SOME NEW METHODS IN MATRIX CALCULATION[1]

By Harold Hotelling

Columbia University

I. Introduction

1. The increased practical importance of matrix calculation. This paper will be concerned chiefly with methods of finding the inverse of a matrix, and of finding the latent roots and latent vectors, which are also known by a variety of other names associated with particular applications, such as principal axes in geometry and mechanics, and principal components in psychology. These two computational problems are of extremely wide application. The first is closely related to the solution of systems of linear equations, which we shall also consider. In the method of least squares the solution of the normal equations is best carried out with the help of the inverse of the matrix of the coefficients, since at least some of the elements of this inverse matrix are needed in evaluating the results in terms of probability, a vitally necessary step, and since the inverse matrix is useful also in various other ways, such as altering the set of predictors used in a regression equation. Modern statistics also utilizes quadratic and bilinear forms such as the generalized Student ratio [15] for discriminating between samples according to multiple variates instead of one only, the associated discriminant functions [10], the closely related figurative distance of Mahalanobis, Bose and Roy [5] and the critical statistic in an investigation by Wald [28] of the efficient classification of an individual into one of two groups. All these may be calculated very easily from the inverse of a matrix of sums of products, or of covariances or correlations, or from the principal components. Consideration of the relations between two sets of variates [18] may utilize both the inverse of a matrix and a process resembling the calculation of principal components. Similar computational problems arise in applying to sets of numerous variates the contributions to multivariate statistical analysis of R. A. Fisher, S. S. Wilks, W. G. Madow, M. A. Girshick, P. L. Hsu and M. S. Bartlett. Among the non-statistical applications of the inverse matrix and of latent roots and vectors are problems of dynamics, both in astronomy and in airplane design [12], the analysis of stresses and strains in structures [26, 27], and electrical engineering problems [24].

Perhaps no objection to attempts at statistical inference is more common than that the variation of this or that relevant factor has been ignored. For example in dealing with time series the need of allowing for trend and seasonal variation, perhaps by means of a sequence of orthogonal polynomials for trend and of

[1] Revision of a paper presented at the Symposium on Numerical Calculation held Dec. 28, 1941 in New York by the Institute of Mathematical Statistics and the American Statistical Association with the cooperation of the Committee on Addresses in Applied Mathematics of the American Mathematical Society. For the program of the Symposium see the *Annals of Mathematical Statistics* for March, 1942, p. 103.

trigonometric functions for seasonal variation, is well recognized. It is indeed desirable to use regression equations with a liberal number of predictors to eliminate spurious influences, as well as to reduce the error variance, and likewise in other statistical methods. But the computational difficulties in the joint analysis of the desired number of variables have frequently seemed too formidable. We shall see how efficient techniques, in conjunction with efficient machines, can go far to facilitate the use of an appropriate number of variables by reducing the labor to modest dimensions.

While the rise of modern multivariate statistical theory has made available new exact tests of hypotheses in terms of probability over a wide range of cases in which multiple measurements are involved, such measurements have been accumulating on a large scale. In many psychological, anthropometric, astronomical, meteorological and economic fields, actual measurements are available on numbers of variates far greater than have been regarded as amenable, within practical limits, to adequate treatment by the numerical methods generally used. In some instances the number of cases in which complete sets of these variates are available is also large. The 1931 census of India included an extensive sample in which fifty physical variates were measured for each individual. Karl J. Holzinger and his collaborators have worked out and circulated privately a complete matrix of correlations among 78 mental tests. Astronomers have indicated the desirability of a recalculation of the elements of the solar system by means of a gigantic least-square solution with 150 or more unknowns, at the same time deploring the seeming impossibility of this ever being carried out. To apply the methods of modern theoretical statistics to derive from such observations all the important information they contain is an enterprise whose feasibility depends on new numerical methods.

The chief computational problems, apart from those of tabulating and providing convenient approximations for the probability distributions, are (1) the calculation of the many sums of products of pairs of p variates when p is large, and (2) operations on the matrices of these sums of products such as finding the inverse and the principal components. The first problem, which in classical applications of the method of least squares to long series has seemed the heavier, has in a sense been solved by the use of punched cards. A card is used for each case, and all p variates are punched into it. By running the cards repeatedly through a machine wired at each run to select a particular pair of variates, multiply them together, and cumulate the products, this part of the work may be disposed of with great speed. The cost of the machines does at present limit the economical use of this method to rather large numbers, both of variates and of cases. This limit has recently been pushed upward by the introduction of improved multiplying calculators, with high-speed automatic multiplication and squaring locks. But these mechanical advances, in combination with recent discoveries in statistical theory, the increasingly felt need to resort to numerous variates, and the actual existence in many cases of data on such multiple variates, emphasize the need for rapid, economical and accurate calculations with matrices whose elements are sums of products.

Modern machine methods, especially those of the punched-card type, but also those using machines such as the Monroe, Marchant and Fridén, tend to reduce the work of formation of sums of products, in comparison with other operations, to such an extent as to enhance the relative value of methods in which such calculation of direct product-sums is important. Thus products of matrices are much simpler to compute than inverses, and positive than negative powers. Indeed, powers and products of matrices can be computed by means of punched-card machines, and for large matrices this is doubtless the most efficient procedure now available, though considerable rewiring is needed. There is also a possibility, which does not seem too remote, of development of further devices to do this rewiring automatically.

2. Iterative and direct methods. Partitioned matrices. In later sections we shall deal chiefly with certain iterative methods, giving particular attention to the neglected question of limits of error in stopping at any point, and considering the rate of approach to the desired solution. For finding the roots of a matrix and the associated vectors, if the matrix has more than about four rows, it seems clear that an iterative method is the most economical of labor in all but very special cases. On the other hand the problems of solving systems of linear equations and finding the inverse of a matrix do not usually yield readily to iterative methods unless an approximation to the solution is available to begin with. This approximation is not necessarily a very close one, but must not be too wild. It may in some cases be obtained from a general knowledge of the subject.

The Mallock electrical device [22] is capable of solving almost instantaneously ten linear equations in ten unknowns with perhaps two significant digits in each result, though this question of accuracy remains to be elucidated. The combination of this device with the iterative method of Section 7 below, and with the use of partitioning for matrices of more than ten rows, offers what seems at present the best hope for the systematic inversion of large matrices. Since only one of the Mallock machines is in existence (it is in Cambridge, England), some adaptation of the Doolittle or related methods will ordinarily be used. By taking advantage of the possibilities in modern calculating machines of accumulating products to reduce the amount of writing required in the Doolittle method, exceedingly compact and efficient methods have been developed for solving systems of linear equations and for evaluating inverse matrices by Dwyer [7, 8, 9], who utilized the earlier work of Waugh, Kurtz, Horst, Dunlap and Cureton cited by him, and for solving systems of linear equations, by Crout [6]. Dwyer gives valuable bibliographies.

By some of these methods, or from a general knowledge of the subject, one may well obtain approximate solutions correct to a very small number of decimal places, and then by iteration get as many more places as are required, with labor far less than would be necessary to carry through from the beginning the requisite number of places. Further applications of iterative methods arise when a least-square solution is to be revised, either on account of new observa-

tions or because of errors discovered in the original observations or calculations. But however a least-square calculation or the evaluation of any inverse matrix begins, and whatever intermediate steps are taken, it seems advisable to terminate it with the method of Section 7. This combines a check on the previous work, at a labor cost equivalent merely to substituting the values found for the unknowns into the equations, with an improvement in accuracy and a useful limit of error for the unknowns.

In the inversion of large matrices there are important possibilities in the properties of partitioning. For example, a square matrix of $2p$ rows may be partioned into four square matrices a, b, c, d, of p rows, and written

$$\begin{bmatrix} a & b \\ c & d \end{bmatrix}.$$

If this is multiplied on the right by another partitioned square matrix of $2p$ rows which may be written

$$\begin{bmatrix} A & C \\ B & D \end{bmatrix},$$

where A, B, C, D are square p-rowed matrices, the product

$$\begin{bmatrix} aA + bB & aC + bD \\ cA + dB & cC + dD \end{bmatrix}$$

is identical with the result of partitioning the product of the two original $2p$-rowed matrices. If the second is the inverse of the first, this product is the identical matrix. Consequently, if the first matrix is given, we have for determining its inverse the four matrix equations in A, B, C, D,

$$aA + bB = 1 \qquad aC + bD = 0$$

$$cA + dB = 0 \qquad cC + dD = 1,$$

where 1 stands for the identical matrix of p rows and 0 for the p-rowed matrix consisting entirely of zeros. These equations may be solved just as in elementary algebra except that care must be used to perform matrix multiplications in correct order. Thus

$$A = (a - bd^{-1}c)^{-1}, \qquad B = -d^{-1}cA$$

$$D = (d - ca^{-1}b)^{-1} \qquad C = -a^{-1}bD.$$

These formulae call for inversion of four p-rowed matrices, namely d, $a - bd^{-1}c$, a, and $d - ca^{-1}b$. Without changing the number of such inversions we may choose alternative sets of matrices to invert, with economy of labor in certain cases. For example, if b is easy to invert, we may use for D the expression

$$D = b^{-1}aAbd^{-1}.$$

The formulae and numerical work are further simplified if the given matrix is symmetric. Other modes of partitioning are also possible, and may be valuable in various kinds of numerical work. Another method of obtaining the inverse of a matrix by partitioning is given by Frazer, Duncan and Collar [12, pp. 112–118], who also give an account of general properties of partitioned matrices. In the treatment of relations between two or more sets of variates [18, 31], partitioned matrices appear.

The most efficient method of calculation of a function of a matrix will depend in part on what else is to be calculated. For example, if the latent roots and vectors are needed for any reason as well as the inverse of a matrix, it is better to calculate the former first, and then the determination of the inverse matrix becomes a trivial task; but if the latent roots and vectors are not needed for some other purpose it is usually better not to calculate them but to use a more direct method to obtain the inverse. If in addition to the inverse the determinant is wanted, or many consecutive powers of a matrix, or if a matrix-multiplying machine considerably speedier than present procedures becomes available, a method [3] based on the Cayley-Hamilton theorem that a matrix satisfies its own characteristic equation may be recommended.

Iterative methods have what Whittaker and Robinson [30] call the pleasing characteristic that mistakes do not necessarily spoil the whole calculation, but tend to be corrected at later stages. This of course does not mean that there is no penalty for mistakes. They have an obvious tendency to prolong the number of repetitions required, and if repeated at late stages may actually prevent realization of a substantially correct result. A less obvious consequence of mistakes near the termination of an iterative calculation is that they tend to vitiate any limits of error that may be derived, including those that will be found below. Great care should be used to insure accurate calculation especially in the last stages of any iterative process.

To insure accuracy even before the last stages, and therefore efficiency, a check column consisting of the sums of the elements in the rows of matrices multiplied and added together may well be carried along. In multiplying two matrices only the check column of the second factor is used; it is multiplied by each row of the first factor to obtain the check column for the product. A computer thoroughly experienced with matrix multiplication may dispense with the check column at all stages but the last of an iterative process, relying on the self-correcting property of the process.

A simple but extremely valuable bit of equipment in matrix multiplication consists of two plain cards, with a re-entrant right angle cut out of one or both of them if symmetric matrices are to be multiplied. In getting the element of the ith row and jth column of the product, the ith row of the first factor and the jth column of the second should be marked by a card beside, above, or below it. In writing a symmetric matrix it is convenient to omit the elements below the principal diagonal. The re-entrant right angle is then utilized to mark off the numbers belonging to a particular row.

A report [13] on certain iterative methods of solving linear and other equations and of calculating latent roots and vectors, with engineering applications, was published by R. von Mises and H. Geiringer in 1929. As part of a discussion of certain problems in psychology [16] the present author in 1933 described iterative processes both for solving systems of linear equations and for finding principal components, and later [17] showed how to accelerate convergence to principal components by repeatedly squaring the matrix. Further acceleration of convergence by other devices has been discovered by A. C. Aitken [2]. Dr. Geiringer has also discussed a method of solution of equations involving iteration by small groups of unknowns [14]. The method of Kelley and Salisbury [20] should be noted. It has been used extensively by psychologists. Definite limits of error and measures of rate of convergence for this method are missing. Certain other iterative methods will be discussed in later sections. It will appear that the most-used methods are by no means the best.

Questions regarding the probability of a matrix of covariances satisfying particular conditions of computational significance may in some cases be illuminated with the help of the theory of the variates as a random sample of a larger aggregate. This theory was outlined in the latter part of the paper [16].

II. Linear Equations and Inverse Matrices

3. Accuracy of direct solution of linear equations. The question how many decimal places should be retained in the various stages of a least-square solution and of other calculations involving linear equations has been a puzzling one. It has not generally been realized how rapidly errors resulting from rounding may accumulate in the successive steps of such procedures as, for example, the Doolittle method. In this popular algorism for solving a system of equations

$$\sum_{j=1}^{p} a_{ij}x_j = g_i \qquad (i = 1, \cdots, p),$$

the equivalent of successive eliminations of $x_1, x_2, \cdots, x_{p-1}$ to obtain an equation in x_p alone is accomplished by calculating successively

$$a_{ij\cdot 1} = a_{ij} - a_{i1}a_{1j}/a_{11}, \qquad g_{i\cdot 1} = g_i - a_{i1}g_1/a_{11} \qquad (i, j = 2, 3, \cdots, p),$$

then

$$a_{ij\cdot 12} = a_{ij\cdot 1} - a_{i2\cdot 1}a_{2j\cdot 1}/a_{22\cdot 1},$$

$$g_{i\cdot 12} = g_{i\cdot 1} - a_{i2\cdot 1}g_{2\cdot 1}/a_{22\cdot 1} \qquad (i, j = 3, \cdots, p),$$

and so forth. Let us suppose that each of the a_{ij}'s and g_i's is subject to an error concerning which it is known only that its absolute value does not exceed ϵ. Thus if they are given accurately to k decimal places only, we have $\epsilon = 10^{-k}/2$. Let the actual errors be represented by δa_{ij} and δg_i. If these are small an estimate of the error in $g_{i\cdot 1}$ may be obtained by expanding in a Taylor series and retaining only the linear terms:

$$\delta g_{i\cdot 1} = \delta g_i - \frac{a_{i1}}{a_{11}}\delta g_1 - \frac{g_1}{a_{11}}\delta a_{i1} + \frac{a_{i1}g_1}{a_{11}^2}\delta a_{11}.$$

The closest upper bound for this error obtainable without special assumptions regarding the values of the given quantities is specified by the inequality

$$|\delta g_{i\cdot 1}| \leq \epsilon\left(1 + \left|\frac{a_{i1}}{a_{11}}\right| + \left|\frac{g_1}{a_{11}}\right| + \left|\frac{a_{i1}g_1}{a_{11}^2}\right|\right).$$

The a's and g's are often correlation coefficients. Any set of normal equations of least squares may be reduced to a form in which this is the case, and this reduction has considerable merits. The various correlation coefficients are frequently of interest in themselves, and their use in the normal equations practically insures that all the quantities appearing at any stage are of the same order of magnitude. This last is a very substantial advantage, partly because of the check column which is customarily carried along, in which each entry is the sum of the other entries in its row. Since the absolute value of a correlation coefficient is less than unity, and since a_{ii} becomes equal to unity, the last inequality gives in this case

$$|\delta g_{i\cdot 1}| < 4\epsilon,$$

and no closer inequality appears possible. In the same way we find for this case in which the a's are correlation coefficients that

$$|\delta a_{ij\cdot 1}| < 4\epsilon.$$

Proceeding from these inequalities in the same way, and neglecting the fact that $|a_{22\cdot 1}| < 1$ though like a_{11} it is put equal to unity in the argument, we find for the errors in $a_{ij\cdot 12}$ and $g_{i\cdot 12}$ the estimated upper bound 16ϵ, with an actual upper bound somewhat higher unless $a_{12} = 0$. Continuing in the same way we find for $a_{ij\cdot 12\cdots(p-1)}$ and $g_{i\cdot 12\cdots(p-1)}$ the estimated limit of error $4^{p-1}\epsilon$, with a possibility of a somewhat higher value up to $4^{p-1}\epsilon/a$, where a is the determinant $|a_{ij}| < 1$. The rapidity with which this increases with p is a caution against relying on the results of the Doolittle method or other similar elimination methods with any moderate number of decimal places when the number of equations and unknowns is at all large. Thus if $p = 11$ the limit of error exceeds a million times ϵ, indicating that if only one decimal place is wanted in the value of x_p the original correlations must be utilized to at least seven decimals, even if we neglect the additional errors introduced by dropping decimals beyond those retained in the intermediate stages of the calculation. The errors accumulate further during the back solution, so that if all the unknowns are wanted with one-place accuracy it is necessary to use the original correlations with substantially more than seven decimal places. For larger values of p the increase in the error limit is startling. Thus for $p = 27$ (the number of tests reported to be involved in a certain current procedure in classifying military personnel) the limit of error even for the first unknown evaluated is $4^{26}\epsilon$, repre-

senting a loss of about 16 decimal places of accuracy, while the correlations in Holzinger's 78-rowed matrix would need to be carried to no less than 46 places to insure even an approximate accuracy in the first decimal place of one of the regression coefficients in a formula derived by least squares for predicting one of his variates in terms of all the others.

These high limits of error may possibly be reduced in the following ways: (a) a more exact study of the error might be made by means of terms of the Taylor series of orders higher than the first; (b) the positive definite character of a correlation matrix (or other matrix of normal equations) might be utilized in an attempt to arrive at lower limits of error; (c) instead of considering the maximum possible error we might depend on some mutual cancellation of different errors and content ourselves with statements in terms of probability. The compounding of different errors of rounding, which may individually be regarded as having a probability distribution of uniform density over a fixed range, quickly gives rise to an almost exactly normal distribution of known mean and variance, so that the probability approach is attractive. However the limits of error obtained in this way with, for example, a five per cent level of probability of a greater error, though somewhat smaller than the limits associated with certainty, are disappointingly large. Investigations of the types (a) and (b) have not been made; they would apparently be very cumbersome, and (a) might have the effect of increasing the error limits considered above instead of cutting them down. Use of the check column does not provide any safeguard against the errors of rounding appearing in the original correlations, though from the probability standpoint, a carefully devised use of the check column may mitigate the accumulation of errors in successive stages.

To control such errors reliance is often placed in a substitution of the solution obtained in the given equations. This is not completely satisfactory, since under some circumstances large errors in the solution may yield only slight deviations of the left from the right members of the equations, and since some deviations must be expected in any case in which only a limited number of decimals is carried along. Moreover this substitution, even if it reveals the existence of errors, does not usually make clear at once what should be done about them. A recalculation to a larger number of decimal places is horribly laborious. There is here a distinct need of using an iterative process for improving on the solution obtained, and setting definite limits for the errors.

4. The classical iterative method. The iterative method which seems to be the oldest and the most used for solving systems of linear equations, and which may like all other methods of doing this be applied to find the inverse of a matrix, is that of Gauss and Seidel. It seems also to be used in the "method of relaxations" [26], which has been recommended to engineers but lacks limits of error and measures of rate of convergence.

This classical method, starting with any assumed values for the unknowns, begins by changing the value for the first unknown so as to satisfy the first equation; this is possible if the coefficient is different from zero. The revised

set of trial values is then further altered by changing the second unknown so as to satisfy the second equation. Then the third unknown is altered so that the third equation will be satisfied, and so forth. When all the unknowns have been thus altered the cycle may be begun again, and repeated until the differences between consecutive values of each unknown become small enough to indicate a satisfactory convergence. The method converges if the matrix A of the coefficients a_{ij} is positive definite, as it is for the normal equations of least squares, and also in certain engineering applications [7, 8, 9]. Moreover the character of being positive definite insures that each a_{ii} differs from zero, so that the successive adjustments indicated are all actually possible. In the published discussions, proofs of convergence have sometimes been omitted, and in some cases (e.g. [30], Sec. 130) the proofs are incomplete. Even the fuller proofs [13] and [16] fail to give explicit limits for the errors in stopping at any particular stage. But from the discussion [16, pp. 502, 504] it is easy to see that positive numbers d_i and k exist, with $k < 1$, such that the error in the mth estimate of x_i is less than $d_i k^m$. This limit of error diminishes in geometric progression with successive iterations; hence the number of decimal places of accuracy increases approximately in arithmetic progression. The progression is however irregular and the trial values may fluctuate considerably. Numerical determination of limits of error does not appear to be easy. Experience with the method indicates that it is satisfactory only in case a really good approximation is available to begin with, in spite of its universal convergence.

5. An acceleration and extension of the classical iteration. This classical scheme may be improved in the following way if numerous cycles of revision of the trial values are expected to be needed for the requisite accuracy. The first step, consisting of replacing the trial value x_1 by

$$x_1' = (g_1 - a_{12}x_2 - \cdots - a_{1p}x_p)/a_{11}$$

and leaving $x_2, \cdots, x_p$ unchanged, amounts to subjecting the $p + 1$ variables $x_0, x_1, \cdots, x_p$ to the homogeneous transformation

$$\begin{aligned} x_0' &= x_0 \\ x_1' &= (g_1 x_0 - a_{12}x_2 - \cdots - a_{1p}x_p)/a_{11} \\ x_2' &= x_2 \\ &\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\ x_p' &= x_p, \end{aligned}$$

where the symbol x_0, introduced for convenience in order to make these equations homogeneous, is always equal to unity. The matrix of the transformation,

$$T_1 = \begin{bmatrix} 1 & 0 & 0 & 0 & \cdots & 0 \\ g_1/a_{11} & 0 & -a_{12}/a_{11} & -a_{13}/a_{11} & \cdots & -a_{1p}/a_{11} \\ 0 & 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 0 & 1 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & 0 & 0 & \cdots & 1 \end{bmatrix},$$

is of course singular. If X_0 denote the one-column, $(p + 1)$-rowed matrix of the initial trial values, with unity at the head of the column, the column matrix $X_1 = T_1X_0$ is the result of this first operation, again with unity at the head of the column. The trial values obtained by the second operation appear likewise in the column matrix $X_2 = T_2X_1 = T_2T_1X_0$, where

$$T_2 = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & \cdots & 0 \\ 0 & 1 & 0 & 0 & 0 & \cdots & 0 \\ g_2/a_{22} & -a_{21}/a_{22} & 0 & -a_{23}/a_{22} & -a_{24}/a_{22} & \cdots & -a_{2p}/a_{22} \\ 0 & 0 & 0 & 1 & 0 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & 0 & 0 & 0 & \cdots & 1 \end{bmatrix}.$$

The result of a complete cycle of substitutions may be written $X_p = T_pT_{p-1} \cdots T_2T_1X_0$, where the matrices T_i are of the same simple character illustrated by T_1 and T_2. This same result will be obtained, because of the associative law of matrix multiplication, if we first calculate numerically the matrix

$$T = T_pT_{p-1} \cdots T_2T_1$$

and then $X_p = TX_0$. (Experience shows that computers need at this point the caution that the matrices must be arranged in their proper order. A good procedure is first to form T_2T_1, then to multiply this by T_3 on the left, etc.). This requires rather more work than the original Gauss-Seidel scheme, and therefore is not worth while if only one cycle of substitutions is needed.

The advantage lies in the fact that T may readily be squared, and T^2X_0 gives a result equivalent to that of two full cycles of iteration by the Gauss-Seidel method. Furthermore, T^2 may be squared to give T^4, which may also be squared, and so on. Obviously k such squarings give a matrix which, when multiplied by X_0, yields the same result as 2^k complete cycles of the original substitutions. In terms of the number k of squarings the number of decimal places of accuracy tends to increase in geometrical instead of arithmetic progression. This modification of the classical method does not seem to have been published heretofore, though both it and the method of Section 7 have been in use by the author and his students since 1936.

R. A. Fisher [11, Sec. 29] has introduced the valuable method of finding the inverse of a matrix A by solving together p systems, each of p equations in p unknowns, with the same matrix A of coefficients, but different columns of unknowns; these several columns of unknowns are the elements of the identical matrix. The technique of carrying this out by any of the methods resembling that of Doolittle is a simple extension involving replacement of the right-hand members of the equations by 1's and 0's and carrying along p such columns instead of one while applying exactly the same linear operations to the rows as in the older problem. This, like the problem of solving linear equations, has been elegantly adapted to efficient calculation with modern machines by Dwyer

[7, 8, 9]. The foregoing iterative methods may also be applied in this case, but the matrix T will be different for the different columns. When the given matrix is symmetric (as is implied by the positive definite character assumed in the proofs of convergence) the number of iterations required is generally cut down because the determination of each column determines also the elements of the corresponding row which lie in other columns. Iteration by groups [14] may well have a place here.

An observation of A. C. Aitken's [1] is noteworthy in connection with the solution of equations with a non-symmetric matrix, and with the finding of the inverse of such a matrix. Writing the equations in the matrix form $AX = G$, we see that the solution $X = A^{-1}G$ is also the solution of the system $(A'A)X = A'G$, where A' is the transverse (also called the transpose or conjugate) of A. Evidently $A'A$ and $A'G$ can be formed by direct multiplications and additions, without divisions. Since $A'A$ is symmetric, any of the methods for solving symmetric equations are applicable to the new system. To find the inverse of A we may first find the inverse of the symmetric matrix $A'A$ and then postmultiply it by A'; for $(A'A)^{-1}A' = A^{-1}$.

6. Roots, norms and convergence of matrices. The *norm* of a matrix A may be defined as the square root of the sum of the products of its elements by their complex conjugates, and denoted by $N(A)$. If A is real and a_{ij} is the element in the ith row and jth column,

$$N(A) = \sqrt{\Sigma\Sigma a_{ij}^2}. \tag{6.1}$$

This is the same function which Wedderburn [29, p. 125] defines as the absolute value of A and denotes by A with a heavy vertical bar on each side. Since it is rather troublesome to avoid confusing this with the determinant of A, we use the notation $N(A)$, though the analogy with the ordinary absolute value of a quantity is very suggestive in connection with proofs of convergence and limits of error obtained by means of the "triangular inequalities" below. Rella [25] gives a different definition of the absolute value of the matrix as the maximum of the absolute values of its roots.

The triangular inequalities, whose proof is easy with the help of the Cauchy inequality, are:

$$N(A + B) \leq N(A) + N(B), \tag{6.2}$$

$$N(AB) \leq N(A)N(B). \tag{6.3}$$

From the last it follows that for any positive integer m,

$$N(A^m) \leq [N(A)]^m. \tag{6.4}$$

Hence if $N(A) < 1$, the limit of $N(A^m)$ as m increases is zero. It then follows that the limit of A^m itself is zero, i.e. that each of its elements approaches zero, because of the definition of the norm.

The identical matrix of p rows, which we shall denote simply by 1, has the

norm $\sqrt{p}$, while a scalar matrix k (i.e. one with the quantity k in each element of the principal diagonal and zeros elsewhere) has the norm $k\sqrt{p}$. The norm of a p-rowed orthogonal matrix is $\sqrt{p}$.

The roots of a square matrix, also known as the latent roots or characteristic roots, are the values $\lambda_1, \cdots, \lambda_p$ of λ for which the determinant obtained by subtracting λ from each element of the principal diagonal vanishes. By expanding this determinant in powers of λ and using a relation between roots and coefficients of an equation, it is evident that the sum of the roots equals the sum of the elements in the principal diagonal. This sum is known as the *trace* of the matrix and denoted by $\mathrm{tr}(A)$. Thus

$$\lambda_1 + \lambda_2 + \cdots + \lambda_p = \mathrm{tr}(A). \tag{6.5}$$

From the definitions of the transverse and norm of A it is plain that

$$[\mathrm{N}(A)]^2 = \mathrm{tr}(AA') \tag{6.6}$$

if A is real.

If $f(x)$ is any polynomial in x, $f(A)$ is a matrix whose roots are known [29, p. 30] to be $f(\lambda_i)$, $(i = 1, 2, \cdots, p)$. In particular, the roots of A^m are λ_i^m. Consequently

$$\lambda_1^m + \lambda_2^m + \cdots + \lambda_p^m = \mathrm{tr}(A^m). \tag{6.7}$$

All the roots of a zero matrix are zero. But the fact that all the roots of a matrix are zero does not necessarily imply that the matrix is zero; for example the roots of

$$\begin{bmatrix} 2 & -1 \\ 4 & -2 \end{bmatrix} \tag{6.8}$$

are both zero. But for real symmetric matrices the vanishing of all the roots does imply the vanishing of the matrix; for the sum of the squares of the elements of a symmetric matrix equals the sum of squares of the roots, since $A = A'$, and by (6.7), (6.6) and (6.1),

$$\Sigma\lambda_i^2 = \mathrm{tr}(A^2) = \mathrm{tr}(AA') = [N(A)]^2 = \Sigma\Sigma a_{ij}^2.$$

Moreover, by continuity considerations, a sequence of p-rowed symmetric matrices must approach zero if all the roots approach zero, and conversely.

From this it is evident that a necessary and sufficient condition that A^m approach zero as m increases, when A is symmetric, is that all the roots of A be less than unity. This provides a sharper criterion of convergence than the requirement that $N(A) < 1$, which is sufficient but not necessary for convergence. The latter is however far easier to apply in most numerical work, since it is far easier to compute $N(A)$ than the greatest root. Moreover it is easy to set an upper bound for $N(A)$ in various ways, of which the crudest is to notice that, by (6.1), $N(A)$ cannot exceed p times the greatest absolute value of any

element of A. Also, the test in terms of the norm is applicable to asymmetric as well as symmetric matrices.

From these considerations regarding the convergence of A^m we deduce at once the following result. If the norm of a square matrix is less than unity, then all the roots are less than unity in absolute value. The converse is not true, as the example (6.8) shows.

For any real square matrix A, symmetric or not,

$$\lambda_1^2 + \lambda_2^2 + \cdots + \lambda_p^2 \leq [N(A)]^2. \tag{6.9}$$

To prove this, we observe first that $2a_{ij}a_{ji} \leq a_{ij}^2 + a_{ji}^2$, and consequently $\mathrm{tr}(A^2) \leq \mathrm{tr}(AA')$. From (6.7) and (6.6) we then have $\Sigma\lambda_i^2 \leq \mathrm{tr}(AA') = [N(A)]^2$. This reasoning shows incidentally that $\Sigma\lambda_i^2$ is real, though the individual roots may be complex.

Not only for investigating convergence, but also in the important but neglected problems of setting definite limits of error after a finite number of steps, the norm is an extremely useful function. If a matrix is to be computed with such accuracy that the error in each element is less than δ, and A is the matrix of errors, the requisite accuracy will according to (6.1) be attained when $N(A) < \delta$. The definition and theorems regarding the norm are valid without any restriction to square matrices, for which alone the roots are defined. For example, we may use the norm to derive an inequality concerning the solution of the system of p linear equations

$$\Sigma a_{ij}x_j = g_i,$$

which may be written in matrix form $AX = G$, where A is a square matrix and X and G are matrices each of one column and p rows. From (6.3) we find $N(G) \leq N(A)N(X)$, whence

$$N(X) \geq N(G)/N(A).$$

We shall now deduce a result which seems to be new to matrix theory and which we shall later apply to find limits of error. If A is any matrix such that $1 - A$ is non-singular the identity

$$(1 - A)^{-1} = 1 + A + A^2 + \cdots + A^{m-1} + A^m(1 - A)^{-1}$$

holds, and may be demonstrated exactly as if A were an ordinary scalar quantity. Suppose that $N(A) \leq k < 1$. Taking the norm and using (6.2), (6.3) and (6.4), we have

$$N[(1 - A)^{-1}] \leq p^{1/2} + k + k^2 \cdots + k^{m-1} + k^m N[(1 - A)^{-1}].$$

Since $k < 1$ we may solve for $N[(1 - A)^{-1}]$. Summing the geometric progression, we obtain:

$$N[(1 - A)^{-1}] \leq \frac{p^{1/2} - 1}{1 - k^m} + \frac{1}{1 - k}.$$

This holds for every positive integral value of m, and therefore in the limit when m becomes infinite. Thus we find that

$$N[(1 - A)^{-1}] \leq p^{1/2} - 1 + \frac{1}{1 - k} \tag{6.10}$$

whenever $N(A) \leq k < 1$.

7. An efficient inversion procedure. Let C_0 be an approximation to the inverse of a matrix A, and consider the following sequence of operations. Calculate

$$C_1 = C_0(2 - AC_0), \tag{7.1}$$

and then in turn $C_2, C_3, \cdots$ where

$$C_{m+1} = C_m(2 - AC_m). \tag{7.2}$$

Let us inquire as to the conditions under which the sequence of matrices C_m converges to A^{-1}, the maximum error that may be committed in stopping at any stage, and the rate of convergence. Suppose that C_0 is a good enough approximation to A^{-1} to make the roots of the matrix

$$D = 1 - AC_0 \tag{7.3}$$

all less than unity in absolute value. Then increasing powers of D approach zero, and the convergence of C_m to A^{-1} will follow from the relation

$$C_m = A^{-1}(1 - D^{2^m}), \tag{7.4}$$

which will now be proved by mathematical induction. From (7.1) and (7.3),

$$C_1 = A^{-1}(AC_0)(1 + D) = A^{-1}(1 - D)(1 + D) = A^{-1}(1 - D^2),$$

so that (7.4) is verified for $m = 1$. Now assume (7.4) for a particular value of m, and substitute it in (7.2). This gives

$$C_{m+1} = A^{-1}(1 - D^{2^m})(1 + D^{2^m}) = A^{-1}(1 - D^{2^{m+1}}),$$

which being of the same form as (7.4) completes the induction.

If $N(D) \leq k < 1$ the roots of D are all less than unity in absolute value, as shown in Sec. 6, and the foregoing result holds. Assuming this to be true we now derive an upper bound for the error in C_m in terms of k and $N(C_0)$. According to (7.3),

$$A^{-1} = C_0(1 - D)^{-1}.$$

Hence, by (7.4),

$$C_m - A^{-1} = -A^{-1}D^{2^m} = -C_0(1 - D)^{-1}D^{2^m}.$$

Therefore, by (6.3), (6.4) and (6.10),

$$N(C_m - A^{-1}) \leq N(C_0)k^{2^m}\left(p^{1/2} - 1 + \frac{1}{1 - k}\right). \tag{7.5}$$

This sets an upper bound for the difference between each element of C_m and the corresponding element of A^{-1}. A slightly looser but simpler limit may be obtained from this in terms of the greatest absolute value c of any element of C_0. Since $N(C_0) \leq cp$,

$$N(C_m - A^{-1}) \leq k^{2^m} cp\left(p^{1/2} - 1 + \frac{1}{1-k}\right). \tag{7.6}$$

The great value of this method, whenever a good enough initial approximation is available to make $N(D)$ less than unity, is that the number of decimal places of sure accuracy increases in *geometric* progression, rather than in arithmetic progression as with the usual methods. Consequently this method will always be the most efficient if a sufficiently large number of decimal places is required. Moreover, a limit can be set in advance for the number of iterations that will be required in order to insure any required degree of accuracy. If certainty of correctness in the sth decimal place is required we may choose m so that the right-hand member of (7.6) is less than $10^{-s}/2$. In terms of logarithms to the base 10 the number of decimal places whose accuracy is assured by m iterations is thus at least

$$2^m \mid \log k \mid - \log 2 - \log cp[p^{1/2} - 1 + (1-k)^{-1}]. \tag{7.7}$$

These limits of error can be bettered after some iterations have actually been made. When C_r becomes available we may calculate $k_r = N(1 - AC_r)$, which may be used in place of k in the formulae just derived if m is replaced by $m - r$, and is generally enough smaller than k to make a marked improvement.

The elements of the matrix of errors will actually, of course, be smaller than the norm of this matrix in every practical case, in a ratio fluctuating about p^{-1}. The limits obtained by our formulae can be reached only in case the entire error of the matrix C_m is concentrated in one element, a very unlikely event. Thus the limits given above will usually be quite conservative.

As the iteration proceeds the elements of the matrix $D_m = 1 - AC_m = D^{2^m}$ will diminish rapidly in case of convergence. For this reason it may sometimes be better to calculate C_{m+1} not directly from (7.2), but from the formula

$$C_{m+1} = C_m + C_m D_m \tag{7.8}$$

in which the last term can be regarded as a correction of C_m which will often be very small. This method, however, lacks the self-checking feature, so that its use at the final stage is dubious.

This iterative process has been noticed previously [12, p. 120], but without a limit of error or observation of the geometric progression in the number of accurate digits.

If the initial approximation is not good enough to make $N(D) < 1$, it may be improved by other methods, such as those of Sections 4 and 5, to the point at which this more rapid method becomes applicable. But in some cases (e.g. the second example of §8) the method converges even though $N(D) > 1$, as

may be demonstrated at a later stage at which the norm of the matrix corresponding to D becomes numerically less than unity.

For the mass of least-square and other problems in which the inverse of a matrix is needed, the best procedure appears to begin with one of the methods described by Dwyer [7, 8, 9], carried to a small number of decimal places, and then to calculate D from (7.3), a step equivalent to substituting the approximate solution obtained into the equations. It may then be evident at a glance that the norm of D is so small that the method of the present section will converge rapidly to give as many more places as desired. If $N(D)$ is too large for this, and if gross errors have been eliminated, there is a choice between recalculation from the beginning, the classical iterative process, and the acceleration of this process by matrix-squaring, with perhaps some iteration by small groups. The choice will depend partly on how much the elements of D need to be reduced. The classical iteration (or sometimes the process of this section) is appropriate for correcting a slight excess of $N(D)$ over unity, its matrix-squaring extension for larger alterations.

Let E_0 be the error in C_0, so that $C_0 = A^{-1} + E_0$. Then by (7.1),

$$C_1 = (A^{-1} + E_0)(1 - AE_0) = A^{-1} - E_0 AE_0 .$$

If E_1 is the error in C_1, so that $C_1 = A^{-1} + E_1$, we thus have

$$E_1 = -E_0 A E_0 .$$

If A is symmetric, we naturally take C_0 as a symmetric matrix, and this will cause E_0, C_1, and E_1 also to be symmetric. If also A is positive definite, it will follow from the last equation that E_1 is *negative* definite, or negative semi-definite. Consequently the diagonal elements of C_1 tend to underestimate the corresponding elements of A^{-1}, and never exceed them. Furthermore, the value of a quadratic form whose matrix is A^{-1} will be at least as great as the estimate of it based on C_1. The squares, both of the multiple correlation coefficient and the generalized Student ratio [15], can be expressed as such quadratic forms. Hence both these statistics are slightly underestimated when C_1 is used in place of the true matrix of coefficients. Later approximations C_m do not change the signs of these biases, though they make their magnitudes approach zero in case the conditions for convergence are satisfied, and definite limits converging to zero are easily found for them in such cases from the results above.

8. Illustrations and further comments. We shall indicate symmetric matrices by writing only the elements on and above the principal diagonals.

To illustrate various methods Dwyer [7] has evaluated the inverse of

$$A = \begin{bmatrix} 1.0 & .4 & .5 & .6 \\ & 1.0 & .3 & .4 \\ & & 1.0 & .2 \\ & & & 1.0 \end{bmatrix}$$

as

$$\begin{bmatrix} 2.0710 & -.1913 & -.7759 & -1.0109 \\ & 1.2842 & -.2186 & -.3552 \\ & & 1.3989 & .2732 \\ & & & 1.6940 \end{bmatrix}.$$

If the accuracy of the calculation had been only such as to insure correctness in the first decimal place the approximation to A^{-1} would have been

$$C_0 = \begin{bmatrix} 2.1 & -.2 & -.8 & -1.0 \\ & 1.3 & -.2 & -.4 \\ & & 1.4 & .3 \\ & & & 1.7 \end{bmatrix}.$$

It is easy by mental arithmetic alone, without the use of a machine or side calculations, to see that

$$D = 1 - AC_0 = \begin{bmatrix} -.02 & .02 & 0 & -.01 \\ 0 & 0 & -.02 & .03 \\ .01 & -.01 & 0 & -.02 \\ -.02 & .04 & -.02 & 0 \end{bmatrix}$$

and further that $N(D) = \sqrt{.0052} = .072$. This is so much less than unity that the iteration process of §7 will converge rapidly. As a matter of fact, without determining the sum of the squares of the elements of D we could have observed at a glance that $N(D)$ must be less than four times the greatest absolute value of an element, and thus have a value less than .16. In the same way $N(C_0)$ is seen to be less than 8.4; actually it equals 3.8588. The latter value, with $k = .072$, $p = 4$, substituted in (7.5) gives for the norms of the successive error matrices $E_m = C_m - A^{-1}$,

$$\begin{aligned} N(E_0) &\leq 8.03k = .578, \\ N(E_1) &\leq 8.03k^2 = .0414, \\ N(E_2) &\leq 8.03k^4 = .000216, \\ N(E_3) &\leq 8.03k^8 = .000\,000\,0058. \end{aligned}$$

This promises merely that after one application of the iterative process the results will be accurate to one decimal place, which we know already but might not have known for sure in such a case; that a second iteration will give results accurate to three places, and that a third will give results accurate to about eight places. These estimates will however be improved after actually computing C_1. This may well be done by (7.1) if a machine is available; otherwise, and almost as easily, by (7.8) we obtain

$$C_1 = \begin{bmatrix} 2.070 & -.190 & -.776 & -1.011 \\ & 1.282 & -.218 & -.355 \\ & & 1.398 & .274 \\ & & & 1.692 \end{bmatrix}$$

and $N(C_1) = 3.8163$. (We have now passed beyond the stage of easy mental calculation, but might alternatively use the easy upper bound 8.28 for $N(C_1)$, obtained as before.) We shall use this value instead of $N(C_0)$ in (7.5) and at the same time use for k the value of $N(D_1)$, where

$$D_1 = 1 - AC_1 = 1 - AC_0(1 + D) = 1 - (1 - D)(1 + D) = D^2.$$

This is most easily found from D, from which it may be written down directly by mental calculation:

$$D_1 = 10^{-4} \times \begin{bmatrix} 6 & -8 & -2 & 8 \\ -8 & 14 & -6 & 4 \\ 2 & -6 & 6 & -4 \\ 2 & -2 & -8 & 18 \end{bmatrix}$$

The norm of D_1 is seen by the crude method to be less than .0072, and is actually .003212. Taking the latter value for k we have, similarly to (7.5),

$$N(E_m) \leq N(C_1)k^{2^{m-1}} \times 2.00323 = 7.645k^{2^{m-1}}$$

Thus,

$$\begin{aligned} N(E_1) &\leq .0246, \\ N(E_2) &\leq .000\,0789, \\ N(E_3) &\leq .000\,000\,000\,8. \end{aligned}$$

The reduction in these limits of error is due to the difference between $[N(D)]^2 = .0052$ and $N(D^2) = .003212$.

Using $C_2 = C_1 + C_1D_1$ we obtain:

$$C_2 = \begin{bmatrix} 2.0710366 & -.1912542 & -.7759568 & -1.0109294 \\ & 1.2841486 & -.2185780 & -.3551910 \\ & & 1.3989056 & .2732260 \\ & & & 1.6939852 \end{bmatrix}.$$

From this we calculate

$$D_2 = 1 - AC_2 = 10^{-8} \times \begin{bmatrix} 112 & -164 & -40 & 168 \\ -164 & 288 & -136 & 88 \\ 64 & -128 & 100 & -104 \\ 48 & -32 & -184 & 364 \end{bmatrix},$$

agreeing with the value obtained from the formula $D_2 = D_1^2$, and finally $C_3 = C_2(1 + D_2) =$

$$\begin{bmatrix} 2.071\,038\,458 & -.191\,256\,831 & -.775\,956\,284 & -1.010\,928\,962 \\ & 1.284\,153\,005 & -.218\,579\,235 & -.355\,191\,257 \\ & & 1.398\,907\,104 & .273\,224\,045 \\ & & & 1.693\,989\,071 \end{bmatrix}$$

which as shown above is correct to at least eight decimal places, and doubtless more, in each element. The estimate of A^{-1} obtained by Dwyer by several

direct methods to four places is corroborated by this result excepting for a slight error in the element in his first row and third column.

(ii) Suppose that the approximation in the foregoing example had been even cruder, with determination of the elements of A^{-1} only to the nearest integer. This would give

$$C_0 = \begin{bmatrix} 2 & 0 & -1 & -1 \\ & 1 & 0 & 0 \\ & & 1 & 0 \\ & & & 2 \end{bmatrix},$$

$$D = \begin{bmatrix} .1 & -.4 & .5 & -.2 \\ -.1 & 0 & .1 & -.4 \\ .2 & -.3 & .5 & .1 \\ 0 & -.4 & .4 & -.4 \end{bmatrix}.$$

The sum of the squares of the elements of D is 1.51, so that the norm is greater than unity, and it is not clear at this stage whether the iterative process we have been using will converge or not. But upon computing

$$D^2 = \begin{bmatrix} .15 & -.11 & .18 & .27 \\ .01 & .17 & -.16 & .19 \\ .15 & -.27 & .36 & .09 \\ .12 & .04 & 0 & .36 \end{bmatrix}$$

we find that $N(D^2) = \sqrt{.6093} = .7806$, and since this is less than unity we are assured that the process will converge. We may write immediately, without use of a machine or written side calculation:

$$C_1 = C_0 + C_0D = \begin{bmatrix} 2.0 & -.1 & -.9 & -1.1 \\ & 1.0 & .1 & -.4 \\ & & 1.0 & .3 \\ & & & 1.4 \end{bmatrix}.$$

Utilizing the value of D^2 already determined, we readily find

$$C_2 = C_1 + C_1D^2 = \begin{bmatrix} 2.032 & -.138 & -.848 & -1.056 \\ & 1.138 & -0.42 & -.372 \\ & & 1.182 & .274 \\ & & & 1.558 \end{bmatrix}.$$

From this point on a machine is needed for efficiency. The next step is to calculate D^4, either by squaring D^2 or by the formula $D^4 = 1 - AC_2$; both methods may be used as a check. The result is:

$$D^4 = 10^{-4} \times \begin{bmatrix} 808 & -730 & 1094 & 1330 \\ 20 & 786 & -830 & 890 \\ 846 & -1560 & 1998 & 540 \\ 616 & 80 & 152 & 1696 \end{bmatrix}.$$

We may now consider the accuracy of further approximations, inserting in (7.5) $N(C_2) = \sqrt{13.385572} = 3.659$ in place of $N(C_0)$, $m - 2$ for m, and $k = N(D^4) = .4119$. Thus

$$\begin{aligned} N(E_2) &\leq (9.8807)(.4119) &&= 4.0699 \\ N(E_3) &\leq (9.8807)(.4119)^2 &&= 1.6764 \\ N(E_4) &\leq (9.8807)(.4119)^4 &&= .2844 \\ N(E_5) &\leq (9.8807)(.4119)^8 &&= .00819 \\ N(E_6) &\leq (9.8807)(.4119)^{16} &&= .000\,006\,79. \end{aligned}$$

Because of the roughness of the initial approximation in this case the convergence is rather slow at first, but later it is much accelerated. So far as the limits found above show, five iterations are necessary to be sure of even approximate two-place accuracy in the results (somewhat better limits could be obtained after actually calculating C_2, still better ones from C_3, etc.), but the sixth iteration gives results sure to be accurate nearly to five places. Perhaps the best treatment of a numerical case of this kind is to work out the solution by Dwyer's method to two, three or four places, and then to apply the iterative process once, and as many more times as necessary to obtain the required accuracy.

The final step should, for the sake of checking, be a calculation of C_{m+1} from $C_m(2 - AC_m)$, rather than from $C_m + C_mD^{2^m}$.

Upon observing that $N(D) > 1$ we might have used the Seidel process to improve each row of C_0. This process is however extremely slow, and in the present example is markedly inferior to that used above.

(iii) If we start from the result which Dwyer gives to four decimal places as C_0, we obtain

$$D = 1 - AC_0 = 10^{-5} \times \begin{bmatrix} 1 & 4 & -3 & -2 \\ 3 & -2 & 1 & 0 \\ -3 & 3 & -1 & 1 \\ 0 & 2 & 0 & -2 \end{bmatrix}.$$

We find $N(C_0) = 3.8188$, and putting $k = N(D) = .00085$ we have from (7.5),

$$N(E_m) \leq 3.8188\ (.00085)^{2^m}(2.00085) \leq (7.6408)(.00085)^{2^m}.$$

Thus $N(E_1) \leq .000\,0055$,

$N(E_2) \leq .000\,000\,000\,0004$.

9. Certain other methods of successive approximation. A class of methods for solving linear equations, which may be extended to find the inverse of a matrix, is given by Frazer, Duncan and Collar [12, pp. 132–133], generalizing a method of J. Morris. In this method the matrix A of the coefficients in the linear equations, or the matrix to be inverted, is written as the sum of an easily inverted matrix V, for example a diagonal or triangular matrix, and another

matrix W. Then

$$A^{-1} = (1 + V^{-1}W)^{-1}V^{-1} = (1 - f)^{-1}V^{-1},$$

where $f = -V^{-1}W$. If the latent roots of f are all less than unity in absolute value, and *a fortiori* if $N(f) < 1$, the series

$$1 + f + f^2 + f^3 + \cdots = (1 - f)^{-1}$$

converges. To solve the equations $AX = G$, where X and G are column vectors (i.e. matrices of one column) is to determine

$$X = A^{-1}G = (1 - f)^{-1}H,$$

where $H = V^{-1}G$. The method of Frazer, Duncan and Collar is to calculate the successive vectors

$$X_0 = H, \quad X_1 = H + fX_0, \quad X_2 = H + fX_1, \cdots, X_r = H + fX_{r-1}, \cdots.$$

It is clear that

$$X_r = (1 + f + f^2 + \cdots + f^r)H.$$

The error in X_r is therefore the vector

$$E_r = f^{r+1}(1 - f)^{-1}H.$$

We may ascertain a limit for the errors if $N(f) \leq k < 1$. Indeed, by (6.3) and (6.10),

$$N(E_r) \leq k^{r+1}\left(p^{1/2} - 1 + \frac{1}{1-k}\right)N(H),$$

where p is the number of unknowns; and no individual unknown will have an error greater than $N(E_r)$.

Convergence of this method, if existent, may be accelerated by matrix-squaring. Indeed, upon calculating in turn $f^2, f^4, f^8, f^{16}, \cdots$ by repeated squarings, we need only to work with the sequence

$$X_0 = H, \quad X_1 = (1 + f)X_0, \quad X_2 = (1 + f^2)X_1,$$

$$X_7 = (1 + f^4)X_3, \quad X_{15} = (1 + f^8)X_7, \cdots,$$

omitting the intermediate approximations. This will be worth while for solving a single set of equations only in case such great accuracy is required as to demand the use of rather high powers of f. Each squaring of f consists of the formation of p^2 sums of products, so that determination of, say, X_{31} by this method requires $4p^2$ such sums after f has been determined, in addition to the $5p$ involved in finding X_1, X_3, X_7, X_{15}, X_{31} after the squarings. By the method of Frazer, Duncan and Collar the corresponding number of sums of products would be $31p$. Since $4p^2 + 5p < 31p$ only in case $p \leq 6$, it appears that the matrix-squaring is justified only for six or fewer unknowns unless a larger number of terms is required. Furthermore, increasingly high powers of a matrix, to be useful, need usually to be expressed with more and more significant digits.

If more than one system of equations with the same matrix A is to be solved, these methods have the advantage that the same matrix f can be used for all the vectors G of right-hand members. In such cases the value of matrix-squaring is enhanced in comparison with that in which only a single system of equations is to be solved. Determination of A^{-1} is equivalent to solving p such systems in which the several column vectors G together constitute the identical matrix. If more than p of these systems of equations are to be solved it is best to find A^{-1} and then form the various solutions $A^{-1}G$ from the columns G of right-hand members.

It is worth noticing that the matrices $1 + f$, $1 + f^2$, etc., are commutative, as are all rational functions of a single matrix. In difficult cases this may occasionally provide a useful check.

This method differs from the other iterative methods with which we are concerned in that errors of calculation are not automatically corrected by it. This is a serious disadvantage, especially for the inexperienced computer, and makes desirable the careful maintenance of a check column. On the other hand, it does not require any preliminary knowledge of the solution. Indeed, it should be classified rather with the direct than with the iterative procedures on this account.

The critical element in determining the success of this method is the possibility or impossibility of finding suitable matrices V and W, such that V^{-1} can be calculated easily, and such that the elements of $f = -V^{-1}W$ are sufficiently small to make the roots all numerically less than unity. Morris uses for V the matrix derived from A by replacing all the elements above the principal diagonal by zeros. This insures that the corresponding positions in V^{-1} are also occupied by zeros. The other elements of V^{-1} are then determined fairly easily. If the non-diagonal elements of A, which appear in W, are sufficiently small, this fact will insure small enough elements in f to make convergence rapid.

A second method, given by Frazer, Duncan and Collar, chooses for V a diagonal matrix (one having only zero elements except in the principal diagonal), or simply the unit matrix. This choice reduces the labor of inversion to a minimum. Successful convergence will take place when the non-diagonal elements of A are sufficiently small in comparison with those in the diagonal, if V is taken as the diagonal matrix containing the diagonal elements of A.

A third method which may be useful in certain cases, particularly when some but not all of the unknowns are required, is the following. Let A be partitioned:

$$A = \left[\begin{array}{c|c} a & b \\ \hline c & d \end{array}\right],$$

where a and d are square submatrices which, being of lower order than A, are more easily inverted. Let V and W be the correspondingly partitioned matrices

$$V = \left[\begin{array}{c|c} a & 0 \\ \hline 0 & d \end{array}\right], \qquad W = \left[\begin{array}{c|c} 0 & b \\ \hline c & 0 \end{array}\right].$$

Putting $s = a^{-1}b$, $t = d^{-1}c$, we have:

$$f = -\left[\begin{array}{c|c} 0 & s \\ \hline t & 0 \end{array}\right], \qquad f^2 = \left[\begin{array}{c|c} st & 0 \\ \hline 0 & ts \end{array}\right], \qquad f^4 = \left[\begin{array}{c|c} (st)^2 & 0 \\ \hline 0 & (ts)^2 \end{array}\right], \cdots .$$

If only the first q of the p unknowns are required, a and b may be taken as matrices of q rows. If G_1 and H_1 consist respectively of the first q rows of G and H, and if G_2 and H_2 consist of the remaining rows, then $H_1 = a^{-1}G_1$ and $H_2 = d^{-1}G_2$. Then, in case of convergence, the first q rows of the solution are given by

$$(1 + st)[1 + (st)^2][1 + (st)^4] \cdots H_1 - s(1 + ts)[1 + (ts)^2][1 + (ts)^4] \cdots H_2 .$$

Convergence to the correct values is assured here if the norm of any power of st is less than unity, as is true if and only if the absolute values of all the roots of st are less than unity. This is easily seen to be true, since as m increases

$$\lim (ts)^m = t[\lim (st)^{m-1}]s.$$

10. A simple iterative method of solving equations. An entirely different method, whose convergence is independent of the initial trial values, is the following. To solve for the column vector X the equation $AX = G$, we may start with an arbitrary column of trial values X_0 and a scalar constant h, and then for $m = 1, 2, \cdots$ calculate X_m from

$$X_m = hG + (1 - hA)X_{m-1} .$$

If X_m is equal to X_{m-1} it is obviously the desired solution. Otherwise there is an error

$$\begin{aligned} X_m - X &= (hG - X) + (1 - hA)X_{m-1} = (hA - 1)X + (1 - hA)X_{m-1} \\ &= (1 - hA)(X_{m-1} - X) = \cdots = (1 - hA)^m(X_0 - X). \end{aligned}$$

This converges to zero as m increases provided the latent roots of $1 - hA$ are all less than unity in absolute value. If A has only real roots this is equivalent to requiring that they all be between 0 and $2/h$. In particular, if A is a correlation matrix, its roots are all real and positive. Since their sum $= tr(A) = p$, where p is the number of rows, all roots of A lie between 0 and p. Consequently the process will converge in this case if $0 \leq h \leq 2/p$. It is desirable, in order to make the error diminish as fast as possible, to take h as large as is consistent with convergence. In some cases a lower limit than p will be known for the greatest root of A, and then a smaller value than $2/p$ can be taken for h. A limit of error is obviously set by

$$N(X_m - X) \leq [N(1 - hA)]^m N(X_0 - X).$$

This method can of course be applied to find the inverse matrix.

It can also be accelerated by matrix-squaring. If we put $D = 1 - hA$ we have for example,

$$X_8 = (1 + D)(1 + D^2)(1 + D^4)hG + D^8X_0 .$$

The last term will approach zero in case of convergence, and may be omitted in this type of calculation.

Thus accelerated, the method gives decimal places of accuracy increasing in geometric instead of arithmetic progression, and is remarkably simple and straightforward. It is at its best when the roots of A are known to be closely clustered about unity. A criterion of this is that $\Sigma(\lambda_i - \bar{\lambda})^2$ shall have a small value, where $\bar{\lambda}$ is the mean of the p roots λ_i. This sum of squares equals $\Sigma\lambda_i^2 - p$ for a correlation matrix A, and $\Sigma\lambda_i^2 = tr(A^2) = \Sigma\Sigma a_{ij}^2 = p + 2\sum_{i<j} a_{ij}^2$, so that

$$\sum (\lambda_i - \bar{\lambda})^2 = 2\sum_{i<j} a_{ij}^2 .$$

Smallness of this quantity is favorable not only to this iterative method but also to those of §§4 and 5.

11. Use of the characteristic equation for inversion and for finding determinants. A method differing greatly from the others is based on the Cayley-Hamilton theorem that every matrix satisfies its own characteristic equation [29, p. 23; 4, p. 296]. This is the equation

$$f(\lambda) = \begin{vmatrix} a_{11} - \lambda & a_{12} & \cdots & a_{1p} \\ a_{21} & a_{22} - \lambda & \cdots & a_{2p} \\ \cdots\cdots & \cdots\cdots & \cdots & \cdots\cdots \\ a_{p1} & a_{p2} & \cdots & a_{pp} - \lambda \end{vmatrix}$$

$$= e_p - e_{p-1}\lambda + e_{p-2}\lambda^2 - \cdots + (-)^{p-1}e_1\lambda^{p-1} + (-)^p\lambda^p = 0,$$

where e_r $(r = 1, 2, \cdots, p)$ is the sum of the products r at a time of the roots, and also equals the sum of the r-rowed principal minors of the matrix A. Substituting A for λ, which by the Cayley-Hamilton theorem is legitimate, multiplying by A^{-1}, and transposing a term, yields

$$(11.1)\quad e_pA^{-1} = e_{p-1} - e_{p-2}A + e_{p-3}A^2 - \cdots + (-)^p e_1A^{p-2} + (-)^{p+1}A^{p-1}.$$

This equation provides a direct method of calculating A^{-1} as soon as the elementary symmetric functions e_r of the roots of $f(\lambda) = 0$ have been evaluated. This evaluation may be accomplished by means of Newton's identities [4, p. 243] connecting the elementary symmetric functions with the power-sums. If s_r is the sum of the rth powers of the roots, these formulae give:

$$e_1 = s_1$$
$$e_2 = \tfrac{1}{2}(e_1 s_1 - s_2)$$
$$e_3 = \tfrac{1}{3}(e_2 s_1 - e_1 s_2 + s_3)$$
$$\cdots\cdots\cdots\cdots\cdots$$
$$e_p = \frac{1}{p}(e_{p-1} s_1 - e_{p-2} s_2 + \cdots \pm s_p).$$

The procedure is to calculate in turn $A^2, A^3, \cdots, A^{p-1}$, then to obtain the s's from the diagonals of these matrices, since $s_r = tr(A^r)$, then to obtain the elementary symmetric functions $e_1, \cdots, e_{p-1}$ of the roots from Newton's formulae, and to substitute these in the right-hand member of (11.1). It is then only necessary to find and divide by e_p, which equals the determinant of A. For this, and for checking the calculations, there is a choice of methods. We may find the diagonal of A^p, without troubling to compute the whole of this matrix, from the product AA^{p-1} and also, to provide a comprehensive check, from $A^{p-1}A$ or possibly from the product of two powers of A of exponents approximating $p/2$. The sum of these diagonal elements of A^p is s_p, which may be substituted in the last of the Newton formulae above with the quantities previously found to give e_p. An alternative method is to multiply A by its adjoint e_pA^{-1}, which is computed by (11.1), to obtain the determinant e_p.

The total number of multiplications, divisions, and additions is distinctly greater by this method than by efficient direct methods such as that of Dwyer [7, 9]. On the other hand, this method is straightforward and easily checked; the divisions involved are of the simplest character, consisting only of the divisions by $2, 3, \cdots, p$ in Newton's formulae and of the final division of the adjoint matrix by one number; and for large matrices it is ideally adapted for matrix multiplication by means of punched cards. A further very important advantage of this characteristic function method is that it yields considerable additional information as a by-product. Not only the determinant of the matrix but the sums s_r of the principal minors of each order r are determined. Moreover the characteristic equation, whose coefficients would be exceedingly difficult to compute directly from definitions for a large matrix, is by this method made available for the study of the latent roots, which have great interest in themselves for numerous purposes.

The characteristic function method is applicable whether A is symmetric or not. If it is symmetric, the same is true of each of the other matrices appearing in the calculation, so that it is necessary to write only about half the elements.

An illustration using a symmetric matrix has been given by M. D. Bingham [3]. In the illustration below the matrix is not symmetric and has complex double roots and non-linear elementary divisors, so that evaluation of the roots by iterative methods, though possible, would be very slow and laborious, as shown by Aitken [2]. This is indeed the same example used by Aitken in this discussion. But it should be noted that the associated latent vectors, which are determined along with the roots in the iterative processes, require the solution of sets of $p - 1$ linear equations if the roots are found directly by solving the characteristic equation.

$$\text{Let } A = \begin{bmatrix} 15 & 11 & 6 & -9 & -15 \\ 1 & 3 & 9 & -3 & -8 \\ 7 & 6 & 6 & -3 & -11 \\ 7 & 7 & 5 & -3 & -11 \\ 17 & 12 & 5 & -10 & -16 \end{bmatrix}.$$

Then

$$A^2 = \begin{bmatrix} -40 & -9 & 105 & -9 & -40 \\ -76 & -43 & 32 & 44 & 23 \\ -55 & -22 & 62 & 20 & -10 \\ -61 & -25 & 65 & 20 & -7 \\ -40 & -9 & 110 & -14 & -40 \end{bmatrix}, \quad A^3 = \begin{bmatrix} -617 & -380 & 64 & 499 & 256 \\ -260 & -189 & -316 & 355 & 280 \\ -443 & -279 & -106 & 415 & 259 \\ -464 & -300 & -136 & 439 & 292 \\ -617 & -385 & 69 & 499 & 256 \end{bmatrix},$$

$$A^4 = (A^2)^2 = \begin{bmatrix} -1342 & -978 & -2963 & 2444 & 2006 \\ 944 & 522 & -1982 & -10 & 503 \\ -358 & -333 & -2435 & 1307 & 1334 \\ -175 & -243 & -2645 & 1247 & 1355 \\ -1312 & -963 & -2978 & 2444 & 1991 \end{bmatrix} = AA^3 \text{ (check).}$$

From the diagonals of these matrices,

$$s_1 = 5, \quad s_2 = -41, \quad s_3 = -217, \quad s_4 = -17.$$

Calculating the sum of the diagonal elements only of A^5 (on a machine, without listing them separately) from AA^4 and also, as a check, from A^2A^3 we find $s_5 = 3185$. Newton's formulae then give

$$e_1 = 5, \quad e_2 = 33, \quad e_3 = 51, \quad e_4 = 135, \quad e_5 = -225,$$

the last value being that of the determinant of A. We readily find from (11.1),

$$A^{-1} = -\frac{1}{225}\begin{bmatrix} -207 & 64 & -124 & 111 & 171 \\ -315 & 30 & 195 & -180 & 270 \\ -315 & 30 & -30 & 45 & 270 \\ -225 & 75 & -75 & 0 & 225 \\ -414 & 53 & 52 & -3 & 342 \end{bmatrix}.$$

So far, all results by this method are exact, but the division by 225 introduces recurring decimals and therefore a limited validity for the form

$$A^{-1} = \begin{bmatrix} .9200 & -.2844 & .5511 & -.4933 & -.7600 \\ 1.4000 & -.1333 & -.8667 & .8000 & -1.2000 \\ 1.4000 & -.1333 & .1333 & -.2000 & -1.2000 \\ 1.0000 & -.3333 & .3333 & 0 & -1.0000 \\ 1.8400 & -.2356 & -.2311 & .0133 & -1.5200 \end{bmatrix}$$

The characteristic equation

$$f(\lambda) = \lambda^5 - 5\lambda^4 + 33\lambda^3 - 51\lambda^2 + 135\lambda + 225 = 0$$

may in this case be solved readily, since

$$f(\lambda) = (\lambda + 1)(\lambda^2 - 3\lambda + 15)^2.$$

III. Latent Roots and Vectors

12. Direct and iterative methods. If the latent roots but not the latent vectors of a matrix are desired, as for example in a preliminary study of vibra-

tions in machinery being designed, where the important question is whether any root has a positive real part, it is only necessary to find the characteristic equation and to work with it by the methods of the theory of equations. The coefficients in the characteristic equation are the sums of the r-rowed principal minors ($r = 1, 2, \cdots, p$), and are expeditiously found directly from this definition for matrices of four or fewer rows. For large matrices, however, the calculation of so many large overlapping determinants is wasteful of effort, since many virtually equivalent calculations must be done repeatedly. Indeed, calculation by determinants in a great many situations, including the solution of linear equations, is open to this objection. The methods of §11 yield the characteristic function in a manner which, for large matrices, appears to be the best available, excepting perhaps the new method of Samuelson [25a].

When, as is commonly the case, the latent vectors are desired, a straightforward calculation directly from the definitions would require not only setting up and solving the characteristic equation, but also the solution, in the case of each root, of the set of linear equations in p unknowns whose matrix is obtained from the characteristic matrix by substituting the particular root for λ. It is this solution of linear equations that aggravates greatly the computational labor when direct methods are used.

An ingenious method has been used by R. A. Fisher [11, pp. 299 ff.]. Starting with a four-rowed determinant whose elements are linear functions of an unknown θ, Fisher calculates the value of the determinant for selected values of θ, and then by interpolation using divided differences finds the largest value of θ making the determinant zero. The point of the divided difference method is that it avoids the direct calculation of the determinant for more than a few values of θ, replacing it essentially by calculation of the fourth-degree polynomial in θ from its differences and using the fact that the fourth divided differences are constant. The linear equations are then solved in a direct manner. If applied to large matrices this would be very laborious, but it compares favorably with calculation directly from definitions in the manner suggested by reading books on algebra and solid analytic geometry. But even with large matrices Fisher's method may perhaps be the best in certain cases, e.g. if all that is desired is the root of median absolute value and if this root is real, or if it is desired to find a few real roots that are close together, with numerous others greater and another numerous group less than these. This is because the iterative methods give the real roots in the order of their absolute values, beginning with the greatest, but with the possibility of obtaining them in the opposite order by first inverting the matrix. The Mallock electrical device [22] may be used to calculate determinants, and thus to apply this method.

If A and B are p-rowed matrices and B is non-singular, the determinantal equation $|A - \lambda B| = 0$ is equivalent to $|AB^{-1} - \lambda| = 0$ and also to $|B^{-1}A - \lambda| = 0$. The column vectors X_i satisfying $(A - \lambda B)X = 0$ also satisfy $(B^{-1}A - \lambda)X = 0$ and the row vectors V_i satisfying $V_i(A - \lambda_i B) = 0$ also satisfy $V_i(AB^{-1} - \lambda_i) = 0$. If A and B are symmetric, $V_i = X_i'$. Thus

any problem of this type is reducible to that of finding latent roots and vectors, upon calculating B^{-1} by any method and multiplying in either order by A.

The fundamental iterative method for finding latent roots and vectors of A begins with an arbitrary matrix X_0 of a single column. This column vector is premultiplied by A to obtain a new column vector X_1. If, as is possible though unlikely, the elements of X_1 are proportional to those of X_0, they constitute one of the latent vectors of A, and the factor of proportionality is the corresponding root, for then X_0 and X_1 are solutions of the matrix equation $(A - \lambda)X = 0$. It should be observed that the latent vector is determined only to within an arbitrary scalar factor of proportionality, though we may sometimes find it convenient to normalize the vector by choosing the factor in such a way that the sum of the squares of the elements, which equals the square of the norm, is unity.

If X_1 is not proportional to X_0, the operation may be repeated by calculating $X_2 = AX_1$, then $X_3 = AX_2$, and so on. If these vectors are then normalized, or if they are divided by, say, their respective first elements, then the other elements will (in the cases of greatest practical importance) gradually approach stable values which will determine one of the latent vectors, while the successive factors of proportionality will approach the corresponding root. The convergence of this process is however apt to be rather slow. Fortunately there are several known ways of accelerating it.

Matrix-squaring is the first of these methods of accelerating convergence [17, 19]. It is clear that $X_t = A^t X_0$. Consequently one application of the iterative process with A^t is equivalent to t iterations with A. It is relatively easy to square A, and then by repeated squarings to form A^4, A^8, A^{16}, etc. The economical limit of this process is determined partly by the necessity of retaining more and more digits in the successively higher powers, but up to a point not yet determined exactly it presents very great advantages. For proceeding to the determination of latent roots of other than the maximum absolute value, with their associated vectors, this method lends itself to further shortcuts [17, 2], which seem to give it an advantage over an older method [13].

Another method of accelerating convergence, introduced by A. C. Aitken, and referred to by him as the δ^2-method, uses the ratio $\phi(t)$ of an element of X_{t+1} to the corresponding element of X_t in the function

$$\frac{\phi(t+1)\phi(t-1) - [\phi(t)]^2}{\phi(t+1) - 2\phi(t) + \phi(t-1)},$$

which converges rapidly toward the root λ_1 of greatest absolute value. If a constant c is subtracted from all three of the quantities $\phi(t+1)$, $\phi(t)$ and $\phi(t-1)$ before computing the foregoing function the result is unchanged. This fact reduces greatly the computational labor, since the decimal places of λ_1 already determined are common to all three.

If A is symmetric and we form the scalar products of $X_t = A^t X_0$ with itself

and with X_{t+1} we have

$$X_t'X_t = X_0'A^{2t}X_0, \qquad X_{t+1}'X_t = X_0'A^{2t+1}X_0.$$

The ratio of these two scalars gives an estimate of λ_1 which on the basis of the ratios of consecutive elements in a given place in the trial vectors would not be reached until a later stage of convergence, corresponding in fact to twice as many iterations. Aitken has pointed out the great value of this procedure for finding the root (but not the latent vector), and has extended the idea to asymmetric matrices, where there is a complication because of the existence of two latent vectors for each root, one determined by premultiplying by A, the other by A'.

The comprehensive paper [2] of Aitken gives an extremely valuable account of the whole problem and processes of finding the latent roots and vectors, including a survey of the various cases arising when there are multiple roots, complex roots, and non-linear elementary divisors. This paper should be studied carefully by anyone with any substantial numerical problem of this kind.

A method using rotations of two variables at a time has been devised by T. L. Kelley [21].

The remainder of this paper will be concerned with some results, believed to be new, by which useful upper limits can be set for the errors of the results yielded by iteration for latent roots and vectors of a symmetric matrix. To find such limits of error for asymmetric matrices appears to be a much more difficult and as yet unsolved problem.

13. Accuracy of iteration with symmetric matrices. If A is symmetric, as it is in most statistical problems (though with some exceptions, as in [18]), the roots are all real and the elementary divisors are linear. Moreover there exist an orthogonal matrix H and a diagonal matrix

$$\Lambda = \begin{bmatrix} \lambda_1 & 0 & 0 & \cdots \\ 0 & \lambda_2 & 0 & \cdots \\ 0 & 0 & \lambda_3 & \cdots \\ \cdots & \cdots & \cdots & \cdots \end{bmatrix}$$

such that

$$A = H\Lambda H'. \tag{13.1}$$

Since H is orthogonal, $HH' = 1$, and therefore

$$\Lambda = H'AH, \qquad A^t = H\Lambda^t H'. \tag{13.2}$$

We may associate with the successive trial vectors $X_t = AX_{t-1} = A^tX_0$ the vectors $Y_t = H'X_t$; then $X_t = HY_t$. From these equations and the second of (13.2) it is clear that

$$Y_t = H'X_t = H'A^tX_0 = \Lambda^tH'X_0 = \Lambda^tY_0.$$

Hence, if the elements of Y_0 are $y_1, \cdots, y_p$,

$$Y_t = \begin{bmatrix} y_1\lambda_1{}^t \\ y_2\lambda_2 \\ \cdot \\ \cdot \\ \cdot \\ y_p\lambda_p{}^t \end{bmatrix}.$$

Now let α_{kt} be the scalar product of X_t and X_{t-k}:

$$\alpha_{kt} = X_t'X_{t-k} = Y_t'Y_{t-k} = \Sigma y_i^2\lambda_i^{2t-k}; \tag{13.3}$$

and let

$$\nu_{kt} = \frac{\alpha_{kt}}{\alpha_{0t}} = \frac{\Sigma y_i^2\lambda_i^{2t-k}}{\Sigma y_i^2\lambda_i^{2t}}. \tag{13.4}$$

If A has a negative root this fact will become evident after a certain stage in the iteration used to obtain this root by an alternation of sign of the numbers in any one position in consecutive trial vectors. However A^2, which as pointed out in §12 may well be calculated anyhow, has only positive roots, which are the squares of the roots of A, and has the same latent vectors as A. Hence we shall have results of sufficient generality for real symmetric matrices if we assume that all roots of the matrix with which we work are positive or zero, i.e. that it is positive definite or semi-definite. Let us choose the notation so that

$$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0.$$

Then if $k \geq 0$,

$$\alpha_{0t} = \Sigma y_i^2\lambda_i^{2t} \leq \lambda_1^k \Sigma y_i^2\lambda_i^{2t-k} = \lambda_1^k\alpha_{kt}.$$

Hence, by (13.4),

$$\lambda_1 \geq [\nu_{kt}]^{-1/k}. \tag{13.5}$$

It is known [23] that if $a_1, \cdots, a_p, c_1, \cdots, c_p$ are any positive numbers, the function

$$\left(\frac{c_1a_1^k + \cdots + c_pa_p^k}{c_1 + \cdots + c_p}\right)^{1/k}$$

increases monotonically with k. Putting $c_i = y_i^2\lambda_i^{2t}$, $a_i = \lambda^{-k}$ if $\lambda_i \neq 0$, and $c_i = a_i = 0$ if $\lambda_i = 0$, we find that the right-hand member of (13.5) decreases monotonically as k increases. Hence the best of these lower bounds for λ_1 is that corresponding to the least value of k that can be used, namely $k = 1$. Consequently the lower bound to be recommended for λ_1 is given by

$$\lambda_1 \geq \frac{1}{\nu_{1t}}. \tag{13.6}$$

From (13.4) it is easily seen that this lower bound approaches λ_1 when t increases, provided $y_1 \neq 0$.

An upper bound for λ_1 is available from the fact that the sum of the tth powers of the roots is the trace of A^t. Since we assume all $\lambda_i \geq 0$ this gives

$$\lambda_1 \leq (\text{tr } A^t)^{1/t}.$$

That this upper limit converges to λ_1 when t increases is easily seen from (6.7) upon consideration of $\log (\Sigma\lambda_i^t)^{1/t}$.

A lower limit alternative to that of (13.6) is also available from tr (A^t), and likewise converges to λ_1. Indeed, since λ_1 is the greatest root, we have

$$\lambda_1 \geq (\text{tr } A^t/p)^{1/t}.$$

We now seek limits of accuracy for the latent vector corresponding to λ_1 and estimated by X_t. If we call this vector X, and define $Y = H^{-1}X$, then $\lim Y_t^* = Y^*$, where Y^* is the normalized form of Y. Y^* has as its ith element

$$\lim_{t\to\infty} \frac{y_i\lambda_i^t}{\sqrt{\Sigma y_j^2\lambda_j^{2t}}}.$$

If $y_1 \neq 0$, and $\lambda_1 > \lambda_2 \geq \lambda_3$, this limit is ± 1 if $i = 1$, and is otherwise 0. We take the value of Y to be

$$\begin{bmatrix} 1 \\ 0 \\ 0 \\ \cdot \\ \cdot \\ 0 \end{bmatrix}$$

in this case. If λ_1 is a multiple root, the limit of X_t^* will depend on the initial values y_i.

A useful measure of the closeness of approach of X_t to X is the "correlation coefficient" $r_t = X^{*\prime}X_t^* = Y^{*\prime}Y_t^* = \dfrac{y_1\lambda_1^t}{\sqrt{\Sigma y_j\lambda_j^{2t}}}$,

which obviously approaches unity as t increases if $y_1 \neq 0$ and λ_1 is a *simple* root, or if $\lambda_1 = \lambda_2 = \cdots = \lambda_s > \lambda_{s+1}$ and we arrange our definitions so that $y_1 \neq 0$ and $y_2 = \cdots = y_s = 0$. In terms of the notation previously introduced, $r_t = \dfrac{y_1\lambda_1^t}{\sqrt{\alpha_{ot}}}$. The sum of the squares of the differences of corresponding elements of the normalized vectors X^* and X_t, i.e. $[N(X^* - X_t^*)]^2$, is $2(1 - r_t)$, and therefore approaches zero as r_t approaches unity. We shall seek for r_t a lower limit approaching unity as t increases.

Let us now put

$$w_{it} = \frac{y_i^2\lambda_i^{2t}}{\Sigma y_j^2\lambda_j^{2t}} = \frac{y_i^2\lambda_i^{2t}}{\alpha_{ot}}. \qquad \text{Then } r_t^2 = w_{1t} \text{ and } \sum_{i=1}^{p} w_{it} = 1.$$

For $k \geq 1$,

$$\alpha_{kt} = \Sigma y_i^2 \lambda_i^{2t-k} \geq y_2^2 \lambda_2^{2t-k} + \cdots + y_p^2 \lambda_p^{2t-k} \geq \lambda_1^{-k}(y_2^2 \lambda_2^{2t} + \cdots + y_p^2 \lambda_p^{2t})$$
$$= \lambda_1^{-k} \alpha_{ot}(w_{2t} + \cdots + w_{pt}) = \lambda_1^{-k} \alpha_{ot}(1 - w_{1t}) = \lambda_1^{-k} \alpha_{ot}(1 - r_t).$$

$$\therefore r_t^2 \geq 1 - \frac{\lambda_1^k \alpha_{kt}}{\alpha_{ot}} = 1 - \nu_{kt} \lambda_1^k.$$

This unfortunately is not a very useful lower bound for r_t, since it approaches *zero*, not unity, as t increases.

A more satisfactory result is obtained as follows. Let $\eta_i = \lambda_i^{-1}$. Then $\nu_{kt} = \sum_{i=1}^{p} w_{it} \eta_i^k$. For any value of t we may consider a distribution of a variate taking the positive values $\eta_1, \cdots, \eta_p$ with the positive weights, or probabilities, w_{it}. The kth moment of this distribution about 0 is ν_{kt}. In particular the first moment is ν_{1t}, and is evidently at least equal to η_1, which is the least of the η_i. The standard deviation is $\sigma = \sqrt{\nu_{2t} - \nu_{1t}^2}$. As t increases, ν_{1t} will approach η_1 and σ will approach zero. Hence, if $\lambda_1 > \lambda_2$, a stage will eventually be reached at which $\nu_{1t} < \eta_2$. Let

$$k = \frac{\eta_2 - \nu_{1t}}{\sigma}.$$

By the Tchebychef-Bienaymé inequality,

$$w_{2t} + \cdots + w_{pt} \leq \frac{1}{k^2},$$

and therefore

$$r_t^2 = w_{1t} \geq 1 - \frac{\nu_{2t} - \nu_{1t}^2}{(\eta_2 - \nu_{1t})^2},$$

provided t is large enough so that $\nu_{1t} < \eta_2$. This lower bound approaches unity, as desired, when t increases.

If $\lambda_1 = \lambda_2, \cdots = \lambda_k > \lambda_{k+1}$, the same proof shows that

$$w_1^{(+)} + \cdots + w_k^{(+)} \geq 1 - \frac{\nu_{2t} - \nu_{1t}^2}{(\eta_{r+1} - \nu_{1t})^2},$$

provided $\nu_{1t} < \eta_{r+1}$. The left member is the correlation of X_t with that one of the k-parameter family of latent vectors corresponding to the multiple root for which the correlation is a maximum.

In order to utilize these results we need a lower bound for η_2, or for η_{r+1}. In case $\lambda_1 \neq \lambda_2$ this requires an upper bound for λ_2. Such an upper bound may be found at the next stage through working with the reduced or "deflated" matrix used in [17]. This is $A_1 = A - \lambda_1 XX'$, where X is the normalized latent vector corresponding to λ_1; and $\lambda_2^t \leq \text{tr}\,(A_1^t)$.

Since we have arrived at a definite lower limit for r_t which approaches unity

as the iterative process proceeds, and since we have found for λ_1 upper and lower bounds converging to it, a solution has been found for the troublesome problem of the degree of accuracy in stopping at any stage of the iteration for finding the greatest root and the associated latent vector. It would be possible to go on to find from these results appropriate inequalities for A_1, and then by repetition of the above arguments, for λ_2 and the second latent vector; and then likewise for the second reduced matrix A_2 and the further roots, vectors, and reduced matrices in this cyclic order. These steps may well be taken by the computer who has mastered the above argument in connection with a numerical example.

BIBLIOGRAPHY

[1] A. C. AITKEN, "Studies in practical mathematics. I. The evaluation, with applications, of a certain triple matrix product," *Proc. Roy. Soc. Edinburgh*, Vol. 57 (1937), pp. 172–181.

[2] A. C. AITKEN, "Studies in practical mathematics, II. The evaluation of the latent roots and latent vectors of a matrix," *Proc. Roy. Soc. Edinburgh*, Vol. 57 (1937), pp. 269–304.

[3] M. D. BINGHAM, "A new method for obtaining the inverse matrix," *Jour. Amer. Stat. Assoc.*, Vol. 36 (1941), pp. 530–534.

[4] M. BÔCHER, *Introduction to Higher Algebra*, New York, 1907, 321 pp.

[5] R. C. BOSE and S. N. ROY, "The distribution of the Studentized D²-statistic," *Sankhya*, Vol. 4 (1938), pp. 19–38.

[6] PRESCOTT D. CROUT "A short method for evaluating determinants and solving systems of linear equations with real or complex coefficients." *Trans. Am. Inst. Elec. Engineers*, Vol. 60 (1941).

[7] PAUL S. DWYER, "The solution of simultaneous equations," *Psychometrika*, Vol. 6 (1941), pp. 101–129.

[8] PAUL S. DWYER, "The Doolittle technique," *Annals of Math. Stat.*, Vol. 12 (1941), pp. 449–458.

[9] PAUL S. DWYER, "Recent developments in correlation technique," *Jour. Amer. Stat. Assoc.*, Vol. 37 (1942), pp. 441–460.

[10] R. A. FISHER, "The use of multiple measurements in taxonomic problems," *Annals of Eugenics*," Vol. 7 (1936), pp. 179–188.

[11] R. A. FISHER, *"Statistical Methods for Research Workers*, London, 7th ed., 1938, 256 pp.

[12] R. A. FRAZER, W. J. DUNCAN and A. R. COLLAR, *Elementary matrices and some applications to dynamics and differential equations*, Cambridge Univ. Press, 1938. 416 pp.

[13] R. VON MISES and HILDA POLLACZEK-GEIRINGER, "Zusammenfassende Berichte. Praktische Verfahren der Gleichungsauflösung." *Zeitschrift für angewandte Math. und Mechanik*, Vol. 9 (1929), pp. 58–77 and 152–164.

[14] HILDA GEIRINGER, "On the numerical solution of linear problems by group iteration," *Bull. Amer. Math. Soc.*, Vol. 48 (1942), p. 370.

[15] HAROLD HOTELLING, "The generalization of Student's ratio," *Annals of Math. Stat.*, Vol. 2 (1931), pp. 360–378.

[16] HAROLD HOTELLING, "Analysis of a complex of statistical variables into principal components," *Jour. Educ. Psych.*, Vol. 24 (1933), pp. 417–441 and 498–520. (Reprints available only from publishers, Warwick and York, Baltimore.)

[17] HAROLD HOTELLING, "Simplified calculation of principal components," *Psychometrika*, Vol. 1 (1936), pp. 27–35.

[18] HAROLD HOTELLING, "Relations between two sets of variates," *Biometrika*, Vol. 28 (1936), pp. 321–377.

[19] R. C. J. HOWLAND, "Note on a type of determinantal equation," *Phil. Mag.*, Series 7, Vol. 6 (1928), pp. 839–842.

[20] TRUMAN L. KELLEY and FRANK S. SALISBURY, "An iteration method for determining multiple correlation constants," *Jour. Amer. Stat. Assn.*, Vol. 21 (1926), pp. 282–292.

[21] TRUMAN L. KELLEY, *Essential traits of mental life*, Harvard, 1935, 145 pp. Chap. 1.

[22] *The Mallock electrical calculating machine.* Reprint from *Engineering* (London), June 22, 1934. 8 pp.

[23] NILAN NORRIS, "Inequalities among averages," *Annals of Math. Stat.*, Vol. 6 (1935), pp. 27–29.

[24] LOUIS A. PIPES, "Matrix theory of oscillatory networks," *Jour. Applied Physics*, Vol. 10 (1939), pp. 849–860.

[25] T. RELLA, "*Über den absoluten Betrag von Matrizen*," International Congress of Mathematicians at Oslo, 1936 (derives λ_1 as the absolute value from 5 postulates, which are proved independent).

[25a] P. A. SAMUELSON, "A method of determining explicitly the coefficients of the characteristic equation," *Annals of Math. Stat.*, Vol. 13 (1942), pp. 424–429.

[26] R. V. SOUTHWELL, *Relaxation methods in engineering science, a treatise on approximate computation*, Oxford Univ. Press, 1940. 252 pp.

[27] G. TEMPLE, "The general theory of relaxation methods applied to linear systems," *Proc. Roy. Soc. London*, Vol. 169 A (1939), pp. 476–500.

[28] A. WALD, "The classification of an individual into one of two groups." (Unpublished).

[29] J. H. M. WEDDERBURN, *Lectures on matrices*, New York, 1934. 200 pp.

[30] E. T. WHITTAKER and G. ROBINSON, *The calculus of observations*, London, 1924. 395 pp.

[31] S. S. WILKS, "Certain generalizations in the analysis of variance," *Biometrika*, Vol. 24 (1932), pp. 471–494.

Of interest in relation to the subject of this paper, though not mentioned in the text, are the following:

ARNOLD DRESDEN, "On the iteration of linear homogeneous transformations," *Bull. Amer. Math. Soc.*, Vol. 48 (1942), pp. 577–579 and 949.

PAUL HORST, "A method for determining the coefficients of a characteristic equation," *Annals of Math. Stat.*, Vol. 6 (1935), pp. 83–84.

A. T. LONSETH, "Systems of linear equations with coefficients subject to error," *Annals of Math. Stat.*, Vol. 13 (1942), pp. 332–337.

RUFUS OLDENBURGER. "Infinite powers of matrices and characteristic roots," *Duke Math. Jour.*, Vol. 6 (1940), pp. 357–361.