

DETERMINATION OF CHARACTERISTIC VALUES*

BY

MARK LOTKIN

Avco Mfg. Corp., Wilmington, Mass.

A method for computing the characteristic values of arbitrary matrices was recently proposed by the author in reference [1]. This method utilizes a sequence of unitary transformations which are designed to triangularize the original matrix, thereby producing the desired characteristic values along the diagonal of the triangularized form. Each of the unitary transformations is constructed in such a way as to reduce the norm of the upper-triangular part of the matrix. The basic function underlying this construction is a certain cubic polynomial whose coefficients depend upon the elements of the matrix to be reduced.

In this paper there is presented a refinement of the cubic polynomial, which is shown to possess properties that make it superior to the previous polynomial, on the basis of theoretical and practical considerations. Theoretically, the modified approach is seen to become identical with the Jacobi method for symmetric matrices, in certain cases; practically, the modification has been found to lead to more rapid convergence, at least for a considerable number of certain matrices that were subjected to both techniques.

A brief résumé of the relevant equations inherent in the procedure seems appropriate here. Let us assume, then, that the sequence of transformed matrices has reached the p th stage A_p , $p = 0, 1, 2, \dots$, $A_0 = A$, and that it is consequently desired to construct $A_{p+1} = T_p^{-1} A_p T_p$, where the unitary matrix T_p has the "norm-reducing" property, *i.e.*, if

$$M_p = \sum_{\substack{r,s=1 \\ r < s}}^{n-1} |a_{rs}^{(p)}|^2$$

denotes the "upper-triangular" norm of A_p , and M_{p+1} denotes the corresponding quantity for A_{p+1} , then

$$M_{p+1} - M_p < 0. \tag{1}$$

Let the element $b = |b| \exp i\beta$, located in the (i, j) the position $i < j$, of A , be the "pivot" for the next transformation T_{p+1} , and let the elements in the position (i, i) , (j, j) , and (j, i) of A_p be denoted by $a = |a| \exp i\alpha$, d , and c , respectively; in general, let $a_{rs}^{(p)} = |a_{rs}^{(p)}| \exp (i\alpha_{rs}^{(p)})$.

It was shown in [1] that, for $R \neq 0$,

$$M_{p+1} - M_p < H(R, \theta) \equiv RF(R, \theta) \tag{2}$$

with

$$F(R, \theta) = C_3 R^3 + C_2 R^2 + C_1 R + C_0 \tag{3}$$

*Received June 10, 1958; revised manuscript received November 28, 1958.

and

$$C_3 = |c|^2,$$

$$C_2 = -2|c| [|d| \cos(\theta + \gamma - \delta) - |a| \cos(\theta + \gamma - \alpha)],$$

$$C_1 = |d - a|^2 - 2|b||c| \cos(2\theta - \beta + \gamma) + \sum_{i+1 \leq k \leq i-1} (|a_{ki}^{(p)}|^2 + |a_{ik}^{(p)}|^2),$$

$$C_0 = 2|b| [|d| \cos(\theta + \delta - \beta) - |a| \cos(\theta + \alpha - \beta)] \\ + 2 \sum_{i+1 \leq k \leq i-1} [|a_{ik}^{(p)}| |a_{jk}^{(p)}| \cos(\theta + \alpha_{jk} - \alpha_{ik}) \\ - |a_{ki}^{(p)}| |a_{kj}^{(p)}| \cos(\theta + \alpha_{ki} - \alpha_{kj})].$$

Any solution (R, θ) of the system

$$F(R, \theta) = 0 \tag{4}$$

$$\partial F / \partial \theta = 0 \tag{5}$$

then guarantees that $M_{p+1} < M_p$. Having determined R, θ , the elements of the transformation matrix T_p are found by means of

$$r = (R^2 + 1)^{-1/2}, \text{sgn } r = \text{sgn } R, t = r \exp i\theta, \tag{6}$$

and the new elements of A_{p+1} are determined from the relationships (8) through (15) of reference [1].

In addition to the relationships (16) between the elements of A_{p+1} and A_p , stated in [1], there exist further identities between these elements, of value in the actual performance of numerical calculations; some of these are exhibited below.

A short calculation shows, for example, that

$$|a_1|^2 + |b_1|^2 + |c_1|^2 + |d_1|^2 = |a|^2 + |b|^2 + |c|^2 + |d|^2, \tag{7}$$

by virtue of $r^2(1 + R^2) = 1$. For the same reason,

$$|a_{ik}^{(1)}|^2 + |a_{jk}^{(1)}|^2 = |a_{ik}|^2 + |a_{jk}|^2 \tag{8}$$

$$|a_{ki}^{(1)}|^2 + |a_{kj}^{(1)}|^2 = |a_{ki}|^2 + |a_{kj}|^2 \tag{9}$$

for $1 \leq k \leq n, k \neq i, j$.

Further, whenever $\theta = 0$,

$$b_1 - c_1 = b - c. \tag{10}$$

As stated in the introductory paragraph, the determination of R, θ was previously based on Eqs. (4) and (5).

Now a possibly larger decrease in the norm M than indicated by (4) and (5) may be obtained by choosing R such that

$$M_1 - M \leq \min_{-\infty < R < \infty} H(R, \theta).$$

The values of R at which $\min H$ occurs must then satisfy

$$\partial H / \partial R \equiv 4C_3R^3 + 3C_2R^2 + 2C_1R + C_0 = 0 \tag{11}$$

$$\partial H / \partial \theta = 0. \tag{12}$$

If $R = R_m$ satisfies (11), then clearly

$$M - M_1 \geq R_m^2(3C_3R_m + 2C_2R_m + C_1).$$

Now for fixed θ , $H(0, \theta) = 0$, $(\partial H/\partial R)_{0,\theta} = 0$, $(\partial^2 H/\partial R^2)_{0,\theta} = 2 C_1$. Thus $\min H < 0$ whenever $C_0 \neq 0$. If $C_0 = 0$, but $C_1 < 0$, then still $\min H < 0$. Only if $C_0 = 0$, $C_2^2 - 4C_1C_3 \leq 0$, is $\min H$ equal to zero. Thus in general (11) may be considered superior to (4) for the reduction of the super-diagonal norm. If Eq. (11) has, for fixed θ , two negative minima, then we choose for the transformation $z = R \exp i\theta$ that root R for which $\min \min H$ is assumed.

The superiority of (11) may be deduced also from the study of certain second order matrices, which are obviously of basic importance in this problem.

I. Let us consider

$$A = \begin{bmatrix} a & b \\ c & d \end{bmatrix} \tag{13}$$

with $a = d$. If A is skew-hermitian, then $a = d = 0$. It may be assumed that $bc \neq 0$. Since now $\alpha = \delta$, (11) becomes

$$4 |c| R(|c| R^2 - |b|) = 0,$$

so that $R = |b/c|^{1/2}$, and $\min H(R, \theta) = -|b|^2$. With $z = |b/c|^{1/2} \exp 2^{-1}(\beta - \gamma)i$, and $r^2 = |c|/(|b| + |c|)$, the transformation equations lead to

$$\begin{aligned} a_1 &= a + (bc)^{1/2} \\ d_1 &= a - (bc)^{1/2} \\ b_1 &= 0 \\ c_1 &= (b/|b|)(|c| - |b|). \end{aligned}$$

The expressions for a_1, d_1 are, naturally, the exact roots for the matrix (13), with $a = d$. Thus $M_1 - M = -|b|^2$.

II. Again let us consider matrix A as defined by (13), now subject to the following conditions:

- (i) $\alpha = \delta$
- (ii) $\beta + \gamma = 2\alpha$
- (iii) $\theta = 2^{-1}(\beta - \gamma)$.

Then

$$\begin{aligned} C_3 &= |c|^2 \\ C_2 &= -2|c|(|d| - |a|) \\ C_1 &= (|d| - |a|)^2 - 2|bc| \\ C_0 &= 2|b|(|d| - |a|), \end{aligned}$$

whence

$$S(R) = [|c|R^2 - (|d| - |a|)R - |b|][2|c|R - (|d| - |a|)] \tag{14}$$

$$H(R) = R[|c|R^2 - (|d| - |a|)R - 2|b|][|c|R - (|d| - |a|)]. \tag{15}$$

If $|c|R^2 - (|d| - |a|)R - |b| = 0$, then $H(R) = -|b|^2$. If, however, $2|c|R - (|d| - |a|) = 0$, then $H(R) = |c|R^2 (|c|R^2 + 2|b|) > 0$. Consequently, $\min H(R) = -|b|^2$ is assumed at the roots of $|c|R^2 - (|d| - |a|)R - |b| = 0$.

Therefore, $b_1 = 0$, so that the characteristic values of A appear immediately as a_1, d_1 .

III. Now let the n th order matrix A be hermitian, i.e., $a_{ji} = \bar{a}_{ij}, a_{ii}$ real.

Then it is seen that the cubic polynomial $F(R, \theta)$ of (2) becomes

$$C_3 = |b|^2$$

$$C_2 = -2|b|[|d|\cos(\theta + \gamma - \delta) - |a|\cos(\theta + \gamma - \alpha)]$$

$$C_1 = |d - a|^2 - 2|b|^2 \cos(2\theta - \beta + \gamma)$$

$$C_0 = 2|b|[|d|\cos(\theta + \delta - \beta) - |a|\cos(\theta + \alpha - \beta)].$$

for the choice of $\theta = 2^{-1}(\beta - \gamma)$ above expressions become

$$C_3 = |b|^2$$

$$C_2 = -2|b|(d - a)$$

$$C_1 = (d - a)^2 - 2|b|^2$$

$$C_0 = 2|b|(d - a),$$

and, consequently,

$$H(R, \theta) \equiv RF(R, \theta) = R[|b|R^2 - (d - a)R - 2|b|][|b|R - (d - a)]$$

$$S(R, \theta) \equiv \partial H / \partial R = 2[|b|R^2 - (d - a)R - |b|][2|b|R - (d - a)].$$

Thus again $H(R) = -|b|^2$ for the roots R of the quadratic factor of $S(R, \theta)$. This, naturally, implies again that $b_1 = 0$.

IV. The matrix

$$B = \begin{bmatrix} 1 & 1 & 0 \\ 0 & 1 & 1 \\ 1 & 0 & 1 \end{bmatrix} \tag{16}$$

has been mentioned as one which defies direct treatment by Greenstadt's method [2], as well as by the method of reference [1]. However, it is stated in [3] that by applying a transformation to B which effects a rotation through $\theta = \pi/4$, the transformed matrix becomes tractable by Greenstadt's method, and that in twelve cyclically executed annihilations of the respective pivot the matrix B becomes triangularized to a sufficient degree of accuracy.

It will be seen that the same preparatory rotation through $\theta_0 = \pi/4$, followed by two transforms of the type determined by (11), exactly diagonalizes the matrix (16).

Since the transpose B^T of B has the lower superdiagonal norm, we subject B^T rather than B to the sequence of unitary transformations. The preliminary rotation is effected by

$$T = \begin{bmatrix} \cos \theta_0 & 0 & -\sin \theta_0 \\ 0 & 1 & 0 \\ \sin \theta_0 & 0 & \cos \theta_0 \end{bmatrix}, \tag{17}$$

leading to

$$B_1 \equiv T^{-1}B^T T = \begin{bmatrix} 3/2 & 2^{-1/2} & 1/2 \\ 2^{-1/2} & 1 & -2^{-1/2} \\ -1/2 & 2^{-1/2} & 1/2 \end{bmatrix}.$$

Let us choose next the element $a_{12}^{(1)} = 2^{-1/2}$ as the pivot b . Then case II discussed previously is found to apply. With $z = 2^{-1/2}$ one obtains

$$B_2 = \begin{bmatrix} 2 & 0 & 0 \\ 0 & 1/2 & -3^{1/2}/2 \\ 0 & 3^{1/2}/2 & 1/2 \end{bmatrix}.$$

The superdiagonal norm $M(B_1) = 5/4$ has thus been reduced by the amount $b^2 = 2^{-1}$ to $M(B_2) = 3/4$.

Next we take $a_{23}^{(2)} = -3^{1/2}/2$ as the pivot b . Here case I obtains. Therefore, $R = 1$, $z = i$, and

$$B_3 = \begin{bmatrix} 2 & 0 & 0 \\ 0 & (1/2)(1 + i/3^{1/2}) & 0 \\ 0 & 0 & (1/2)(1 - i/3^{1/2}) \end{bmatrix}, \tag{18}$$

so that B has actually been reduced to diagonal form.

The diagonalization of B^T has thus been achieved by subjecting it to the unitary transformation $B_3 = P^{-1}B^T P$, with

$$\begin{aligned} P &= \begin{bmatrix} 2^{-1/2} & 0 & -2^{-1/2} \\ 0 & 1 & 0 \\ 2^{-1/2} & 0 & 2^{-1/2} \end{bmatrix} \begin{bmatrix} (2/3)^{1/2} & -3^{-1/2} & 0 \\ 3^{-1/2} & (2/3)^{1/2} & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ 0 & 2^{-1/2}i & -2^{-1/2} \\ 0 & 2^{-1/2} & -2^{-1/2}i \end{bmatrix} \\ &= \begin{bmatrix} 3^{-1/2} & -(1/2) - (2 \cdot 3^{1/2})^{-1}i & (2 \cdot 3^{1/2})^{-1} + i/2 \\ 3^{-1/2} & 3^{-1/2}i & -3^{-1/2} \\ 3^{-1/2} & (1/2) - (2 \cdot 3^{1/2})^{-1}i & (2 \cdot 3^{1/2})^{-1} - i/2 \end{bmatrix}. \end{aligned} \tag{19}$$

According to a theorem of Toeplitz (see, e.g., [4]) a matrix M can be reduced by unitary transformations to the diagonal form if and only if the matrix M is normal: $M^{cT} M = M M^{cT}$, where M^{cT} denotes the conjugate transpose of M . Thus the matrix B is seen to be normal. The interesting question then arises whether a class of normal matrices of which B is a member can be reduced to diagonal form by the general technique of this paper. This question can be answered in the affirmative; the results will be published elsewhere.

V. While the choice of a particular value of θ may be appropriate, in special situations, in general the condition (12) may have to be considered, for optimum results. In the example discussed here the values of $\theta = 0, \pi/6, \pi/3, \dots, 5\pi/6$ were applied to each pivot, which was always chosen to be the element of largest modulus. For the transformation $z = R \exp i\theta$ that pair (R, θ) was selected for which $\min H(R, \theta)$ is assumed.

The following matrix is taken from [1]:

$$A = \begin{bmatrix} 1 & 0 & -2 \\ 2 & -1 & 2 \\ 2 & 1 & 0 \end{bmatrix}; \quad (20)$$

its characteristic values are $-2, 1 \pm 2i$. Using the method of [1], eight iterations of A produced the diagonal terms shown under (33) of reference [1], viz.

$$-1.97139 - .07432i, \quad .95014 + 2.11992i, \quad 1.02125 - 2.04650i.$$

The use of the refinement based on Eq. (11) led to

$$-1.99489 + .01100i, \quad .99161 + 2.00158i, \quad 1.00328 - 2.01258i,$$

clearly a considerably improved result over the previous one. The super-diagonal norm at this stage had been reduced from 8.000 to 1.424×10^{-3} .

VI. The matrix

$$C = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 \\ \frac{1}{2} & \frac{1}{3} & \frac{1}{4} & \frac{1}{5} & \frac{1}{6} & \frac{1}{7} \\ \frac{1}{3} & \frac{1}{4} & \frac{1}{5} & \frac{1}{6} & \frac{1}{7} & \frac{1}{8} \\ \frac{1}{4} & \frac{1}{5} & \frac{1}{6} & \frac{1}{7} & \frac{1}{8} & \frac{1}{9} \\ \frac{1}{5} & \frac{1}{6} & \frac{1}{7} & \frac{1}{8} & \frac{1}{9} & \frac{1}{10} \\ \frac{1}{6} & \frac{1}{7} & \frac{1}{8} & \frac{1}{9} & \frac{1}{10} & \frac{1}{11} \end{bmatrix} \quad (21)$$

is one of a sequence of non-symmetric matrices of extremely bad "condition" [5]. Matrix C is nearly singular; the absolute value of its determinant is $(31\ 05\ 2236\ 7232 \cdot 10^{-5})^{-1} = .3220\ 3799 \cdot 10^{-16}$. It is well known—see, e.g., Todd [6]—that certain calculations with these matrices, such as inversions, determination of characteristic values, etc., suffer from "numerical instability". For machines with twelve decimals, employing floating point arithmetic, attempts to invert matrices of even the eighth order have been doomed to failure.

The characteristic values λ_n of C , calculated by means of a determinantal method, and arranged in order of decreasing magnitude, are listed in Table 1.

TABLE 1.
Characteristic values of C.

n	λ_n
1	.2132 3763 $\times 10^1$
2	-.2214 0681 $\times 10^0$
3	-.3184 3305 $\times 10^{-1}$
4	-.8983 2330 $\times 10^{-3}$
5	-.1706 2788 $\times 10^{-4}$
6	-.1397 4990 $\times 10^{-6}$

It is seen from this table that $\lambda_1 \cdot \lambda_2 \cdots \lambda_6 = .3220\ 3799 \times 10^{-16} = \det C$, to eight significant figures.

The reduction technique described above was programmed for the IBM 704 machine. Single precision arithmetic, and floating point with eight significant figures was used. At each step, the super-diagonal element of largest absolute value was taken as the pivot $b = a_{ij}$ of the next transformation. Some of the results are shown in Table 2.

TABLE 2.
Triangularization of Matrix C

p	M_p	$a_{ii}^{(M)}$	$a_{jj}^{(m)}$
0	$.494 \times 10^0$	1.0000 0000	.9090 9090 $\times 10^{-1}$
10	$.236 \times 10^{-2}$.2124 5647 $\times 10^1$.1250 4932 $\times 10^{-2}$
20	$.553 \times 10^{-4}$.2130 8000 $\times 10^1$.2123 4414 $\times 10^{-2}$
30	$.296 \times 10^{-6}$.2132 2888 $\times 10^1$	-.6031 1972 $\times 10^{-4}$
40	$.385 \times 10^{-6}$.2132 3244 $\times 10^1$	-.7919 3310 $\times 10^{-4}$
50	$.616 \times 10^{-7}$.2132 3678 $\times 10^1$	-.2023 3858 $\times 10^{-5}$
60	$.544 \times 10^{-8}$.2132 3763 $\times 10^1$	-.2646 2287 $\times 10^{-5}$
70	$.980 \times 10^{-9}$.2132 3705 $\times 10^1$	-.1103 9584 $\times 10^{-5}$
80	$.149 \times 10^{-9}$.2132 3751 $\times 10^1$	-.1097 4208 $\times 10^{-5}$
90	$.239 \times 10^{-10}$.2132 3772 $\times 10^1$	-.1057 6060 $\times 10^{-5}$
100	$.448 \times 10^{-11}$.2132 3763 $\times 10^1$	-.1451 8099 $\times 10^{-6}$
110	$.647 \times 10^{-12}$.2132 3765 $\times 10^1$	-.1315 6736 $\times 10^{-6}$
120	$.143 \times 10^{-12}$.2132 3766 $\times 10^1$	-.1652 8475 $\times 10^{-6}$
130	$.245 \times 10^{-13}$.2132 3765 $\times 10^1$	-.1634 9713 $\times 10^{-6}$
140	$.581 \times 10^{-14}$.2132 3765 $\times 10^1$	-.1371 2160 $\times 10^{-6}$
150	$.834 \times 10^{-15}$.2132 3765 $\times 10^1$	-.1409 3837 $\times 10^{-6}$

In this table p denotes the number of iterations, M_p the super-diagonal norm of A_p , $a_{ii}^{(M)}$, $a_{jj}^{(m)}$ the diagonal elements in A_p of largest and smallest absolute value, respectively.

The diagonal elements at $p = 150$, arranged in order of decreasing magnitude, are:

$$\begin{aligned}
 &.2132\ 3765 \times 10^1 \\
 &-.2214\ 0677 \times 10^0 \\
 &-.3184\ 3361 \times 10^{-1} \\
 &-.8983\ 2775 \times 10^{-3} \\
 &-.1705\ 5897 \times 10^{-4} \\
 &-.1409\ 3837 \times 10^{-6}
 \end{aligned}$$

Thus the dominant characteristic value is determined at this stage to about seven significant figures, while the smallest one is known to about three. The rate of decrease of M_p , from $.5 \times 10^0$ to $.8 \times 10^{-15}$, would seem to indicate a "linear" type of convergence. It is obvious that further improvements of the results will be achieved once a number of basic routines that are presently in the program have been sharpened. Such routines are concerned with the conversion of numbers from the decimal to binary system, the calculation of trigonometric functions, the location of roots of polynomials, and other operations required in the method.

VII. Among the many other matrices that have been reduced satisfactorily we mention here the Hilbert matrices. These are symmetric matrices whose elements are

$a_{ij} = (i + j - 1)^{-1}$, $i, j = 1, 2, 3, \dots$. Some results for the eighth order matrix are exhibited in Table 3.

TABLE 3.
Characteristic values of a Hilbert matrix

p	M_p	$a_{ii}^{(M)}$	$a_{ii}^{(m)}$
0	$.882 \times 10^0$	1.0000 0000	.6666 6667 $\times 10^{-1}$
10	$.262 \times 10^{-1}$.1693 0662 $\times 10^1$.4273 0205 $\times 10^{-2}$
20	$.388 \times 10^{-3}$.1695 9118 $\times 10^1$.4030 7163 $\times 10^{-2}$
30	$.543 \times 10^{-5}$.1695 9389 $\times 10^1$.3309 0313 $\times 10^{-4}$
40	$.111 \times 10^{-6}$.1695 9390 $\times 10^1$.1518 9860 $\times 10^{-4}$
50	$.123 \times 10^{-8}$.1695 9391 $\times 10^1$.6202 6355 $\times 10^{-5}$
60	$.211 \times 10^{-11}$.1695 9391 $\times 10^1$.3315 4293 $\times 10^{-6}$
70	$.364 \times 10^{-13}$.1695 9391 $\times 10^1$.7117 3257 $\times 10^{-8}$
75	$.424 \times 10^{-15}$.1695 9391 $\times 10^1$.6802 6828 $\times 10^{-8}$

The intermediate diagonal elements in A_{75} are:

$$\begin{aligned} &.2981 \ 2524 \times 10^0 \\ &.2621 \ 2851 \times 10^{-1} \\ &.1467 \ 6944 \times 10^{-2} \\ &.5437 \ 2030 \times 10^{-4} \\ &.1297 \ 1307 \times 10^{-5} \\ &.1589 \ 9581 \times 10^{-7} \end{aligned}$$

The trace of the matrix, which is theoretically equal to the sum of the characteristic values, is 2.0218 0042. The sum of the diagonal elements at $p = 75$, on the other hand, is found to be 2.0218 006. The well-known Givens method for the characteristic values of real symmetric matrices results in a corresponding value of 2.0218 002.

REFERENCES

1. M. Lotkin, *Characteristic values of arbitrary matrices*, Quart. Appl. Math. XIV, 267-275 (1956).
2. J. Greenstadt, *A method for finding roots of arbitrary matrices*, MTAC 9, 47-52 (1955).
3. R. L. Causey, *Computing eigenvalues of non-hermitian matrices by methods of Jacobi type*, J. Soc. Ind. Appl. Math. 6, 172-181 (1958).
4. C. C. MacDuffee, *The theory of matrices*, Chelsea, New York, 1956, p. 76
5. M. Lotkin, *A set of test matrices*, MTAC 9, 153-161 (1955)
6. J. Todd, *The condition of the finite segments of the Hilbert matrix*, NBS Appl. Math. Ser. 39, 1954