated] at the same time?" ([18], p. 65). To which Thurstone replied that the only reason he did not consider doing so was because it would get too unwieldy to handle. It is the problem suggested by Tukey that is the subject of the present paper.

## 2. *The Problem*

The basic problem with which factor analysis is concerned is the resolution of $n$ variables $z_j$ linearly in terms of some smaller number $m$ of common factors, with error or unique factors introduced to account for any unexplained variance. This may be expressed algebraically by

$$z_j = \sum_{p=1}^{m} a_{jp}F_p + d_jU_j \qquad (j = 1, 2, \cdots, n) \tag{1}$$

or, in matrix form,

$$Z = AF + DU, \tag{1'}$$

in which only the common-factor loadings $a_{jp}$ (or the matrix $A$) are the parameters to be estimated. It is important to distinguish model (1) from the alternative model

$$z_{ji} = \sum_{p=1}^{m} a_{jp}F_{pi} + d_jU_{ji} \qquad (j = 1, \cdots, n; i = 1, \cdots, N) \tag{2}$$

in which the factor values $F_{pi}$ as well as the factor loadings $a_{jp}$ are treated as parameters to be estimated. In addition, varying assumptions may be made about the nature of the residuals and the distribution function of the common factors (see Lawley [14]).

Our concern in this paper is exclusively with model (1), but the alternative (2) is noted because that is the model employed by Eckart and Young [4] and Whittle [20], whose work otherwise is in the spirit of this paper. Starting with model (2), their least-squares fit to the "score" matrix leads to the $m$ highest eigenvalues and corresponding eigenvectors of the correlation matrix.

This conclusion was stated by Eckart and Young: "Thus, if the best approximation to the score matrix is found, then the correlational matrices calculated from it will automatically be the best approximations to the correlational matrices calculated from the orginal score matrix" ([4], p. 216). Whittle also concludes: ". . . the solution turns out to be nothing else than an analysis of the covariance matrix $XX'$ into principal components . . ." ([20], p. 226). These methods, as well as the basic work of Hotelling [10], maximize explained variance (or minimize uniqueness) of the variables in standard measure. They do not maximally reproduce the original correlations, which is the aim of the present paper.

Once a solution (1) is obtained, the fundamental theorem of factor analysis gives (assuming, without any loss in generality, that the factors are uncorrelated):

$$R^{\dagger} = AA', \tag{3}$$

where $R^{\dagger}$ is a matrix of reproduced correlations with communalities in the principal diagonal. A solution to the problem, then, is to get a "best" fit to the observed correlation matrix $R$ by the reproduced correlations employing model (1). A least-squares fit can be obtained in either of two senses:

$$\text{fitting } R \text{ by } (R^{\dagger} + D^2) \tag{4}$$

or

$$\text{fitting } (R - I) \text{ by } (R^{\dagger} - H), \tag{5}$$

where $H = I - D^2 = \text{diag}\,(AA')$ is the diagonal matrix of reproduced communalities. In the case of (4), the minimization of residuals of the total matrix leads to the principal-component solution. In the case of (5), however, minimizing only the off-diagonal residuals leads to the minres solution. The condition for the minres solution may be expressed more precisely by:

$$\min_{A} ||[R - I] - [AA' - \text{diag}\,(AA')]|| \tag{6}$$

in which it is emphasized that both $A$ and $H$ vary. The norm as expressed in (6) may be written out algebraically as

$$f = \sum_{j=1}^{n-1} \sum_{k=j+1}^{n} \left( r_{jk} - \sum_{p=1}^{m} a_{jp} a_{kp} \right)^2 = \min. \tag{7}$$

To call attention to the fact that this is a function of the $n(n - 1)/2$ off-diagonal residual correlations, which are dependent upon the elements in the factor matrix $A$, it may be designated $f(A)$. The objective of minres is to minimize the function $f(A)$, for a specified $m$, by varying the values of the factor loadings. The diagonal matrix of communalities,

$$H = (h_i^2) = \left(\sum_{p=1}^{m} a_{ip}^2\right),$$

is obtained as a by-product of the method.

### 3. *Canonical Form*

Before the actual discussion of alternative computing procedures for the solution of the problem, a general observation should be made. It is a well-known fact that for a given correlation matrix a factor solution usually yields a unique common-factor space but not a unique set of factor loadings (an exception is a principal-factor solution). Because of the arbitrariness of representation of a factor solution in a fixed space, it is desirable to select a "canonical" form. Of course, rotation to such form has nothing to do with the "rotation problem" in factor analysis to attain "psychological meaningfulness" (say, in Thurstone's "simple structure" sense). Rotation to canonical form is merely a suggestion to bring an arbitrary solution to a well-defined form in a mathematical sense. Among other values, it may be useful in resolving the question of the meaning of equivalence of two solutions—they may look different, but if they are truly equivalent then, when each is brought to canonical form, they will be identical.

The canonical form adopted in this paper has the property that successive factors account for maximum possible variance. It is arrived at in the following manner. Let

$A$ = arbitrary form of factor matrix ($n \times m$) produced by criterion (7),
$B$ = canonical form of factor matrix ($n \times m$),
$T$ = orthogonal transformation matrix ($m \times m$);

then

$$B = AT \tag{8}$$

will yield the desired form of the factor solution, and the immediate problem is to determine the matrix $T$. Premultiplying (8) by the transpose of $B$ and then pre- and postmultiplying by $T$ and $T'$, respectively, produces the final result

$$A'A = TLT', \tag{9}$$

where $L = B'B$. It is apparent that $L$ is the diagonal matrix of eigenvalues and $T$ the matrix of corresponding eigenvectors of the matrix $A'A$. Thus, it is only necessary to determine the eigenvectors of an $m \times m$ matrix to obtain the transformation matrix which carries the arbitrary pattern matrix $A$ into the canonical form $B$.

Of course, the matrix of residuals remains unchanged whether computed from $A$ or $B$ since

$$BB' = ATT'A' = AA',$$

and the objective function $f$ has the same value whether computed from $A$ or $B$.

## 4. *Experimental Solutions*

Strangely enough, there are a variety of approaches to the problem, each yielding an appropriate solution but with varying degrees of efficiency. In order to save the researcher the unnecessary labor of retracing some of the less efficient methods, we shall briefly review several techniques we tried before discussing in greater detail the recommended procedure. The experimental work on minres may be grouped into the following categories.

Method I. Principal-factor Iterations
Method II. Gradient Methods
Parallel tangents (partan)
Higher-order cycles
Method III. Successive Approximations of Factor Loadings (Gauss-Seidel)

### *Method I. Principal-factor Iterations*

All the early experimental work (actually three versions) are included under Method I. The basic procedure starts with the choice of the number of factors $m$ and the selection of an arbitrary set of communality estimates for the $n$ variables. A factor matrix is determined from the $m$ largest eigenvalues and associated eigenvectors of the observed correlation matrix with these communalities. Then, the communalities reproduced from this factor matrix are employed as the new set of trial values for the calculation of the next factor matrix. This process is continued until acceptable stability is attained. The algorithm for this method may be expressed in the following form.

Step 1. Start with the observed correlation matrix with arbitrary diagonal: $(R - I + H)$.

Step 2. Compute: $(R - I + H) = Q\Lambda Q'$, where $\Lambda$ is the diagonal matrix of the eigenvalues and the columns of $Q$ are the associated eigenvectors.

Step 3. Determine the first $m$ principal factors: $A = Q_m \Lambda_m^{\frac{1}{2}}$ where $\Lambda_m$ is the $m \times m$ submatrix of $\Lambda$ containing the $m$ largest eigenvalues, and $Q_m$ is the corresponding $n \times m$ submatrix of $Q$. (10)

Step 4. Determine the reproduced communalities: $H = \text{diag}\,(AA')$.

Step 5. Repeat Steps 2–4 until the convergence criterion is met. (The criterion used was $\Delta f_i + 10^{-3} \geqq 0$ where $\Delta f_i = f_i - f_{i-1}$ for iteration $i$.)

Step 6. Convert the arbitrary solution to canonical form by transformation (8), with $T$ determined from (9).

Essentially, what we have grouped under Method I are procedures in which an arbitrary diagonal matrix $H$ of communalities is selected, and by successive calculations of a principal-factor matrix $A$ and its associated communalities $H$, improvements in $f$ are obtained. While this process works, it has been found to be rather slow, and thus we were led to explore alternative procedures.

*Method II. Gradient Methods*

The problem of finding a minimum for the function $f$ might be considered from a purely mathematical viewpoint. A minimum value for a nonlinear function $f(A)$, as in (7), is well defined: It occurs at that point where its partial derivatives with respect to the $nm$ independent variables are zero and its matrix of second derivatives is positive definite. The practical determination of a minimum is another matter. Even with modern-day computers, the solution of $nm$ nonlinear equations can prove to be a horrendous task. Iterative procedures provide the answer to this problem by means of a class of mathematical techniques known as "gradient methods" (see [16]). The general approach is described immediately following. Then two modifications to accelerate convergence are introduced.

An optimal value (maximum or minimum) of a function is obtained, iteratively, by proceeding from a trial solution to the next approximation in the direction of maximal change in the function. That direction can best be described by use of the concept of the gradient, defined by the matrix

$$G = (g_{jp}) \qquad (j = 1, 2, \cdots, n; p = 1, 2, \cdots, m) \tag{11}$$

whose elements are the partial derivatives of the function, namely,

$$g_{jp} = \frac{\partial f(A)}{\partial a_{jp}}. \tag{12}$$

Now, the direction of the $nm$-dimensional gradient vector $G$ is that in which the derivative of the function has its maximum. Hence, by taking the "normal derivative," $df/dn$, which is in the direction of the normal at the point where the derivative is taken, the maximum rate of increase of $f$ is obtained. The direction $G$ is perpendicular to the level surface through the point $A$ (the surface being $f(A) = \text{constant}$). The gradient method, or the method of steepest descent, seeks the minimum value of $f$ by moving along the negative gradient (but not necessarily to the minimum on the line).

The general gradient procedures yield successive points by means of the iteration equation,

$$A_{i+1} = A_i + tG_i, \qquad (i = \text{iteration number}) \tag{13}$$

which represents parametric equations of a line in matrix form. Starting at an arbitrary point, determining the gradient at that point and moving along it,

yields the next point. After arriving at a new point, this process is repeated again and again, until a point is reached at which no improvement is possible. That, then, is the minimum, at which the gradient vector is zero. Unfortunately, the gradient method may converge very slowly.

As a first step toward hastening the convergence, a procedure known as "optimum gradient" [16] was tried. This name comes from the choice of iteration step size. In seeking a minimum for $f(A)$, the size of the move in iteration $(i + 1)$ is chosen to minimize $f(A_i + tG_i)$ over values of $t$. This requires minimizing a fourth degree polynomial in $t$ for the case of the function defined in (7). The optimum gradient method, while showing some improvement over Method I, still was disappointing. Therefore, the search for better methods was continued.

*Parallel tangents (partan).* The next improvement was suggested in an article by Shah, Buehler, and Kempthorne [15] as the method of "parallel tangents" or "partan." This is a variant of the gradient method in which the direction followed is alternatively that of the gradient and that determined from points already reached. To grasp an understanding of this method, the following heuristic approach beginning with the determination of an optimal point on a two-dimensional surface may be useful.

(1) The contours of a positive-definite quadratic function of two variables are concentric and similar ellipses.
(2) Corresponding points of these ellipses lie on lines through their common center (in this case, the location of the minimum).
(3) Contour tangents at corresponding points are parallel.
(4) The minimum of the function along any line is at the point of tangency of the line to a contour.

Thus, minimizing the function on any two parallel lines furnishes two points which are collinear with the location of the over-all minimum.

This process is generalized to $n$ variables by showing that the minimum lies on a line parallel to $n - 1$ tangent hyperplanes. A proof of the convergence of the process to a minimum within $2n$ steps for a quadratic function of $n$ variables is also presented ([15], pp. 75–79). For other than quadratic functions there is no assurance of convergence in a finite number of steps, but we have found partan to be superior to simple gradient methods.

The essence of the algorithm for this method consists simply of alternating gradient and partan approximations, as follows.

Step 1. Start with arbitrary $A_0$.
Step 2. Calculate: $A_2 = A_0 + tG_0$ (to simplify the notation, $A_1$ is not used).
Step 3. Gradient: $A_{2i+1} = A_{2i} + tG_{2i}$ $\quad (i = 1, 2, 3, \cdots)$.

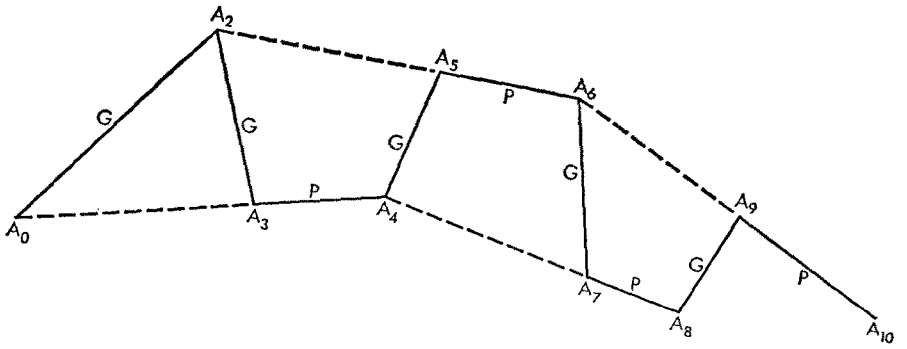



FIGURE 1

Schematic Representation of Partan Procedure

Step 4. Partan: $A_{2i+2} = A_{2i-2} + t(A_{2i+1} - A_{2i-2})$ $(i = 1, 2, 3, \cdots)$.

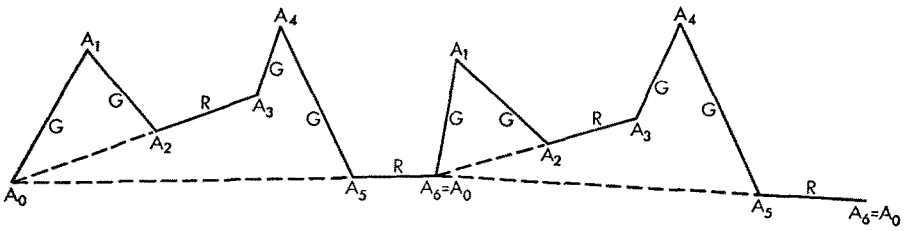
(14) Step 5. Repeat Steps 3 and 4, obtaining successive odd-numbered and even-numbered points, until the convergence criterion is met. (The convergence criterion required the normal derivative to be less than half a unit in the fifth decimal place, with the computation stopped at 100 iterations.)

Step 6. Convert the arbitrary solution to canonical form by transformation (8), with $T$ determined from (9).

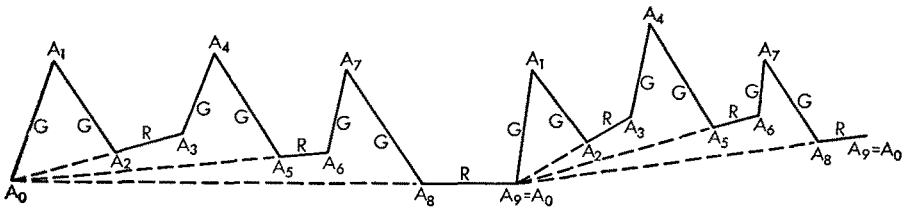
The minimum of the function $f(A)$ is obtained, to the desired degree of accuracy, by proceeding along the $nm$-dimensional path $A_0A_2A_3A_4 \ldots$ given by (14). This is illustrated schematically in Fig. 1, where the gradient and partan steps are indicated by $G$ and $P$, respectively.

The parallel tangents method was tried on many examples. The results were markedly superior to those obtained by Method I, as measured by the speed with which apparent convergence was obtained. Nonetheless, the actual time required to calculate a fair-sized problem (say, 36 variables and 12 factors) still seemed excessively long.

*Higher-order cycles.* In another attempt to hasten convergence of the gradient method, we followed the method of resultant descents proposed by Finkel [5] as improved by Jöreskog [12]. It is similar to partan in that it employs the gradient as well as points already reached in the process. However, it differs from partan by using two gradient steps and one resultant step. Furthermore, a single point is used several times as a base point for resultant extrapolation. The number of times that the base point is employed in this way gives rise to the order of the cycles. This is illustrated in Fig. 2, where cycles of order 2 and of order 3 are indicated, and in which gradient and resultant steps are represented by $G$ and $R$, respectively.




(a) Cycles of Order 2




(b) Cycles of Order 3

FIGURE 2
Schematic Representation of Higher-order Cycles Procedure

The algorithm for cycles of order $k$ may be outlined as follows.

Step 1. Start with arbitrary $A_0$ .

Step 2. Calculate three points (which constitute a composite step) as follows:

$$A_1 = A_0 + tG_0 ,$$

$$A_2 = A_1 + tG_1 ,$$

$$A_3 = A_0 + t(A_2 - A_0).$$

Step 3. Calculate additional composite steps, from the same pivot point, as follows: (15)

$$A_4 = A_3 + tG_3 ,$$

$$A_5 = A_4 + tG_4 ,$$

$$A_6 = A_0 + t(A_5 - A_0).$$

Step 4. After $k$ composite steps (for cycles of order $k$), reset $A_{3k}$ to $A_0$ as the new pivot point.

Step 5. Repeat Steps 2–4 until the convergence criterion is met. (The convergence criterion was the same as that used in partan.)

Step 6. Convert the arbitrary solution to canonical form by transformation (8), with $T$ determined from (9).

The gradient method with higher-order cycles (1 through 5) was tried on many examples. From the viewpoint of time requirements, the best results were generally found with cycles of order 5, although cycles of orders 1, 2, and 3 also were most effective for some of the examples. The relative efficacy of this method did not show any improvement over partan. Hence, an alternative to the gradient methods was sought.

*Method III. Successive Approximations to Factor Loadings (Gauss-Seidel)*

It will be recalled that the gradient methods require the solution of a fourth degree polynomial in the course of minimizing the objective function (7). Believing that this might account for a considerable fraction of the computing time, we were led to consider techniques that would involve lower-order polynomials. A classical iterative method for solving systems of linear equations—Gauss-Seidel (see [19], sec. 130)—seemed to be generalizable to our nonlinear problem. The Gauss-Seidel process is sometimes called a "method of successive displacements" because, as a change is made in a variable and the new corresponding variable is determined, it then replaces the original variable. We then looked for ways of varying the values of the factor loadings in the process of minimizing the function $f$ without giving rise to high powers of the variables.

From the basic theorem of factor analysis (3) it is evident that if changes or displacements are introduced in only one row of $A$, the reproduced correlations will be linear functions of these displacements, and the objective function $f$ will be quadratic only. More explicitly, for any row $j$ in $A$ an increment $x_p$ ($p = 1, 2, \cdots, m$) is added to each element:

$$a_{j1} + x_1 , a_{j2} + x_2 , \cdots , a_{jp} + x_p , \cdots , a_{jm} + x_m . \tag{16}$$

Then the reproduced correlation of variable $j$ with any other variable $k$ is

$$r_{jk}^{\dagger} = \sum_{p=1}^{m} a_{kp}(a_{jp} + x_p) \tag{17}$$

and the sum of squares of residual correlations of variable $j$ with all other variables $k$ is given by

$$f_j = \sum_{\substack{k=1 \\ k \neq j}}^{n} (r_{jk} - r_{jk}^{\dagger})^2 = \sum_{\substack{k=1 \\ k \neq j}}^{n} \left( r_{jk}^{*} - \sum_{p=1}^{m} a_{kp} x_p \right)^2 , \qquad (j \text{ fixed}) \tag{18}$$

where $r_{jk}^{*}$ is the residual correlation of variables $j$ and $k$ without the incremental change in the factor loadings of variable $j$, that is,

$$r_{jk}^{*} = r_{jk} - \sum_{p=1}^{m} a_{jp} a_{kp} . \tag{19}$$

To determine the values of the $x_p$ which minimize the objective function $f$,

take the partial derivatives of (18) with respect to each of these $x$'s, set them equal to zero, and solve for the $x$'s, getting the result

$$\sum_{p=1}^{m}\left(\sum_{\substack{k=1\\k\neq j}}^{n} a_{kp}a_{kq}\right)x_p = \sum_{\substack{k=1\\k\neq j}}^{n} r_{jk}^{*}a_{kq}\,, \qquad (q = 1, 2, \cdots, m). \tag{20}$$

This may be put in matrix form,

$$X_j A'_{j(}A_{)j(} = R_j^0 A, \tag{21}$$

where $X_j$ is the row vector of incremental changes of the factor loadings for variable $j$, $A_{)j(}$ is the factor matrix with the elements in row $j$ replaced by zeros, and $R_j^0$ is the row vector of residual correlations of variable $j$ with all other variables (and 0 for the self-residual). Then the solution for the displacements to the factor loadings (for a given variable) that will minimize the objective function, is

$$X_j = R_j^0 A(A'_{j(}A_{)j(})^{-1}. \tag{22}$$

The foregoing process is carried out systematically for all variables, in turn. Thus successive approximations of rows of factor loadings are obtained which will yield a minimum value for the function $f$ to any desired degree of accuracy. The computing algorithm for Method III can be summarized by the following.

Step 1. Start with arbitrary $A_0$ .

Step 2. Calculate $X_1$ from (22) in which each $A$ is the arbitrary $A_0$ and $R_1^0$ is the row vector of observed correlations to start the process. Designate the factor matrix with the first row replaced by the computed values (16) as $A_1$ (the subscript on the $A$ represents the result of iteration 1, not the pivot variable).

Step 3. Compute a major iteration cycle by pivoting on each variable, in turn. Thus, $A_1$ , $A_2$ , $\cdots$ , $A_n$ will be determined from $X_1$ , $X_2$ , $\cdots$ , $X_n$ . (23)

Step 4. Apply the convergence test after computing $A_{cn}$ , where $c = 1, 2, 3, \cdots$ is the number of major iteration cycles.

Step 5. Repeat Steps 3 and 4 until the convergence criterion is met. (The criterion used was the rate of change in $f$, namely, $\Delta f/f + 10^{-10} \geq 0$, which is actually the limit of accuracy of the 48-bit computer in single precision floating point, and the process was stopped arbitrarily if convergence did not take place by 800 iterations.)

Step 6. Convert the arbitrary solution to canonical form by transformation (8), with $T$ determined from (9).

The method of successive approximations to factor loadings seemed to provide the answer to the problem. It produced a factor matrix satisfying the condition (7) more effectively than any of the other methods tried; and, in absolute terms, obtained such a solution in very reasonable time. For small problems (5, 8, and 9 variables with 2 or 3 factors) the calculations were accomplished in 12 to 68 seconds when pressed to the limit of accuracy of the computer, and in 3 to 18 seconds with resonable and practical standards. For larger problems (24 variables, 4 factors; and 36 variables, 12 factors) minres solutions were still obtained in reasonable time (40 seconds and 8.4 minutes, respectively). While eleven different problems were employed for the empirical study of the different methods, the results for the example (see [7]) of 24 psychological tests and 4 factors is typical of all the others. Method I did not converge for this example, and after 100 iterations and 48 minutes the objective function was reduced to .46006 from an initial value of .92777 (based on the first 4 principal components). The several solutions covered by Methods II and III all converged, taking from about 7 minutes to less than 2 minutes, with $f = .45989$. After relaxing the standard of convergence, the same level for the objective function (to six decimal places) was reached in 39 seconds.

## 5. *Minres Solution*

Having experimented with the several mathematical approaches to a minres solution, and tested them empirically on many problems, we concluded that Method III offered the most efficacious solution. This method is then implied when we simply use the term "minres solution." In the study of the effect of different computing procedures, the calculations were carried far beyond any practical needs—with very rigorous standards for convergence. After the choice of the preferred method was made, consideration had to be given to practical requirements for a minres solution.

Of course, it is apparent that the "best" fit of a model to empirical data, in the sense of the objective function $f$ being a minimum, may not appear very convincing. If the model specifies two factors for a set of 20 variables, it is to be expected that the residuals may be of sizable magnitude, although the sum of squares of off-diagonal residuals has been minimized. All that can be said is that for the given hypothesis, the resulting minres solution best satisfies the least-squares criterion. Unfortunately, a good statistical resolution of the question of the number of factors $m$ is not yet available. (There is a good chance that under certain conditions a statistic that is asymptotically distributed as $\chi^2$ may be available as a large sample test of the hypothesis of the number of factors.)

In any event, for a given hypothesis regarding the number of factors, it is desirable to obtain a stable factor solution. The objective, then, is to set a convergence criterion so as to guarantee the accuracy of the factor loadings. After considering convergence criteria that employed actual changes in the

function $f$, relative changes in $f$, changes in the normal derivative (although this does not otherwise have to be computed in Method III), and changes in the communalities of the variables, we finally came back to consideration of changes in the factor loadings themselves. To be sure that the resulting minres solution had stabilized, we required that the maximum change from one iteration to the next of all factor loadings be bounded by some preassigned small number. In other words, the convergence criterion was set as follows:

$$\max_{j,p} |{}_{(i)}a_{jp} - {}_{(i-1)}a_{jp}| < \epsilon \qquad (j = 1, \cdots, n; p = 1, \cdots, m) \tag{24}$$

where $i$ is the iteration number. For most problems $\epsilon = .001$ is satisfactory, and as a precautionary measure it might be wise to terminate the process if the number of iterations exceeds 1000.

For purposes of illustrating the minres solution, the familiar example [7] of eight physical variables is employed. The correlations among the variables appear in the upper triangle of Table 1. In the experimentation these data were analyzed for the hypothesis of 2 factors and 3 factors by each of the methods discussed above. The details of a minres solution will be shown only for $m = 2$.

The starting point for the minres calculations is an arbitrary factor matrix, and for this purpose the first two principal components were selected (i.e., calculated from the correlation matrix with unities in the diagonal). Of course, the first two principal components account for more variance (6.444) than any other two factors, but the sum of squares of off-diagonal residuals produced from this initial matrix is not a minimum ($f = .07663$). For the minres solution of Table 2 the objective function was improved considerably ($f = .01205$). The two factors which provide the "best" fit to the

TABLE 1

Correlations and Residuals for Eight Physical Variables *

| Variable | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
|---|---|---|---|---|---|---|---|---|
| 1. Height | | 846 | 805 | 859 | 473 | 398 | 301 | 382 |
| 2. Arm span | -014 | | 881 | 826 | 376 | 326 | 277 | 415 |
| 3. Length of forearm | -020 | 027 | | 801 | 380 | 319 | 237 | 345 |
| 4. Length of lower leg | 037 | -020 | -011 | | 436 | 329 | 327 | 365 |
| 5. Weight | 016 | -025 | 007 | 008 | | 762 | 730 | 629 |
| 6. Bitrochanteric diameter | 018 | -006 | 011 | -027 | 008 | | 583 | 577 |
| 7. Chest girth | -021 | 004 | -014 | 029 | 012 | -026 | | 539 |
| 8. Chest width | -024 | 044 | -002 | -020 | -027 | 021 | 015 | |

* Correlations in upper triangle, residuals in lower triangle; decimal points omitted.

TABLE 2

Minres Solution with Two Factors

| Variable j | $F_1$ | $F_2$ | $h_j^2$ |
|---|---|---|---|
| 1 | .856 | -.324 | .838 |
| 2 | .848 | -.412 | .889 |
| 3 | .808 | -.409 | .821 |
| 4 | .831 | -.342 | .808 |
| 5 | .750 | .571 | .889 |
| 6 | .631 | .492 | .640 |
| 7 | .569 | .510 | .583 |
| 8 | .607 | .351 | .492 |
| Variance | 4.449 | 1.510 | 5.959 |

off-diagonal correlations account for almost 75% of the total variance, as can be seen from the last line of Table 2.

In meeting the criterion for the minres solution, no factor loading in Table 2 changed by as much as .001 in going from major iteration cycle 5 (ending with iteration 40) to major iteration cycle 6 (ending with iteration 48). While this clearly indicates that for the model of two common factors the solution is stable, the related question of the adequacy of this hypothesis still remains. If the model were to fit the data precisely, all residuals would vanish—but, of course, this is not to be expected of empirical data. The actual residuals (shown in the lower half of Table 1) range from −.027 to .044 with a mean of zero to more than four decimal places and a standard deviation of .021. A crude criterion for "when to stop factoring" is the requirement that the standard deviation of residuals be less than the standard error of a zero correlation for the given size sample. For a sample of 305 cases, this statistic is .057, and supports the hypothesis of only two factors. This does not prove that there are only two significant common factors; on the contrary, there is much evidence ([7], p. 377) that more than three common factors are required for adequate (statistical) explanation of the observed data. However, from a practical point of view, the magnitude of the residuals in Table 1 may be considered too small to provide another meaningful factor.

Before leaving the numerical example, some simple comparisons of the different computing methods may be of interest. For practical purposes, all methods produced the same results—even though Method I did not converge (according to the strict experimental criterion)—with the individual factor loadings and communalities agreeing to within a couple of units in the third decimal place. Moreover, the objective function came down to .01205390 for Methods II and III and was only one unit larger in the eighth decimal place

for the operational program with the somewhat relaxed convergence standard. While the objective function was .012146 for Method I when the computations were terminated at 100 iterations, it was actually lower (.012063) after 99 iterations. The vacillation around the apparent point of convergence was another disturbing factor about Method I. The important difference, of course, was the time required to reach a solution. From over a minute for Method I, the time was reduced to 22 seconds by Method II (actually, 21 seconds by cycle of order 5), 12 seconds by Method III, and only 4 seconds by the operational version of the last method.

## 6. *Discussion*

There are certain features of a minres solution that deserve emphasis. Some of these may be considered highly desirable; others present puzzling problems, as yet unresolved.

While a principal-factor solution, in general, is not a minres solution, the converse is always true—a minres solution is a principal-factor solution with appropriate diagonal entries. A principal-factor solution for a correlation matrix with minres communalities cannot be different from the minres solution that produced those communalities; if it were, both the off-diagonal and the diagonal sums of squares of residuals would be increased—in the former case because the off-diagonal sum is minimized by minres, and in the latter case because the diagonal sum is zero for minres. This means that the principal-factor and minres solutions are equivalent, and if put in canonical form are identical. In schematic form,

$$(R - I + H_{min}) \xrightarrow{\text{PFA}} A_{min} , \tag{25}$$

the theorem states that a principal-factor analysis of a correlation matrix with minres communalities produces a minres factor solution. A corollary property is that a principal-factor solution of a correlation matrix with $m$-factor minres communalities will have a sum of the $m$ largest eigenvalues equal to the sum of the communalities, while the remaining $n - m$ eigenvalues will be positive and negative and add to zero.

Just as a minres solution reproduces itself through PFA, so does a maximum-likelihood solution, viz.,

$$(R - I + H_{ML}) \xrightarrow{\text{PFA}} A_{ML} . \tag{26}$$

Of course, the principal-factor analysis of the correlation matrix with maximum-likelihood communalities is in canonical form, while the original maximum-likelihood solution (from which the communalities were taken) probably is not, and must first be put in that form in order to verify the equivalence. Now, the factor matrix obtained by (26) is a least-squares fit to $(R - I + H_{ML})$ with perfect fit of the diagonal, and must therefore be a minres solution. Does this mean that minres (with its computational ad-

vantages) may be used in place of the highly desirable maximum-likelihood solution (for which an efficient computing algorithm has not yet been found)?

According to the conditions for minres (7) and for maximum-likelihood ([7], p. 369), the two solutions should be identical only if the communalities are equal. How much the actual communalities for a set of $n$ variables may differ and still produce practically equivalent results is a matter for empirical investigation (so long as it is understood that it is not mathematical equality that is sought). Such an investigation was made for the example of 8 physical variables, but postulating 3 factors instead of 2 as in the preceding section. Specifically, the following solutions were computed.

1. ML (arbitrary initial matrix),
2. PFA (ML communalities),
3. Minres (ML initial matrix),
4. ML (minres initial matrix).

After the maximum-likelihood solutions were put in canonical form, all four solutions were practically identical. Communalities and contributions of factors differed only in the third decimal place, and individual factor loadings, with only a few exceptions, agreed to within a few units in the third place. The equivalence of these solutions was found in spite of the fact that the 8 communalities ranged from .5 to 1.0.

It should be noted, however, that the minres solution found to be equivalent to a maximum-likelihood solution was for a particular local maximum, while another maximum-likelihood solution differed from it. This problem—local versus global maximum or minimum—is unresolved for either the maximum-likelihood or the minres method. Even when a computing procedure converges there is no assurance that the optimal point (maximum or minimum) is for the entire surface in the multidimensional space or only for a local area. In practical usage, this question may be immaterial as long as a reasonably good solution is obtained, but from a theoretical viewpoint it is necessary to know whether the ultimate solution has been reached for the given conditions.

Still another problem requiring resolution may be called the "generalized Heywood case" ([7], pp. 125-6). Occasionally, a minres solution is obtained for which one of the communalities exceeds unity. While this happens only infrequently, nonetheless it is a situation that should be eliminated if the factor analysis is to be acceptable. In its present form, the minres solution satisfies (7) but no side condition restricting the communalities to numbers less than one. If a Heywood case arises in a minres solution, then the computing program merely adjusts the loadings for the unruly variable so that its communality is one, without disturbing any of the other factor loadings. Of course, the objective function must become larger than the minimum reached with the unacceptable communality, and the residuals involving that variable also increase. Forcing any excessive communality back to unity was a solution of

expediency. In the near future, we hope to explore a minres procedure with built-in restrictions on the communalities.

## REFERENCES

[1] Bargmann, R. *Factor analysis program for 7090 preliminary version.* IBM Research Center, Inter. Doc. 28-126, October 1963.
[2] Boldt, R. F. Factoring to fit off diagonals. *Psychometrika* (in press).
[3] Comrey, A. L. The minimum residual method of factor analysis. *Psychol. Reports,* 1962, **11**, 15–18.
[4] Eckart, C. and Young, G. The approximation of one matrix by another of lower rank. *Psychometrika,* 1936, **1**, 211–18.
[5] Finkel, R. W. *The method of resultant descents for the minimization of an arbitrary function.* Paper 71, preprint of paper presented at 14th National Meeting of Association of Computing Machinery, 1959.
[6] Gengerelli, J. A. A simplified method for approximating multiple regression coefficients. *Psychometrika,* 1948, **13**, 135–46.
[7] Harman, H. H. *Modern factor analysis.* Chicago, Ill.: Univ. Chicago Press, 1960.
[8] Horst, P. A method of factor analysis by means of which all coordinates of the factor matrix are given simultaneously. *Psychometrika,* 1937, **2**, 225–36.
[9] Householder, A. S. and Young, G. Matrix approximation and latent roots. *Amer. math. Monthly,* 1938, **45**, 165–71.
[10] Hotelling, H. Analysis of a complex of statistical variables into principal components. *J. educ. Psychol.,* 1933, **24**, 417–41, 498–520.
[11] Howe, W. G. *Some contributions to factor analysis.* Report No. ORNL-1919. Oak Ridge, Tenn.: Oak Ridge National Laboratory, 1955. (Ph.D. Thesis, Univ. North Carolina.)
[12] Jöreskog, K. G. *Testing a simple structure hypothesis in factor analysis.* Res. Bull. RB-65-1. Princeton, N. J.: Educ. Test. Serv., January 1965.
[13] Keller, J. B. Factorization of matrices by least-squares. *Biometrika,* 1962, **49**, 239–42.
[14] Lawley, D. N. A modified method of estimation in factor analysis and some large sample results. *Uppsala Symposium on Psychological Factor Analysis.* Uppsala: Almqvist & Wiksell, 1953, 35–42.
[15] Shah, B. V., Buehler, R. J., and Kempthorne, O. Some algorithms for minimizing a function of several variables. *J. Society for Industrial and Applied Mathematics,* 1964, **12**, 74–92.
[16] Spang, H. A., III. A review of minimization techniques for nonlinear functions. *SIAM Review,* 1962, **4**, 343–65.
[17] Thurstone, L. L. *Multiple-factor analysis.* Chicago, Ill.: Univ. Chicago Press, 1947.
[18] Thurstone, L. L. A method of factoring without communalities. *1954 Invitational Conference on Testing Problems.* Princeton, N. J.: Educ. Test. Serv., 1955, 59–62, 64–66.
[19] Whittaker, E. and Robinson, G. *The calculus of observations.* London: Blackie & Son, 1944.
[20] Whittle, P. On principal components and least square methods of factor analysis. *Skand. Aktuar.,* 1952, **35**, 223–39.

*Manuscript received 7/19/65*