A Method for Finding Roots of Arbitrary Matrices

1. Jacobi's Method for Real Symmetric Matrices. There is a well known method due to JACOBI[1] for diagonalizing real symmetric matrices. It consists of performing a sequence of orthogonal transformations (rotations), each one on a two-dimensional subspace of the underlying vector space in which the matrix is considered to be defined. Thus, if we call the original matrix A, we transform with an orthogonal matrix $S_{(km)}$ as follows:

$$B = S^T_{(km)} A S_{(km)}. \tag{1}$$

(The superscript T indicates the transposed matrix.) The main idea of Jacobi's method is to annihilate one of the off-diagonal elements (A_{km}) by this rotation. After we have annihilated A_{km} with this rotation, we go to element $A_{k'm'}$ and find a transformation $S_{(k'm')}$ which will annihilate $A_{k'm'}$. This last transformation, if it affects A_{km}, will in general "deannihilate" it. Nevertheless, one selects each off-diagonal element in turn and performs a transformation which will annihilate it. After one has covered every off-diagonal element once, one has performed what we will call an *iteration*. The essence of Jacobi's method is that even though one undoes the annihilation (in general) of all the off-diagonal elements except the last one, with a sufficient number of iterations the off-diagonal elements will converge to zero, leaving a diagonal matrix. One obtains the final transformation matrix by post-multiplying by the $S_{(km)}$ at each annihilation, starting with the unit matrix. Thus:

$$S_{\text{final}} = S^{(1)}_{k_1 m_1} S^{(1)}_{k_2 m_2} \cdots S^{(1)}_{k_R m_R} S^{(2)}_{k_1 m_1} \cdots S^{(N)}_{k_R m_R} \tag{2}$$

where the superscript refers to the number of the iteration and R is equal to the number of off-diagonal elements ($= \frac{1}{2}n(n-1)$ for an nth order matrix). This S has the property that:

$$S^T A S = D \tag{3}$$

where D is the diagonal matrix whose elements are the roots of the matrix. With each root, there corresponds an eigenvector, which is the column of the S-matrix which occupies the same position in the S-matrix as the column containing the eigenvalue in question occupies in the D matrix.

2. Generalization for Arbitrary Matrices. There are various fundamental theorems which show that a real symmetric (or, for that matter, Hermitian) matrix has real roots and can always be diagonalized by an orthogonal transformation of some kind.

These theorems all break down for non-Hermitian matrices, real or otherwise. In fact, it is known that there exist matrices which no *collineatory* transformation can diagonalize. These are all degenerate, and the best reduction that can be effected is to the Jordan canonical form. However, if we give up the ideal of

diagonalizing matrices, and restrict ourselves to *triangularizing* them, then we have at our disposal a theorem of SCHUR[2] which states that an *arbitrary* matrix can be triangularized by *unitary* transformations. Therefore, instead of specifying $S_{(km)}$ to be orthogonal, we shall require it to be merely unitary, and perform the following transformation on the arbitrary complex matrix, A:

$$B = S^*_{(km)} A S_{(km)}. \tag{4}$$

The elements of $S_{(km)}$ are now complex, and it is in general impossible to characterize S by one parameter. There are really *two* parameters necessary to characterize the transformation. We shall define the 2×2 non-trivial submatrix of $S_{(km)}$ as follows:

$$\begin{bmatrix} s_{kk} & s_{km} \\ s_{mk} & s_{mm} \end{bmatrix} = \begin{bmatrix} a & -\bar{c} \\ c & a \end{bmatrix} \tag{5}$$

where a is real and positive, and the following relation holds:

$$a^2 + |c|^2 = 1. \tag{6}$$

It can be shown that there is no loss in generality in thus restricting $S_{(km)}$.

The elements of the transformed matrix are given by:

$$\begin{aligned} B_{ik} &= aA_{ik} + cA_{im} \\ B_{im} &= -\bar{c}A_{ik} + aA_{im} \end{aligned} \tag{7}$$

$$\begin{aligned} B_{kj} &= aA_{kj} + \bar{c}A_{mj} \\ B_{mj} &= -cA_{kj} + aA_{mj} \end{aligned} \tag{8}$$

$$\begin{aligned} B_{kk} &= a^2A_{kk} + |c|^2A_{mm} + acA_{km} + a\bar{c}A_{mk} \\ B_{km} &= a^2A_{km} - \bar{c}^2A_{mk} + a\bar{c}(A_{mm} - A_{kk}) \\ B_{mk} &= a^2A_{mk} - c^2A_{km} + ac(A_{mm} - A_{kk}) \\ B_{mm} &= a^2A_{mm} + |c|^2A_{kk} - acA_{km} - a\bar{c}A_{mk}. \end{aligned} \tag{9}$$

In view of the non-hermiticity of A it is in general not possible to make both B_{mk} and B_{km} vanish. However, it is always possible to make B_{mk} alone vanish, since there are three equations in the three unknowns a, $Rl(c)$, $Im(c)$ if we bear in mind that setting $B_{mk} = 0$ gives two equations and that (6) must be satisfied.

By setting:

$$a = [1 + |\mu|^2]^{-\frac{1}{2}} \tag{10a}$$

$$c = \mu[1 + |\mu|^2]^{-\frac{1}{2}} \equiv \mu a \tag{10b}$$

we can reduce the number of unknowns and equations to two. We then have, for μ:

$$A_{km}\mu^2 - (A_{mm} - A_{kk})\mu - A_{mk} = 0; \tag{11}$$

hence (setting $A_{mm} - A_{kk} \equiv 2\Delta_{mk}$), we obtain:

$$\mu = \Delta_{mk} \pm \sqrt{\Delta^2_{mk} + A_{km}A_{mk}}. \tag{12}$$

Of these two roots, we choose the one of smaller modulus, in order to avoid large rotations when possible. (This can be done without much trouble in a digital program.)

There are properties of these unitary transformations which are of importance, viz.:

$$(13)\quad |B_{ik}|^2 + |B_{im}|^2 = |A_{ik}|^2 + |A_{im}|^2$$

$$(14)\quad |B_{kj}|^2 + |B_{mj}|^2 = |A_{kj}|^2 + |A_{mj}|^2$$

$$(15)\quad |B_{kk}|^2 + |B_{km}|^2 + |B_{mk}|^2 + |B_{mm}|^2 = |A_{kk}|^2 + |A_{km}|^2 + |A_{mk}|^2 + |A_{mm}|^2.$$

It is no longer possible, as in the real symmetric case, to prove that the sum of squares of absolute values (S.S.A.V.) of the off-diagonal elements decreases. What is necessary to prove is that the S.S.A.V. of the sub-diagonal elements decreases to zero. It is not even true that this sum decreases monotonically.[3]

An example to illustrate this is the following:

$$(16)\quad A \equiv \begin{bmatrix} 1 & 1 & 0 \\ 0 & 1 & 1 \\ 1 & 0 & 1 \end{bmatrix}.$$

One may annihilate the lower left element with the transformation:

$$(17)\quad S_{(31)} \equiv \begin{bmatrix} 0 & 0 & 1 \\ 0 & 1 & 0 \\ -1 & 0 & 0 \end{bmatrix}$$

but the result is:

$$(18)\quad S_{(31)}^T A S_{(31)} = \begin{bmatrix} 1 & 0 & -1 \\ -1 & 1 & 0 \\ 0 & 1 & 1 \end{bmatrix}$$

and we see that the S.S.A.V. of the sub-diagonal elements has increased from 1 to 2. However, one simply keeps on transforming away sub-diagonal elements without regard to the fluctuations of the S.S.A.V. of these elements. Finally they should decrease to zero and the process converge.

The justification for the last statement is based on "experimental" evidence, gained by triangularizing a variety of matrices with a program written in the "Speedcoding" floating-point system for the IBM 701. Floating-point is unnecessary in this problem and slows the machine by a factor of more than ten, but this program was written originally as much to test Speedcoding as to test triangularization. The average time per multiplication or addition in this system is 4 milliseconds.

With a series of matrices whose elements (both real and imaginary parts) were picked at random, the following are the number of iterations and times required for convergence to 10 decimal places (actually, with floating point, to 10^{-30} for sub-diagonal elements).

These data, though incomplete, indicate that the number of iterations in all probability does not increase more strongly than the first power of the order of

TABLE I. *Time for Convergence*

Order	Time per Iteration	Number of Iterations
2	2	3
3	5	6
4	10	9
5	20	12
6	40	14
7	55	14
8	85	17

the matrix. For degenerate matrices, the number of iterations increases considerably, rising to about 25 for a fourth order matrix with four equal roots. It is noteworthy, also, that the roots (which are, of course, the diagonal elements when triangularization is complete) are not accurate. This is because the unitary transformation really preserves the invariants of the matrix, i.e., the coefficients in its characteristic equation, and these do not specify the roots uniquely because of rounding error. For example, let λ_0 be a double root of:

$$(\lambda - \lambda_0)^2 = \lambda^2 - 2\lambda_0\lambda + \lambda_0^2 = 0. \tag{19}$$

What we are given originally, in general, is the expanded form of this equation. Now let us substitute for (λ_0, λ_0) the pair $(\lambda_0 + \epsilon, \lambda_0 - \epsilon)$. The characteristic equation is then:

$$(\lambda - \lambda_0 - \epsilon)(\lambda - \lambda_0 + \epsilon) = \lambda^2 - 2\lambda_0\lambda + \lambda_0^2 - \epsilon^2 = 0. \tag{20}$$

Now if $|\lambda_0| \sim 1$ and $|\epsilon| \sim 10^{-5}$, then the characteristic equation in (20) will differ from that in (19) by ϵ^2, or a quantity $\sim 10^{-10}$ which is of the same order of magnitude as the rounding error in a ten-digit machine. Double precision operations will not increase the accuracy, since the coefficients of the characteristic equation are not known with more than single accuracy. The roots, as we see, may be specified with an ambiguity much greater than the least count of the machine. In general, if there are m equal roots and D digits, the roots will be accurate to about D/m digits. Therefore, it is in general impossible to decide numerically whether roots are equal or just very closely spaced.

These remarks do not apply to Hermitian matrices, where equal roots will come out equal within the actual rounding error of the machine. There is evidently some connection between the "degree of Hermiticity" of a matrix and the accuracy with which its repeated roots can be found.

It is of interest at this point to indicate the complications that arise in attempting to find the eigenvectors of the triangularized matrix, which distinguish markedly the Hermitian and non-Hermitian cases. In particular, it will be shown that, when a root is repeated, it is sometimes impossible to obtain the full number of eigenvectors corresponding to that root. First we look at the Hermitian case.

The original eigenvalue equation was:

$$A\psi^{(k)} = \alpha^{(k)}\psi^{(k)} \tag{21}$$

which now becomes:

$$D\varphi^{(k)} = \alpha^{(k)}\varphi^{(k)} \tag{22}$$

with

$$\psi^{(k)} = S\varphi^{(k)}. \tag{23}$$

Since $D_{ij} = \alpha^{(i)}\delta_{ij}$ we have:

$$\sum_j \alpha^{(i)}\delta_{ij}\varphi_j{}^{(k)} = \alpha^{(i)}\varphi_i{}^{(k)} = \alpha^{(k)}\varphi_i{}^{(k)}. \tag{24}$$

Hence, $\varphi_i{}^{(k)}$ must be equal to δ_{ki}, i.e., all components of $\varphi^{(k)}$ vanish except the kth. Hence also, from (23):

$$\psi_i{}^{(k)} = \sum_j S_{ij}\varphi_j{}^{(k)} = \sum_j S_{ij}\delta_{kj} = S_{ik}. \tag{25}$$

This shows why $\psi^{(k)}$ is the kth column of S.

Now, in the case where A is arbitrary, we can at most triangularize it to obtain:

$$S^*AS = T \tag{26}$$

where T is a triangular matrix. Analogously to the Hermitian case, the transformed eigenvalue equation is:

$$T\varphi^{(k)} = \alpha^{(k)}\varphi^{(k)} \tag{27}$$

which is written out as:

$$\sum_{j=i}^{N} T_{ij}\varphi_j{}^{(k)} = \alpha^{(k)}\varphi_i{}^{(k)}. \tag{28}$$

We subtract one side from the other to obtain:

$$\sum_{j=i}^{N} (T_{ij} - \alpha^{(k)}\delta_{ij})\varphi_j{}^{(k)} = 0. \tag{29}$$

In eq. (24), we had in place of T_{ij}, $\alpha^{(i)}\delta_{ij}$. This gave $(\alpha^{(i)} - \alpha^{(k)})\varphi_i{}^{(k)} = 0$. Here, however, we have terms other than the diagonal ones coming in, so the situation is not so straightforward. Eq. (29) is, in fact, a triangular system of equations for $\varphi_j{}^{(k)}$ which must be solved for each value of k. Let us take $k = N$, so that the last diagonal element vanishes. Let also $i = N$ and consider the last equation, which is:

$$(\alpha^{(N)} - \alpha^{(N)})\varphi_N{}^{(N)} = 0. \tag{30}$$

Since the first factor vanishes, $\varphi_N{}^{(N)}$ is arbitrary. Hence we may solve the triangular system for all the other components of $\varphi^{(N)}$ in terms of $\varphi_N{}^{(N)}$ (assuming no degeneracy, i.e., that there is no other vanishing diagonal term). Next we set $k = N - 1$. The last equation of the system (29) is:

$$(\alpha^{(N)} - \alpha^{(N-1)})\varphi_N^{(N-1)} = 0. \tag{31}$$

Since the first factor does not vanish (by the assumption of no degeneracy), the second factor must. The $(N - 1)$th equation of (29) is:

$$(\alpha^{(N-1)} - \alpha^{(N-1)})\varphi_{N-1}^{(N-1)} + T_{N-1,N}\varphi_N^{(N-1)} = 0. \tag{32}$$

Here, the second term vanishes, and so does the first factor of the first term. Hence $\varphi_{N-1}^{(N-1)}$ is arbitrary. We may then solve for the other components of $\varphi^{(N-1)}$ in terms of $\varphi_{N-1}^{(N-1)}$. Similarly, we may calculate all the vectors $\varphi^{(k)}$ in terms of arbitrary scale factors, which may be chosen so as to normalize the (transformed) eigenvectors $\varphi^{(k)}$. Finally, from $\psi^{(k)} = S\varphi^{(k)}$, we obtain all the eigenvectors of the original problem.

When we do have degeneracy, let us, for definiteness, place one of the repeated roots in the last position of the diagonal. Now we proceed as above, except that for some value M (say) of the index k we will come across a diagonal coefficient in our triangular system of linear equations which vanishes just as the last diagonal element did. Two possibilities are now open. Either the rest of the linear combination (exclusive of the diagonal term) which constitutes the Mth equation vanishes, or it does not. If the former is true, then we are entitled to choose for $\varphi_M^{(N)}$ another arbitrary number and hence obtain another eigenvector. If the latter is true, then we are forced to make it equal to zero by setting $\varphi_N^{(N)} = 0$. This causes us to lose one of the basis vectors of the degenerate subspace defined by the repeated root. We may then start as before by setting $\varphi_M^{(N)}$ equal to some arbitrary number and solve for the rest of the components of φ^N as before. The case when we lose one of the eigenvectors belonging to an eigenvalue corresponds to the case in the Jordan canonical form when a 1 appears attached to two equal roots. When there is no 1 attached to a repeated root, we then have the full complement of eigenvectors for that root. Similar considerations apply to roots of higher multiplicity than 2.

J. GREENSTADT

International Business Machines Corporation
New York, New York

[1] See, for example, R. T. GREGORY, *MTAC*, v. 7, 1953, p. 215.
[2] F. MURNAGHAN, *Theory of Group Representations*. Johns Hopkins Univ. Press. Chap. 1.
[3] Unless a suggestion of J. von NEUMANN (private communication) is followed, according to which one attempts to minimize the S.S.A.V. of the sub-diagonal elements in the mth row and kth column between the pivotal subdiagonal element and the diagonal (including the pivotal element itself). This method, however, is much more time-consuming than that proposed in this paper.

Double Interpolation Formulae and Partial Derivatives in Terms of Finite Differences

1. Abstract. For making interpolations at different parts of a table, double interpolation formulae for mixed forward, backward, and central differences have been derived. However, these formulae are rather cumbersome and time consuming in use. For interpolation with more than one variable, formulae in terms of the tabular entries f_{ij} directly instead of differences are much simpler to use. SALZER[1] has derived such a formula for double-forward interpolation in terms of f_{ij}. This formula works satisfactorily for interpolation near the head of a table. For the purpose of interpolating at other parts of a table, double interpolation formulae for other than double-forward interpolation formula have been derived.

Expressions for partial derivatives of different orders in terms of both double