SYSTEMS OF EQUATIONS, MATRICES AND DETERMINANTS

(Concluded)

Olga Taussky and John Todd

CHAPTER II

The numerical treatment of problems in this field is not entirely straightforward, and requires particular care even when we are handling problems in quite a moderate number of unknowns e.g., for about 10. This whole subject is at present under active investigation in view of the development of high speed automatic digital computing machines, by means of which it is possible to contemplate the solution of problems in which the number of unknowns is of the order of hundreds. These investigations are in various directions. One of these is a re-examination of old methods and includes their history, classification and unification, together with detailed studies of special methods and attempts at the eradication of the superstitions and the justification of the suspicions which are still rife in the rather primitive field of numerical analysis. A second is the devising and examination of newer methods, such as the gradient or finite iteration methods. A third direction is the study of methods appropriate for special systems, such as those which arise by the discretization of differential equations.

We shall show some of the advantages and disadvantages of the classical methods, indicate some of the newer methods, and show some of the difficulties which can arise. This will be done mainly by discussions of special numerical examples.

For up-to-date accounts of this aspect of the subject, reference should be made to Proceedings of a Symposium on Simultaneous Linear Equations and the Determination of Eigenvalues[1]. The article in this by G. E. Forsythe contains a comprehensive bibliography of the first of the two topics. For a more detailed account of methods for the solution of equations and the inverting of matrices, illustrated with numerical examples and directed more towards those who do not have access to high-speed equipment, we refer to the report by L. Fox[2].

II. 1. *SOLUTION OF SYSTEMS OF SIMULTANEOUS EQUATIONS*

Methods of solution can be classified as direct, or as indirect (or iterative). In the first, in theory, we obtain the exact solution

[1]To be published as National Bureau of Standards, Applied Mathematics Series, Vol. 29, 1952.

[2]To appear in National Bureau of Standards Journal of Research, 1952.

after a finite number of steps. In the second we produce an infinite sequence of numbers which converge to the exact solution. In exceptional circumstances the sequence may be stationary, i.e. all terms are equal after a certain stage. There have recently been developed methods which, in theory, are stationary in general. In practice however, because of the fact that our calculations have to be approximate, they are not stationary.

DIRECT METHODS

1. a. *Elimination Method*

This is one of the oldest methods, and, it will be seen, one of the best, at least for general systems. Consider

(1) $R_1 = 12x - 3y + 2z - 96 = 0$ Check sum = -85

(2) $R_2 = -3x - 8y + z - 68 = 0$ Check sum = -78

(3) $R_3 = x + 2y + 6z - 3 = 0$ Check sum = 6

Eliminate x by subtracting 12 times the third equation from the first and by adding 3 times the third equation to the second. We get

(4) $-27y - 70z - 60 = 0$ Check sum = -157

(5) $-2y + 19z - 77 = 0$ Check sum = -60

Eliminate y by multiplying (5) by 27/2 and subtracting from (4). We find

(6) $-\frac{653}{2}z + \frac{1959}{2} = 0$ Check sum = +653

which gives $z = 3$. We now substitute this value of z in (4) (or (5)) to get $y = -10$ and then put these values in (1) (or (2) or (3)) to get $x = 5$.

In this, as in all numerical work, a description of a proposed solution is not complete unless some sort of checking system is incorporated. This should reveal errors as soon as possible after they occur and not at the last stage, if at all. There is available in this case, the following method which can be applied with minor modifications in most of the manipulations in this field. We shall not mention it explicitly again.

We carry an additional column in our work sheet as the sum of the numbers in that row. After writing down our system we compute the numbers -85, -78, 6 labelled "Check sum" above. We perform on these sums the same operations as we do on the equations. Thus we take $-85 - 12 \times 6 = -157$ and $-78 + 3 \times 6 = -60$ and compare these with the sums of the coefficients in (4) and (5) which are $-27 - 70 - 60 = -157$ and $-2 + 19 - 77 = -60$. Proceeding, as a check sum we compute

$-157-[\frac{27}{2}\times(-60)]=+653$ which agrees with the coefficient sum in (6): $-\frac{653}{2}+\frac{1959}{2}=+653$. A final check might be the substitution of the values of x, y, z in the two equations which were not used in the determination of x.

1. *b. Determinantal Solution*

This is often called Cramer's Rule. Applied to our example it gives

$$\begin{array}{cccc} x & & -y & & z & & -1 \\ \left|\begin{array}{ccc} -3 & 2 & -96 \\ -8 & 1 & -68 \\ 2 & 6 & -3 \end{array}\right| & = & \left|\begin{array}{ccc} 12 & 2 & -96 \\ -3 & 1 & -68 \\ 1 & 6 & -3 \end{array}\right| & = & \left|\begin{array}{ccc} 12 & -3 & -96 \\ -3 & -8 & -68 \\ 1 & 2 & -3 \end{array}\right| & = & \left|\begin{array}{ccc} 12 & -3 & 2 \\ -3 & -8 & 1 \\ 1 & 2 & 6 \end{array}\right| . \end{array}$$

i.e.

$$\frac{x}{3265}=\frac{-y}{6530}=\frac{z}{1959}=\frac{-1}{-653}$$

giving $x=5$, $y=-10$, $z=3$. Thus the problem is reduced to the evaluation of determinants.

The evaluation of a determinant of order n from its explicit definition becomes rapidly more tedious as n increases. For there are $n!$ terms in the expansion and each involves n factors; we have therefore to carry out about $n!\times n$ multiplications as well as about $n!$ additions. A more efficient method of evaluating determinants is therefore essential. We shall show how it is possible to transform a determinant by means of the transformations of the form discussed in I. 2, which do not alter its value, into one whose expansion contains but one non-zero term. Take the determinant

$$D=\left|\begin{array}{ccc} 12 & -3 & 2 \\ -3 & -8 & 1 \\ 1 & 2 & 6 \end{array}\right|$$

Denote its rows by r_1, r_2, r_3. Subtracting twelve times r_3 from r_1 and adding three times r_3 to r_2 we obtain:

$$D=\left|\begin{array}{ccc} 0 & -27 & -70 \\ 0 & -2 & 19 \\ 1 & 2 & 6 \end{array}\right|$$

Multiplying r_2 by 27/2 and subtracting from r_1 we obtain

$$D = \begin{vmatrix} 0 & 0 & -\frac{653}{2} \\ 0 & -2 & 19 \\ 1 & 2 & 6 \end{vmatrix}$$

The expansion of this determinant contains the single non-zero term

$$-(1)\ (-2)\ (-\frac{653}{2}) = -653.$$

What is the expense of this operation in the general case? Again neglecting additions we see that the first stage contains one division and $n \times (n-1)$ multiplications. Neglecting the divisions we see that we have in all about $\Sigma\, n(n-1) \doteqdot \frac{1}{3} n^3$ multiplications. This is a considerable improvement on the $n! \times n$ multiplications, even for small values of n.

This method, which we see is essentially equivalent to the elimination method, is recommended for the evaluation of determinants of general type. It also seems reasonable to discard the determinantal method of solution, even if the determinants are evaluated by the method just described, in favor of the elimination method. This method, as applied to determinants often goes by the name of Chió's method of pivotal condensation. The pivots in our example are the numbers 1, -2. There is, of course, plenty of freedom in the choice of the pivots (i.e., the order in which we eliminate the variables). The most efficient choice is being investigated.

1. *c. Finite Iteration Scheme*

The standard method for finding the center of an ellipse

$$(1) \qquad ax^2 + 2hxy + by^2 + 2gx + 2fy + c = 0,\ ab > h^2$$

is to solve the system of equations

$$(2) \qquad ax + hy + g = 0,\ hx + by + f = 0.$$

Its solution gives the center of (1) for any c. We shall now reverse the process: given a system of equations of the form (2) we shall construct the center of the family of ellipses (1). We use the fact that the chord of contact of a pair of parallel tangents passes through the center. Let θ_0 be any direction. Denote by l_0 the chord of contact of the tangents in the direction θ_0. We denote by θ_1 the direction of l_0 and by l_1 the corresponding chord of contact. The intersection of the lines l_0, l_1 is the center. We reach it as follows: Let S_0 be any point. Proceed from S_0 in the direction θ_0 until we meet l_0, at S_1 say. Proceed from S_1 in the direction θ_1 until we meet l_1 at S_2. Then S_2 is the center.

The method can be extended to higher dimensions; for instance, consider the enveloping cylinders, in three directions, of an ellipsoid. The three planes of contact intersect at the center. This can now be approached in three steps.

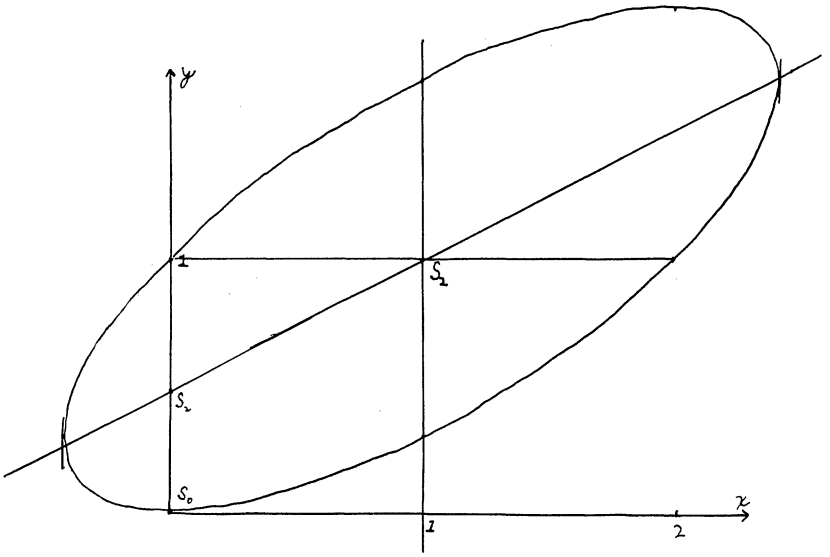

Fig. 1

We shall, for simplicity, in the diagram work out a two-dimensional example. Compare Fig. 1. To solve

$$x - y = 0, \quad -x + 2y - 1 = 0, \tag{3}$$

we determine the center of the ellipse

$$x^2 - 2xy + 2y^2 - 2y = 0.$$

We take $S_0 = (0,0)$ and observe that the *residuals* in (3), i.e. the values of the left hand sides of the two equations for $x = 0$, $y = 0$, are 0, -1. We choose the direction of the residual vector $(0,-1)$, in this case the vertical direction, to be ϑ_0. The chord of contact l_1 is indicated in the diagram. We find the point $S_1 = (0,\frac{1}{2})$. Then we obtain the chord of contact l_1 of tangents parallel to l_0. We then proceed from S_1 along l_0 until we meet l_1 at $S_2 = (1,1)$, the center required.

The general algebraic description of this method for solving $Ax = b$ is the following. We choose any S_0 and observe the residual vector $R_0 = AS_0 - b$ (if $R_0 = 0$ we are finished). We choose $Z_0 = R_0$ and define successively

$$S_{i+1} = S_i + a_i Z_i$$

$$Z_{i+1} = R_i - a_i AZ_i + b_i Z_i, \quad R_{i+1} = R_i - a_i AZ_i$$

where the a_i, b_i are certain scalars: a_i is chosen to make S_{i+1} the proper distance from S_i in the direction Z_i and then b_i is chosen to make Z_{i+1} parallel to the planes of contact of all preceding Z_i. We have to take

$$a_i = R_i^2/Z_i AZ_i, \quad b_i = R_{i+1}^2/R_i^2.$$

It is found that R_{i+1} is indeed the next residual and that R_n is certainly zero in theory. In practice R_n will be small, and it may be necessary to repeat the process.

INDIRECT OR ITERATIVE METHODS

1. d. *Relaxation*

We now return to the 3-dimensional example of 1.*a*. The solution of the equations (1), (2), (3) is accomplished by choosing values of x, y, z to make the residuals R_1, R_2, R_3 zero theoretically or very small practically. We consider the effect of unit changes on the values of the residuals and obtain an operations table

	R_1	R_2	R_3
x	12	-3	1
y	-3	-8	2
z	2	1	6

The matrix here is the transpose of the matrix of the system. The relaxation process, in its most naive form, starts off with arbitrary values of x, y, z and at each stage liquidates, as nearly as possible, by altering one variable, the largest residual. Considerable virtuosity in the art can be achieved by practice, e.g. instead of altering a single variable several can be altered in a "block-relaxation". This may be suggested internally from the behavior of the residuals or externally from some knowledge of the symmetries in the physical problem which gives rise to the numerical problem.

x	y	z		R_1	R_2	R_3
0	0	0		-96	-68	-3
8				0	-92	5
	-11			33	-4	-17
-3				-3	5	-20
		3		3	8	-2
	1			0	0	0
Sum 5	-10	3	Check	0	0	0

Starting with a guess 0, 0, 0 we obtain the residuals -96, -68, -3. Looking at the operations table we see that the most economical way to liquidate $R_1 = -96$ is to change x by 8 thereby altering R_1, R_2, R_3 by 96, -24, 8. We obtain the second row above. Now R_2 is the largest residual and we can liquidate it approximately by a change of -11 in y, which means altering R_1, R_2, R_3 by 33, 88, -22. We thus get the third row above. We now reduce R_1 by a change of -3 in x, then R_2 by a change of 3 in z and then R_2 by a change of 1 in y. At this stage, in view of our specially simple example, we get zero residuals. We check our work by adding up the changes in x, y, z to get $x = 5$, $y = -10$, $z = 3$ and check that the residuals are actually 0, 0, 0.

In practical cases we rarely reach an exact solution. What usually happens is that the residuals are reduced by an order of magnitude. If this is not sufficient we change the scale in our residuals and the variables by a factor 10 or 100 and begin again from this first approximation. This can be repeated until a satisfactory solution is obtained.

Thus described, relaxation is a paper and pencil method and we need only work with small integers. The process can be mechanized, but it would be difficult to codify all the tricks of the trade. We point out that the method is particularly effective when the matrix has a dominant main diagonal.

1. e. *"Seidel" Iteration Scheme*

We again take the three-dimensional case. We begin with any guess at the solution, say $x_0 = 1$, $y_0 = 1$, $z_0 = 1$. We obtain first a revised value for x by substituting the old values of y and z in the first equation and solving for x. This gives $x = 8.08$. We then use this value of x and the old value of z in the second equation to obtain a revised value of y: $y_1 = -11.40$. We then use x and y in the third equation to get a revised value of $z = z_1 = 2.95$. We improve this first approximation 8.08, -11.40, 2.95 in the same way. And so on. A few stages of the process are indicated below:

$$\begin{matrix}1\\1\\1\end{matrix}\quad \begin{matrix}12x - 3 + 2 = 96\\ -24.24 - 8y + 1 = 68\\ 8.08 - 22.80 + 6z = 3:\end{matrix}\quad \begin{matrix}8.08\\-11.40\\2.95\end{matrix};$$

$$\begin{matrix}12x + 34.20 + 5.90 = 96:\\ -13.98 - 8y + 2.95 = 68\\ 4.66 - 19.76 + 6z = 3\end{matrix}\quad \begin{matrix}4.66\\-9.88\\3.02\end{matrix}, \ldots \begin{matrix}5.03\\10.01\\3.00\end{matrix}$$

This process is convenient both manually and mechanically. Convergence is assured if the matrix is positive definite.

1. f. *Gradient Methods*

We return to the example discussed in 1.c. We observe that the

solution to the (consistent) system

$$x - y = 0, \quad -x + 2y = 1$$

is the point which minimizes the quadratic form

$$\epsilon(x,y) = (x - y)^2 + (-x + 2y - 1)^2$$

the minimum being zero.

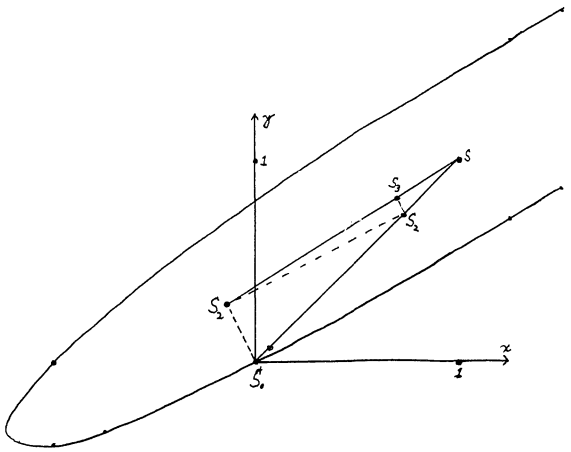

Fig. 2

We take an arbitrary point e.g. $(x_0, y_0) = (0,0)$ as our initial approximation. Compare Fig. 2. We wish to proceed as rapidly as possible to the point where $\epsilon(x,y)$ is minimum. Now $\epsilon(x,y)$ decreases most rapidly at (x_0, y_0) in the direction of the inward normal to that ellipse of the family

$$\epsilon(x,y) = \text{constant} \tag{1}$$

passing through the point. We therefore begin in this direction; it seems reasonable to continue moving in it until we have reached the minimum value of $\epsilon(x,y)$ on this normal: this will occur when the normal becomes a tangent to a member of the family (1). With the choice $(x_0, y_0) = (0,0)$ the direction is $-\left[\left[\frac{\partial \epsilon}{\partial x}\right]_{0,0}, \left[\frac{\partial \epsilon}{\partial y}\right]_{0,0}\right] = (-2,4)$. Our next approximation is

$$x_1 = x_0 - 2r, \quad y_1 = y_0 + 4r$$

where r is such that

$$(x_1 - y_1)^2 + (-x_1 + 2y_1 - 1)^2 - k$$

is a perfect square (in r) for a suitable value of k. We find

$$k = \frac{9}{34}, \quad r = \frac{5}{68}, \quad x_1 = \frac{-5}{34}, \quad y_1 = \frac{10}{34}, \quad \epsilon(x_1, y_1) = \frac{9}{34}.$$

We then proceed from (x_1, y_1) in the new direction of steepest descent: $\left(\frac{6}{17}, \frac{3}{17}\right) \equiv (2,1)$.

We find $(x_2, y_2) = \left(\frac{25}{34}, \frac{25}{34}\right)$, $\epsilon(x_2, y_2) = \left(\frac{9}{34}\right)^2$. In a similar way we find

$$(x_3, y_3) = \left(\frac{805}{1156}, \frac{940}{1156}\right), \quad \epsilon(x_3, y_3) = \left(\frac{9}{34}\right)^3.$$

and

$$(x_4, y_4) = \left(\frac{1075}{1158}, \frac{1075}{1158}\right), \quad \epsilon(x_4, y_4) = \left(\frac{9}{34}\right)^4.$$

The approach of (x_n, y_n) to $(1,1)$ is indicated geometrically in the diagram. The points S_0, S_1, ... lie alternately on the lines $y = x$ and $8x - 13y + 5 = 0$.

The question as to whether it would not be advisable to go beyond or stop short of the point of tangency has been investigated in special cases, but no general decision has been made.

II. 2. *INVERSION OF MATRICES*

The solution of the system of equations $Ax = b$ can be obtained by an application of the matrix A^{-1} to the vector b, an operation involving n^2 multiplications. If it is desired, as often happens in practice, to solve the system $Ax = b$ for various values of b (A being unchanged), then it is frequently advantageous to invert the matrix A, once for all.

Many methods are available for this inversion. They can be classified as direct and indirect or iterative as in the previous section. Many of the methods for the solution of systems of equations can be adapted to the more general problem of inversion.

2. a. *Cholesky's Method*

This is essentially based on the elimination method. In its simplest form it applies to a symmetric matrix. We shall discuss the inversion of a matrix A by using its representation as the product of an upper triangular matrix U and its transpose. Thus

$$A = \begin{pmatrix} 1 & 2 & 3 \\ 2 & 3 & 4 \\ 3 & 4 & 4 \end{pmatrix} = U'U = \begin{pmatrix} a & 0 & 0 \\ b & d & 0 \\ c & e & f \end{pmatrix} \begin{pmatrix} a & b & c \\ 0 & d & e \\ 0 & 0 & f \end{pmatrix} = \begin{pmatrix} a^2 & ab & ac \\ ab & b^2 + d^2 & bc + de \\ ac & bc + de & c^2 + e^2 + f^2 \end{pmatrix}$$

With this assumption we find in turn $a^2 = 1$, $a = 1$; $ab = 2$, $b = 2$;

$ac = 3$, $c = 3$; $b^2 + d^2 = 3$, $d = i$, $bc + de = 4$, $e = 2i$ and $c^2 + e^2 + f^2 = 4$, $f = i$, where we have made arbitrary decisions on the ambiguous signs of c, d and f, which are determined by their squares. It is to be observed that no simultaneous equations have to be solved.

We next observe that

$$A^{-1} = (U'U)^{-1} = (U^{-1})(U')^{-1} = U^{-1}(U^{-1})'$$

and that U^{-1} is easy to find. In fact U^{-1} is also an upper triangular matrix which can be computed from the relation

$$\begin{bmatrix} 1 & 2 & 3 \\ 0 & i & 2i \\ 0 & 0 & i \end{bmatrix} \begin{bmatrix} \alpha & \beta & \gamma \\ 0 & \delta & \epsilon \\ 0 & 0 & \phi \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

from which we obtain successively $\alpha = 1$, $\beta = 2i$, $\gamma = -i$, $\delta = -i$, $\epsilon = 2i$, $\phi = -i$. Hence we have

$$A^{-1} = \begin{bmatrix} 1 & 2i & -i \\ 0 & -i & 2i \\ 0 & 0 & -i \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ 2i & -i & 0 \\ -i & 2i & -i \end{bmatrix} = \begin{bmatrix} -4 & 4 & -1 \\ 4 & -5 & 2 \\ -1 & 2 & -1 \end{bmatrix}$$

The method is sometimes called, for obvious reasons, the square root method. We observe that in it we may introduce surds and complex numbers but that these do not appear in the final result – inversion is a strictly rational process.

The second remark is this, that if we apply the Cholesky process to the solution of equations we can reduce our work somewhat. We replace the equation $Ax = b$ by $U'Ux = b$ and this we can replace by $y = Ux$ and $U'y = b$. We therefore solve first for y and then for x. In the case when A is the matrix under consideration and b the vector (14, 20, 23) we can arrange the results of the work compactly as follows:

1	2	3	14	1	0	0	14	1
2	3	4	20	2	i	0	$8i$	2
3	4	4	23	3	$2i$	i	$3i$	3
	A		b		U'		y	x

We also note, that, in addition to linear checks of the type mentioned in 1.2, we can use a quadratic one. We have

$$x'b = x'Ax = x'U'Ux = (Ux)'(Ux) = y'y$$

In the present example we have

$14 \times 1 + 20 \times 2 + 23 \times 3 = 123$ and $14^2 + (8i)^2 + (3i)^2 = 123$.

2.b. *Iteration*

It is well-known that the sequence of real numbers defined by the recurrence relation

$$x_{n+1} = x_n(2 - Nx_n)$$

converges to N^{-1} for suitable x_0, for any real number $N \neq 0$. It can be shown that, for suitable X_0, the sequence of matrices X_n defined by

$$X_{n+1} = X_n(2I - AX_n)$$

converges to A^{-1}, it being assumed that A is non-singular. As an example take the matrix A to be that of 2.a. If we take

$$X_0 = \begin{pmatrix} -3.0 & 4.1 & -0.9 \\ 4.1 & -5.1 & 1.9 \\ -0.9 & 1.9 & -1.1 \end{pmatrix}$$

we obtain

$$X_1 = \begin{pmatrix} -4.26 & 4.14 & -0.86 \\ 4.14 & -5.06 & 1.94 \\ -0.86 & 1.94 & -1.06 \end{pmatrix} \quad X_2 = \begin{pmatrix} -3.9976 & 3.9864 & -1.0138 \\ 3.9864 & -4.9896 & 2.0104 \\ -1.0136 & 2.0104 & -0.9896 \end{pmatrix}$$

This process is useful for the improvement of approximate inverses.

II. 3. *CHARACTERISTIC ROOTS OF MATRICES*

It has already been remarked that the problem of determining the characteristic roots of a symmetric matrix is essentially the problem of finding the principal axes of a quadric surface. These pure mathematical problems underlie many physical and engineering problems concerned, for instance, with the theory of vibrations.

There are two different problems in this section; according as to whether we wish to determine a few of the dominant characteristic roots (i.e. those with largest moduli), or whether we wish to find all the roots. The first of these problems is the simpler. The following comments refer to the second problem. In the first place, our remarks about the evaluation of determinants indicate that a direct expansion of the determinant, and the assembly of terms into the characteristic polynomial is impractical even if the solution of a polynomial equation could be obtained readily (which is far from being the case). Again, let us assume available a practical method for obtaining a dominant root and a method for obtaining a matrix (of one lower order) having the remaining roots of the original as its characteristic roots. It is tempting to say that repeated application of this process will produce all the characteristic roots. In point of fact, however, the dominant root will only be obtained approximately, and the new matrix will therefore be inaccurate on two counts: due to the inaccuracy of the root and due to the inaccuracies inherent in the numerical calculation involving this approximation to the root. Very soon indeed all significance will be lost. Among the more satisfactory solutions to the

problem is one due to Jacobi: it is described in II. 3. c.

We shall confine our attention to the case of real characteristic roots. This covers the case of symmetric matrices completely.

DOMINANT CHARACTERISTIC VALUES

3.a. *Iteration*

Choose an arbitrary vector $v^{(0)}$ normalized in some way, e.g. to have one of its coordinates unity or to have the sums of the squares of the coordinates unity. Apply the matrix A repeatedly to the vector $v^{(0)}$, expressing each product vector as a scalar multiple of a vector in the chosen normalization. Specifically, if $v^{(i)}$ is normalized then we define $\mu^{(i+1)}$ by the equation

$$Av^{(i)} = \mu^{(i+1)} v^{(i+1)}$$

where $v^{(i+1)}$ is normalized. It can be shown that if A has a single dominant root then these multipliers tend to this value and the (normalized) vectors tend to the corresponding (normalized) characteristic vector.

Let us consider the case

$$A = \begin{pmatrix} 0.2 & 0.9 & 1.32 \\ -11.2 & 22.28 & -10.72 \\ -5.8 & 9.45 & -1.94 \end{pmatrix}$$

and choose $v^{(0)} = (1,0,0)$ and normalize by making the first coordinate unity for simplicity. We obtain the following results:

$Av^{(0)} = (0.2,-11.2,-518) = \mu^{(1)} v^{(1)}$ where $\mu^{(1)} = 0.2$,

$v^{(1)} = (1, -56, -29)$

$Av^{(1)} = (-88.48,-948,-478.74) = \mu^{(2)} v^{(2)}$ where $\mu^{(2)} = -88.48$,

$v^{(2)} = (1, 10.7143, 5.4107)$

$Av^{(2)} = (16.9850,169.5119,84.9534) = \mu^{(3)} v^{(3)}$ where $\mu^{(3)} = 16.9850$,

$v^{(3)} = (1, 9.9801, 5.0017)$

$Av^{(3)} = (15.7843, 157.5384,78.8086) = \mu^{(4)} v^{(4)}$ where $\mu^{(4)} = 15.7843$,

$v^{(4)} = (1, 9.9807, 4.9928)$

$Av^{(4)} = (15.7731,157.6472,78.8316) = \mu^{(5)} v^{(5)}$ where $\mu^{(5)} = 15.7731$,

$v^{(5)} = (1, 9.9947, 4.9979)$

$Av^{(5)} = (15.7925,157.9044, 78.9540) = \mu^{(6)} v^{(6)}$ where $\mu^{(6)} = 15.7925$,

$v^{(6)} = (1, 9.9987, 4.9995)$.

It can be verified that the exact results are $\lambda_1 = 15.8$ and $v_1 = (1, 10, 5)$.

The justification of this process is simple. It is known that, commonly, a matrix A has n different characteristic roots λ_i and n distinct characteristic vectors c_i which are linearly independent and which therefore span the whole space. An arbitrary vector $v^{(0)}$ can be expressed in the form

$$v^{(0)} = \Sigma a_i c_i .$$

Since $Ac_i = \lambda_i c_i$ for $i = 1, 2, \ldots, n$ we have

$$v^{(n)} = A^n v^{(0)} = \Sigma a_i \lambda_i^n c_i$$

and from this, if $\lambda_1 > \lambda_2 \geq \cdots \geq \lambda_n$ we have, for sufficiently large n (depending on the separation of the λ's) that

$$v^{(n)} = A^n v^{(0)} \doteq a\lambda_1^n c_1$$

From this the statements made above follow: that $v_n^{(n)}$ is approximately a multiple of c_1 and that the ratio of corresponding components of $v^{(n)}$ and $v^{(n-1)}$ is approximately λ_1.

3.b. *Relaxation using Rayleigh's Approximation*

For this method we recall the fact that the dominant characteristic root λ_1 of a symmetric matrix A is the upper bound of the Rayleigh quotient

$$R(x) = \frac{xAx'}{xx'}$$

taken over all vectors x and that this bound is assumed when x is the corresponding characteristic vector, c_1. It is also known that the difference between $R(x)$ and λ_i is comparable with the square of the distance between x and c_1, when x is near enough to c_1. We apply the Relaxation method to systems of equations which are not quite compatible since they involve an approximation to the characteristic root sought.

Take the case when

$$A = \begin{bmatrix} 3.5000 & 0.750 & 1.299 \\ 0.750 & 1.625 & 1.083 \\ 1.299 & 1.083 & 2.875 \end{bmatrix}$$

We guess $v^{(0)} = (1,1,1)$ and estimate the dominant characteristic value

$$\mu^{(1)} = R(v) = \frac{3.500 + 1.625 + 2.875 + 2[0.750 + 1.299 + 1.083]}{1^2 + 1^2 + 1^2} =$$

$$\frac{14.264}{3} = 4.755$$

We consider the homogeneous system of equations with matrix $A - \mu^{(1)}I$, i.e.

$$\begin{bmatrix} -1.255 & .750 & 1.299 \\ .750 & -3.130 & 1.083 \\ 1.299 & 1.083 & -1.880 \end{bmatrix}$$

We relax according to the following scheme

$x = 1$	$y = 1$	$z = 1$	Residuals	.794	-1.297	.502
	-0.5			.419	.268	-.040
		-0.1		.289	.160	.148
Sum: $x = 1$	$y = 0.5$	$z = 0.9$	Check:	.2891	.1597	.1485

We now use $v^{(1)} = (1, 0.5, 0.9)$ as a revised approximation to the characteristic vector and re-estimate the characteristic values $\mu^{(2)} = R(v^{(1)}) =$

$$\frac{3.500\times1+1.625\times0.25+2.875\times0.81+2[0.750\times1\times0.5+1.299\times1\times0.9+1.083\times0.5\times0.9]}{1+0.25+0.81}$$

$$= 4.999$$

We now consider the revised system of equations with matrix $A - \mu^{(2)}I$: i.e.

$$\begin{bmatrix} -1.499 & .750 & 1.299 \\ 0.750 & -3.374 & 1.083 \\ 1.299 & 1.083 & -2.124 \end{bmatrix}$$

We relax, beginning with $(x, y, z) = v^{(1)}$, thus:

$x = 1$	$y = 0.5$	$z = 0.9$	Residuals	.045	.038	-.071
		-0.3		.006	.006	-.007
		-0.003		.002	.003	-.001
Sum: $x = 1$	$y = 0.5$	$z = 0.867$	Check:	.0020	.0020	-.0010

Thus we can say that 4.999 is an approximation to the dominant characteristic value of the matrix and that the corresponding characteristic vector is (1, 0.5, 0.867).

ALL CHARACTERISTIC ROOTS

3.c. *Jacobi's Method.*

The method to be described is based on the fact that by an orthogonal transformation from variables x,y to variables x', y' which we can describe in the form

$$\begin{pmatrix} x' \\ y' \end{pmatrix} = \begin{pmatrix} \cos\theta, & \sin\theta \\ -\sin\theta, & \cos\theta \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix}$$

we can express the quadratic form

$$ax^2 + 2hxy + by^2$$

as a sum of squares

$$Ax'^2 + By'^2;$$

we have only to choose θ such that $\tan 2\theta = 2h(a - b)^{-1}$.

We take the matrix

$$A = \begin{pmatrix} 2.879 & -0.841 & -0.148 & 0.506 \\ -0.841 & 3.369 & -0.111 & 0.380 \\ -0.148 & -0.111 & 1.216 & -0.740 \\ 0.506 & 0.380 & -0.740 & 3.536 \end{pmatrix}$$

We observe that the largest off-diagonal element is -0.841. In view of the result just quoted it follows that if

$$\tan 2\theta = -2 \times 0.841/(2.879 - 3.369) \text{ and}$$

$$T_\theta = \begin{pmatrix} \cos\theta & \sin\theta & 0 & 0 \\ -\sin\theta & \cos\theta & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

then the element in $T_\theta A T'_\theta$ corresponding to -0.841 will be zero, or rather, very small in view of the fact that our calculations are not made exactly. We find, in fact, that

$$A_1 = T_\theta A T'_\theta \begin{pmatrix} 2.250 & 0.001 & -0.186 & 0.636 \\ 0.001 & 4.000 & 0.002 & -0.002 \\ -0.186 & 0.002 & 1.216 & -0.740 \\ 0.636 & -0.002 & -0.740 & 3.536 \end{pmatrix}$$

The largest off-diagonal element is -0.740 and we reduce it by transformation by T_ϕ where

$$T_\phi = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & \cos\phi & \sin\phi \\ 0 & 0 & -\sin\phi & \cos\phi \end{pmatrix}$$

and where $\tan 2\phi = 2\times(-0.740)/(1.216 - 3.536)$. We find

$$A_2 = T_\phi A\, T'_\phi = \begin{pmatrix} 2.250 & 0.001 & 0.000 & 0.663 \\ 0.001 & 4.000 & 0.001 & -0.002 \\ 0.000 & 0.001 & 1.004 & 0.001 \\ 0.663 & -0.002 & 0.001 & 3.755 \end{pmatrix}$$

We next reduce the element 0.663 by transformation by T_ψ where

$$T_\psi = \begin{pmatrix} \cos\psi & 0 & 0 & \sin\psi \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ -\sin\psi & 0 & 0 & \cos\psi \end{pmatrix}$$

where $\tan 2\psi = 2\times 0.663/(2.250 - 3.755)$. We find

$$T_\psi A_2 T'_\psi = \begin{pmatrix} 4.005 & -0.002 & 0.000 & -0.001 \\ -0.002 & 4.000 & 0.001 & -0.002 \\ 0.000 & 0.001 & 1.004 & 0.000 \\ -0.001 & -0.002 & 0.000 & 1.999 \end{pmatrix}$$

suggesting that the characteristic roots, which are invariant under the above transformation since the T's are orthogonal, are approximately 4.005, 4.000, 1.004, 1.999.

The proofs that this process is convergent and that it is a practical one are not difficult.

II.4. *ROUNDING-OFF ERRORS*

The question of the effect of rounding errors in calculations is rather difficult but of extreme importance in practice. We shall confine our attention to the case of matrix inversion. What is required in this case are bounds for the difference (in some sense) between the true inverse of a matrix and the result obtained by carrying out a particular process, working always to a fixed number of decimal (or binary) places. These bounds should be readily obtainable and may, of course, be estimates, not necessarily very precise. The bounds may be absolute ones, or probabilistic ones: in the sense that $n\epsilon$ is the absolute error in the sum $\sum_{i=1}^{n} a_i$ due to errors ϵ_i, $|\epsilon_i| \leq \epsilon$, in

a_i while the probabilistic error is estimated as $\epsilon\sqrt{n/12}$.

No satisfactory solutions to these problems exist and the results now available are too complicated to discuss here. To underline the importance of these problems we shall discuss a particular numerical example.

Consider the application of some of the methods just described to the case of the solution of the system of equations

$$\left.\begin{aligned} 10x + 7y + 8z + 7w &= 32 \\ 7x + 5y + 6z + 5w &= 23 \\ 8x + 6y + 10z + 9w &= 33 \\ 7x + 5y + 9z + 10w &= 31 \end{aligned}\right\}$$

which has the solution $x = y = z = w = 1$. This example was constructed by T. S. Wilson.

If we eliminate first y and then w and work only to one place of decimals we obtain

$$0.1x + 0.3z = 0.4$$

$$0.4x + 1.2z = 1.6$$

These equations are not independent so that elimination of x also eliminates z.

If we carry three decimals we find

$$0.060\ z = 0.061$$

which gives a poor determination of $z = 1.0$??

If we apply the relaxation process, we can arrive at the following results, where R_1, R_2, R_3, R_4 denote the successive residuals

x	y	z	w	R_1	R_2	R_3	R_4
16.6	-7.2	-2.5	3.1	0.1	-0.1	-0.1	0.1
2.36	0.18	0.65	1.21	0.01	-0.01	-0.01	0.01
1.136	0.918	0.965	1.021	0.001	-0.001	-0.001	0.001

If we apply the Seidel process we obtain the following sequence of approximations:

$$\begin{bmatrix}0\\0\\0\\0\end{bmatrix};\ \begin{bmatrix}3.20\\0.12\\0.67\\0.20\end{bmatrix};\ \begin{bmatrix}2.44\\0.18\\1.06\\0.35\end{bmatrix};\ \begin{bmatrix}1.98\\0.21\\1.28\\0.46\end{bmatrix};\ \begin{bmatrix}1.70\\0.22\\1.39\\0.55\end{bmatrix};\ \begin{bmatrix}1.55\\0.21\\1.44\\0.61\end{bmatrix};\ \cdots$$

The behavior of this system is blamed on its lack of "condition". Various attempts have been made to measure this lack of condition. The most primitive of these is the smallness of its determinant (in comparison, for example, with the individual terms in its expansion). More satisfactory measures have been introduced. Among these are the M–, N– and P– condition numbers defined by

$M(A) = n\,\mathfrak{M}(A)\,\mathfrak{M}(A^{-1})$ where $\mathfrak{M} = (C)$ denotes $\max_{i,j} |c_{ij}|$

$N(A) = n^{-1}$ norm A norm A^{-1}

$P(A) = \lambda/\mu$ where λ, μ are the greatest and least of the absolute values of the characteristic roots of A.

When these numbers are large, trouble is to be expected. The actual values of the numbers in the case under discussion are respectively

$$2720, \qquad 752, \qquad 2984$$

which are to be compared with the values

$$5.60, \qquad 3.23, \qquad 9.47$$

for the matrix

$$\begin{pmatrix} -2 & 1 & 0 & 0 \\ 1 & -2 & 1 & 0 \\ 0 & 1 & -2 & 1 \\ 0 & 0 & 1 & -2 \end{pmatrix}$$

which is typical of those which arise by the discretization of differential equations.

Estimates for the error in matrix inversion can be given in terms of the condition numbers but it is clear that these numbers cannot be calculated without a knowledge of the inverse (in the first two cases) or of the characteristic roots (in the third).

ERRATA:

Page 13, Sept.-Oct., 1952. The equation

$$A^{-1} = \frac{1}{|A|}\left((-1)^{i+k}A_{ik}\right)$$

should have read

$$A^{-1} = \frac{1}{|A|}\left((-1)^{i+k}A_{ik}\right)'.$$