SIMPLIFIED CALCULATION OF PRINCIPAL COMPONENTS

HAROLD HOTELLING

The resolution of a set of n tests or other variates into components γ_n, each of which accounts for the greatest possible portion γ_1, $\gamma_2, \cdots$, of the total variance of the tests unaccounted for by the previous components, has been dealt with by the author in a previous paper (2). Such "factors," on account of their analogy with the principal axes of a quadric, have been called *principal* components. The present paper describes a modification of the iterative scheme of calculating principal components thère presented, in a fashion that materially accelerates convergence. The application of the iterative process is not confined to statistics, but may be used to obtain the magnitudes and orientations of the principal axes of a quadric or hyperquadric in a manner which will ordinarily be far less laborious than those given in books on geometry. This is true whether the quadrics are ellipsoids or hyperboloids; the proof of convergence given in an earlier paper is applicable to all kinds of central quadrics. For hyperboloids some of the roots k_i of the characteristic equation would be negative, while for ellipsoids all are positive. If in a statistical problem some of the roots should come out negative, this would indicate either an error in calculation, or that, if correlations corrected for attenuation had been used, the same type of inconsistency had crept in that sometimes causes such correlations to exceed unity.

Another method of calculating principal components has been discovered by Professor Truman L. Kelley, which involves less labor than the original iterative method, at least in the examples to which he has applied it (5). How it would compare with the present accelerated method is not clear, except that some experience at Columbia University has suggested that the method here set forth is the more efficient. It is possible that Kelley's method is more suitable when all the characteristic roots are desired, but not the corresponding correlations of the variates with the components. The present method seems to the computers who have tried both to be superior when the components themselves, as well as their contributions to the total variance, are to be specified. The advantage of the present method is enhanced when, as will often be the case in dealing with numerous variates, not all the characteristic roots but only a few of the largest are required.

Iterative processes of various kinds are capable of acceleration by means of the matrix-squaring device here used. In particular, the simultaneous determination of a linear function of a set of variates, and of another linear function of another set, such that the correlation of these two functions is a maximum, may be facilitated in this way. This problem of the most predictable criterion and the best predicter has been discussed briefly by the author, (3) and will be treated more fully in a forthcoming paper.

Let r_{ij} be the covariance of the ith and jth of a set of n variates $x_1, \cdots, x_n$; then if units have been chosen such that each standard deviation is unity, each $r_{ii} = 1$, and the r_{ij}'s are the correlations. If we take any arbitrary set of numbers $a_1, \cdots, a_n$ and substitute in the formula

$$a_i' = \sum_{j=1}^{n} r_{ij} a_j \;, \quad (i = 1, 2, \cdots, n) \;. \tag{1}$$

The new set of numbers $a_i', \cdots, a_n'$ will be proportional to the old if, and only if, they are also proportional to the correlations of one of the principal components with the original variates. These correlations are also the coefficients of the particular principal component in the equations which gives the x's in terms of the γ's. If $a_i' = ka_i$, for each i, then k is the sum of the squares of the correlations of the x's with the particular γ.

If the a_i' are not proportional to the a_i, they may be substituted in the right-hand members of (1), and will then give rise to another set of values $a_i'', \cdots, a_n''$, such that

$$a_m'' = \sum_{m} r_{mi} a_i' \;. \tag{2}$$

If the new quantities are treated in the same way, and this process is repeated a sufficient number of times, the ratios among the quantities obtained will eventually become and remain arbitrarily close to those among the coefficients of one of the γ's. This was demonstrated in the fourth section of the previous paper on principal components. The component thus specified in the limit will, apart from a set of cases of probability zero, be that having the greatest sum of squares of correlations with the x's. This sum will equal the limit k_i of the ratio of any one of the trial values to the corresponding one in the previous set.

Now if we substitute (1) in (2), and define

$$c_{mj} = \sum_{i} r_{mi} r_{ij} \;, \tag{3}$$

we shall have

$$a_m'' = \sum_{j} c_{mj} a_j \;. \tag{4}$$

Consequently if we first calculate the quantities c_{mj}, we may use (4) instead of (1), and then *each iteration is precisely equivalent to two iterations with the original correlations*. Thus the number required for any given degree of accuracy is cut in half.

Let R denote the matrix of the covariances r_{ij}. Then from (3), c_{mj} is the element in the mth row and jth column of the symmetrical matrix R^2. Substitution of a set of trial values in (1) is equivalent to multiplying it by the rows of R, while substitution in (4) amounts to multiplication by the rows of R^2.

But we need not stop with this improvement. Having doubled the speed of convergence by squaring R, we can double it again by squaring R^2. If we square a third time we have a matrix R^8, by which a multiplication is equivalent to eight multiplications by the original matrix, and so forth. We can square as many times as we like; if we square s times successively and denote 2^s by t, we obtain R^t, with which one step of the iterative process is equivalent to t steps of the process with the original matrix. The only limit to this acceleration is reached when the convergence is so rapid that an additional squaring of the matrix is not worth while.

The ultimate ratio of consecutive values, such as a_i'/a_1, was k_1 in the original process. In the accelerated process, using R^t, this ratio is $k_1{}^t$. Instead of extracting the tth root to find k_1, it is better to make a final multiplication of the trial values by the rows of R itself, and so upon division to find k_1. This saves labor and also provides a final check upon the calculations, including the squaring of the matrices.

An additional check upon the squaring operations may be accomplished by carrying along an extra column as in the method of least squares. Each entry in this check column is the sum of those preceding it in the same row. The check column is multiplied by each row of the matrix to obtain the check column for the square of the matrix. This check is not so essential as in the method of least squares, in view of the final substitution just mentioned, and since the calculations are so simple that an experienced computer with a good machine is not likely to make a mistake. However, for an ordinary computer, especially if the variates are numerous and the squaring is repeated several times, there is likely to be an eventual saving of labor if this check is made at each step.

In the determination of the second and later principal components by this method, the convergence may be accelerated in the same manner by the use of the tth power of the matrix of the reduced covariances. However there is a further saving of labor here if we form this power, not directly as in the case of R^t by repeated squarings, but with the help of the determination already made by R^t, and the following results obtained with the help of the algebra of matrices. (Bôcher, 1907, 1921).

Putting C_1 for the matrix in which the element in the ith row and jth column is $a_{i1}a_{j1}$ $(i, j, = 1, 2, \cdots, n)$, the matrix of the reduced covariances used in finding the second principal component is

$$R_1 = R - C_1 \ .$$

From relations established in the former memoir (2) (p. 424, equation (16), and p. 425) we have that

$$\sum_h r_{hm} a_{h1} = k_1 a_{m1} \ ,$$

and

$$\sum_h a_{h1}{}^2 = k_1 \ .$$

These lead to the matrix relations

$$RC_1 = C_1R = k_1C_1 \ ,$$

and

$$C_1{}^2 = k_1C_1 \ .$$

From these it is easy to show for any integer n that

$$R^nC_1 = RC_1{}^n = k_1{}^nC_1 \ ,$$

$$C_1{}^n = k_1{}^{n-1}C_1 \ .$$

Hence we readily obtain

$$R_1{}^2 = R^2 - 2RC_1 + C_1{}^2 = R^2 - k_1C_1 \ ,$$

and in general

$$R_1{}^t = R^t - k_1{}^{t-1}C_1 \ . \qquad (t = 2^s)$$

The partial cancellation of the middle by the last term in the squaring is strikingly reminiscent of some of the formulae in the method of least squares, with which the method of principal components presents many analogies.

From the last matrix equation we derive the following simplified method of obtaining numerical values of the desired power of the reduced matrix:

Having determined $k_1{}^t$ as the ratio of consecutive trial values with the matrix R^t, and k_1 as the ratio of consecutive trial values with R, find $k_1{}^{t-1}$ by division. Multiply this by each of the quantities $a_{i1}a_{j1}$ $(i, j = 1, \cdots, n)$ and subtract the products from the corresponding elements of R^t to obtain the elements of $R_1{}^t$.

The elements of R_1 themselves are found as in the former paper, i. e., by subtracting $a_{i1}a_{j1}$ from the corresponding elements of R. The second principal component is found from R_1 and $R_1{}^t$ in exactly the same manner as the first component from R and R^t. To obtain the matrices R_2 and $R_2{}^t$ from which the third principal component is to be found, the elements of R_1 and $R_1{}^t$ are diminished respectively by $k_2a_{i2}a_{j2}$ and by $k_2{}^{t-1}a_{i2}a_{j2}$; and similarly for the later components, if enough is left of the aggregate variance to make these worth computing.

If we subtract k from each element of the principal diagonal of R the resulting determinant may be called $f(k)$. Now multiply $f(k)$ by $f(-k)$, rows by rows. The resulting determinant is identical with that obtained from the matrix R^2 by subtracting k^2 from each element of the principal diagonal. But if, in the equation $f(k)\ f(-k) = 0$, we substitute $k^2 = x$, we obtain an equation of degree n in x whose roots are the squares of those of $f(k) = 0$. This fact shows that not only the greatest root but all the roots of the characteristic equation of R^2 are the squares of the roots of the characteristic equation of R. Our new method is thus brought into colligation with the classical root-squaring method of solving algebraic equations whose fundamental principle is to increase the separation between roots. (6). The iterative process will in general converge rapidly only when the roots are well separated.

In the use of the original iterative method by several workers it was observed that it was often impossible to determine the last digit accurately without carrying the iteration considerably further than at first seemed necessary, and of course using more decimal places than were finally to be retained. This difficulty largely disappears with the use of the method of the present note, since it is so easy to make the equivalent of 8, 16 and 32 iterations in a single operation. However it suggests the theoretical problem of finding limits of error in the determination of the coefficients and the k's, in terms of the differences between consecutive trial values. This problem is very intriguing; but a solution valid *with certainty* under all circumstances appears upon consideration to be impossible. Indeed, as was pointed out in the earlier paper, if the trial values first taken happen to be the coefficients of the tests in a linear function of those whose correlation with γ_1 is exactly zero, we shall never get γ_1, no matter how many times we iterate. If the correlation with γ_1 is almost but not quite zero we shall usually seem to have convergence for a time to another set of values, the coefficients of γ_2, but eventually the discrepancies between consecutive trial values will increase, and in the end the coefficients of γ_1 will be approached. But although an exact limit of error is thus seen to be impossible if we insist on certainty, we shall attain to a very high probability of having the right limit if we carry the iteration far enough to reach stability in three or four decimals; and this is easy when, as in the example below, an additional decimal place is obtained accurately at each trial.

An additional safeguard against spurious convergence to the wrong principal component possibly useful in some cases would be to use two or more different sets of trial values. If all converged to the same result, it would be incredible that this was anything other than the greatest component. But of course the calculation of the later components, if carried out, would in any case reveal such an error.

The symmetry of the matrices makes it unnecessary to write the elements below and to the left of the principal diagonal. The *i*th row is to be read by beginning with the *i*th element of the first row, reading down to the diagonal, and then across to the right.

Each set of trial values is divided by an arbitrary one of them, which may well be taken to be the greatest. This division may well be performed with a slide rule for the first few sets, which do not require great accuracy.

EXAMPLE

The correlations in the matrix R below were obtained by Truman L. Kelley from 140 seventh-grade school children, and have been corrected for attenuation. (4). The variates, in order, are: memory for words; memory for numbers; memory for meaningful symbols; memory for meaningless symbols. At the right of each matrix (which are supposed to have the vacant spaces filled out so as to be symmetrical) is a check column consisting of the sums of the entries made and understood in the several rows.

MATRIX OF CORRELATIONS

$$R = \left\| \begin{array}{cccc} 1. & .9596 & .7686 & .5427 \\ & 1. & .8647 & .7005 \\ & & 1. & .8230 \\ & & & 1. \end{array} \right\| \begin{array}{c} \text{Check column} \\ 3.2709 \\ 3.5248 \\ 3.4563 \\ 3.0662 \end{array}$$

SQUARE OF MATRIX OF CORRELATIONS

$$R^2 = \left\| \begin{array}{cccc} 2.8061 & 2.9640 & 2.8136 & 2.3902 \\ & 3.1592 & 3.0435 & 2.6334 \\ & & 3.0158 & 2.6688 \\ & & & 2.4626 \end{array} \right\| \begin{array}{c} 10.9738 \\ 11.8001 \\ 11.5417 \\ 10.1550 \end{array}$$

$$R^4 = \left\|\begin{array}{cccc||c} 30.289 & 32.538 & 31.780 & 27.908 & 122.515 \\ & 34.964 & 34.161 & 30.012 & 131.678 \\ & & 33.397 & 29.361 & 128.699 \\ & & & 25.835 & 113.115 \end{array}\right\|$$

$$R^8 = \left\|\begin{array}{cccc||c} 3765 & 4046 & 3955 & 3476 & 15242 \\ & 4349 & 4251 & 3736 & 16382 \\ & & 4154 & 3651 & 16011 \\ & & & 3209 & 14072 \end{array}\right\|$$

Multiplying the rows of R^8 by an initial set of trial values all equal to 1 we obtain the sums of the rows of this matrix, namely,

$$15\ 242, \quad 16\ 382, \quad 16\ 011, \quad 14\ 072\ .$$

The last three digits of these numbers are unimportant. We can divide each of these values by the second, since this is the greatest, retaining only a single decimal place, and multiply the values thus obtained by the rows of R^8: in dividing this time we retain two decimal places. With the next iteration we retain two decimal places; with the next, three; with the next, four; and with all later iterations, five. The seventh and eighth sets of trial values thus obtained are exactly identical in all five decimal places; they are

$$.93042, \quad 1.00000, \quad .97739, \quad .85903\ . \tag{1}$$

The products of this set by the rows of R^8 are

$$14\ 402, \quad 15\ 479, \quad 15\ 129, \quad 13\ 297\ . \tag{2}$$

Their products by the rows of R itself, divided by the second of them, are

$$.93045, \quad 1.00000, \quad .97738, \quad .85901\ . \tag{3}$$

These remain exactly stationary under further iteration with R; their products by the rows of R are

$$3.10744, \quad 3.33972, \quad 3.26418, \quad 2.86884\ .$$

From the second of these values, which corresponds to the value unity in the preceding set, we have $k_1 = 3.33972$. From the second of (2) $k_1{}^8 = 15\ 479$. Hence, by division, $k_1{}^7 = 4635.1$.

Multiplying each of the quantities (3) by the square root of the ratio of k_1 to the sum of the squares of these quantities, we obtain the correlations of the first principal component γ_1 with the several tests; these are

$$a_{11} = .9013, \quad a_{21} = .9687, \quad a_{31} = .9468, \quad a_{41} = .8321\ . \tag{4}$$

These are also the coefficients of γ_1 in the expressions for the tests in terms of the four principal components γ_1, γ_2, γ_3, and γ_4 .

The products of the four quantities (4) by themselves and each other are the elements of the matrix

$$C_1 = \left\| \begin{array}{ccccc} .812 & .8732 & .8534 & .7500 & 3.2890 \\ & .9384 & .9172 & .8061 & 3.5349 \\ & & .8964 & .7879 & 3.4549 \\ & & & .6925 & 3.0365 \end{array} \right\|$$

The column at the right consists of the products of (4) by their sum, 3.6490, and since it gives also the sums of the rows of C_1 provides a check. We next calculate, as a basis for the determination of γ_2 ,

$$R_1 = R - C_1 = \left\| \begin{array}{ccccc} .1876 & .0864 & -.0848 & -.2073 & -.0181 \\ & .0616 & -.0525 & -.1056 & -.0101 \\ & & .1036 & .0351 & .0014 \\ & & & .3075 & .0297 \end{array} \right\|$$

Upon multiplying each element of C_1 by the value previously found for $k_1{}^7$, and then subtracting from the corresponding element of R^8 without the necessity of further matrix squaring. In this particular example, no element of R^8, so far as this matrix was calculated, differs by any significant amount from the corresponding element of $k_1{}^7C_1$. Hence we cannot use $R_1{}^8$ to determine γ_2. This condition, however, points to rapid convergence with the matrix R_1. Indeed, starting with the trial values

$$-2, \quad -1, \quad 0, \quad 3,$$

which are approximately proportional to the elements of the check column of R_1, we find after only six iterations that

$$k_2 = .5252, \ a_{12} = -.4187, \ a_{22} = -.2159, \ a_{32} = .1551, \ a_{42} = .5284 \ ,$$

correct to four decimal places. This labor could have been slightly diminished by first calculating the matrix $R_1{}^2 = R^2 - k_1C_1$.

The third principal component, found from R_2 and $R_2{}^2$ with a total of seven iterations, is specified by $k_3 = .1168$, $a_{13} = -.0735$, $a_{23} = -.0637$, $a_{33} = .2818$, $a_{43} = -.1670$.

In summary, it appears that the first principal component, which accounts for 83.5 per cent of the sum of the variances of the tests, and has high positive correlations with all of them, represents general ability to remember; the second, accounting for 13 per cent of the total variance, is correlated with memory both for words and for numbers in a sense opposite to that of its correlations with symbols of both kinds; and the third principal component, with 3 per cent of the total variance ascribable to it, is most highly correlated with memory for meaningful symbols.

The foregoing calculations are carried to the maximum numbers of decimal places possible with the four-place correlations given. Not all these places are significant in the sense of random sampling. If only the small number (one or two) of places significant in the probability sense, relative to the sampling errors of these 140 cases, had been retained at each stage, the number of iterations would have been reduced even further.

Columbia University,
New York.

BIBLIOGRAPHY

1. BOCHER, MAXIME, Introduction to Higher Algebra, New York: Macmillan, 1907, 1921. pp. xi + 321. Chapter VI.
2. HOTELLING, HAROLD, Analysis of a Complex of Statistical Variables into Principal Components. *J. Educ. Psychol.*, 1933, 24, 417-441, 498-520.
3. HOTELLING, HAROLD, The Most Predictable Criterion, *J. Educ. Psychol.*, 1935, 26, 139-142.
4. KELLEY, TRUMAN L., Crossroads in the Mind of Man. Stanford University: Stanford University Press, 1928. pp. 238. p. 100.
5. KELLEY, TRUMAN L., Essential Traits of Mental Life. Cambridge: Harvard University Press, 1935. pp. 145.
6. WHITTAKER, E. T., and ROBINSON, G., The Calculus of Observations. London and Glasgow; Blackie & Son, Limited, 1924, 1929. pp. xvi + 395. p. 106.