ROUNDING-OFF ERRORS IN MATRIX PROCESSES

By A. M. TURING

(*National Physical Laboratory, Teddington, Middlesex*)

[Received 4 November 1947]

SUMMARY

A number of methods of solving sets of linear equations and inverting matrices are discussed. The theory of the rounding-off errors involved is investigated for some of the methods. In all cases examined, including the well-known 'Gauss elimination process', it is found that the errors are normally quite moderate: no exponential build-up need occur.

Included amongst the methods considered is a generalization of Choleski's method which appears to have advantages over other known methods both as regards accuracy and convenience. This method may also be regarded as a rearrangement of the elimination process.

THIS paper contains descriptions of a number of methods for solving sets of linear simultaneous equations and for inverting matrices, but its main concern is with the theoretical limits of accuracy that may be obtained in the application of these methods, due to rounding-off errors.

The best known method for the solution of linear equations is Gauss's elimination method. This is the method almost universally taught in schools. It has, unfortunately, recently come into disrepute on the ground that rounding off will give rise to very large errors. It has, for instance, been argued by Hotelling (ref. **5**) that in solving a set of n equations we should keep $n \log_{10} 4$ extra or 'guarding' figures. Actually, although examples can be constructed where as many as $n \log_{10} 2$ extra figures would be required, these are exceptional. In the present paper the magnitude of the error is described in terms of quantities not considered in Hotelling's analysis; from the inequalities proved here it can immediately be seen that in all normal cases the Hotelling estimate is far too pessimistic.

The belief that the elimination method and other 'direct' methods of solution lead to large errors has been responsible for a recent search for other methods which would be free from this weakness. These were mainly methods of successive approximation and considerably more laborious than the direct ones. There now appears to be no real advantage in the indirect methods, except in connexion with matrices having special properties, for example, where the vast majority of the coefficients are very small, but there is at least one large one in each row.

The writer was prompted to carry out this research largely by the practical work of L. Fox in applying the elimination method (ref. **2**). Fox

found that no exponential build-up of errors such as that envisaged by Hotelling actually occurred. In the meantime another theoretical investigation was being carried out by J. v. Neumann, who reached conclusions similar to those of this paper for the case of positive definite matrices, and communicated them to the writer at Princeton in January 1947 before the proofs given here were complete. These results are now published (ref. 6).

1. Measure of work in a process

It is convenient to have a measure of the amount of work involved in a computing process, even though it be a very crude one. We may count up the number of times that various elementary operations are applied in the whole process and then give them various weights. We might, for instance, count the number of additions, subtractions, multiplications, divisions, recordings of numbers, and extractions of figures from tables. In the case of computing with matrices most of the work consists of multiplications and writing down numbers, and we shall therefore only attempt to count the number of multiplications and recordings. For this purpose a reciprocation will count as a multiplication. This is purely formal. A division will then count as two multiplications; this seems a little too much, and there may be other anomalies, but on the whole substantial justice should be done.

2. Solution of equations versus inversion

Let us suppose we are given a set of linear equations $\mathbf{Ax} = \mathbf{b}$ to solve. Here $\mathbf{A}$ represents a square matrix of the nth order and $\mathbf{x}$ and $\mathbf{b}$ vectors of the nth order. We may either treat this problem as it stands and attempt to find $\mathbf{x}$, or we may solve the more general problem of finding the inverse of the matrix $\mathbf{A}$, and then allow it to operate on $\mathbf{b}$ giving the required solution of the equations as $\mathbf{x} = \mathbf{A}^{-1}\mathbf{b}$. If we are quite certain that we only require the solution to the one set of equations, the former approach has the advantage of involving less work (about one-third the number of multiplications by almost all methods). If, however, we wish to solve a number of sets of equations with the same matrix $\mathbf{A}$ it is more convenient to work out the inverse and apply it to each of the vectors $\mathbf{b}$. This involves, in addition, n^2 multiplications and n recordings for each vector, compared with a total of about $\frac{1}{3}n^3$ multiplications in an independent solution. There are other advantages in having an inverse. From the coefficients of the inverse we can see at once how sensitive the solution is to small changes in the coefficients of $\mathbf{A}$ and of $\mathbf{b}$. We have, in fact,

$$\frac{\partial x_i}{\partial b_j} = (\mathbf{A}^{-1})_{ij}, \qquad \frac{\partial x_i}{\partial a_{jk}} = -(\mathbf{A}^{-1})_{ij} x_k.$$

This enables us to estimate the accuracy of the solution if we can judge the accuracy of the data, that is, of the matrix **A** and the vector **b**, and also enables us to correct for any small changes which we may wish to make in these data.

It seems probable that with the advent of electronic computers it will become standard practice to find the inverse. This time has, however, not yet arrived and some consideration is therefore given in this paper to solutions without inversion. A form of compromise involving less work than inversion, but including some of the advantages, is also considered.

3. Triangular resolution of a matrix

A number of the methods for the solution of equations and, more particularly, for the inversion of matrices, depend on the resolution of a matrix into the product of two triangular matrices. Let us describe a matrix which has zeros above the diagonal as 'lower triangular' and one which has zeros below as 'upper triangular'. If in addition the coefficients on the diagonal are unity the expressions 'unit upper triangular' and 'unit lower triangular' may be used. The resolution is essentially unique, in fact we have the following

THEOREM ON TRIANGULAR RESOLUTION. *If the principal minors of the matrix* **A** *are non-singular, then there is a unique unit lower triangular matrix* **L**, *a unique diagonal matrix* **D**, *with non-zero diagonal elements, and a unique unit upper triangular matrix* **U** *such that* $\mathbf{A} = \mathbf{LDU}$. *Similarly there are unique* $\mathbf{L}'$, $\mathbf{D}'$, $\mathbf{U}'$ *such that* $\mathbf{A} = \mathbf{U}'\mathbf{D}'\mathbf{L}'$.

The kth diagonal element of **D** will be denoted by d_k. The $1k$ coefficient of the equation $\mathbf{A} = \mathbf{LDU}$ gives us $l_{11}d_1u_{1k} = a_{1k}$ and since $l_{11} = u_{11} = 1$ this determines d_1 to be a_{11} and u_{1k} to be a_{1k}/d_1; these choices satisfy the equations in question. Suppose now that we have found values of l_{ij}, u_{jk} with $j < i_0$ (that is, we have found the first i_0-1 rows of **L** and columns of **U**) and the first i_0-1 diagonal elements d_k, so that the equations arising from the first i_0-1 rows of the equation $\mathbf{A} = \mathbf{LDU}$ are satisfied; and suppose further that these choices are unique and $d_k \neq 0$. It will be shown how the next row of **L** and the next column of **U**, and the next diagonal element $d_{i_0} \neq 0$ are to be chosen so as to satisfy the equations arising from the next row of $\mathbf{A} = \mathbf{LDU}$, and that the choice is unique. The equations to be satisfied in fact state

$$l_{i_0i_0}d_{i_0}u_{i_0k} = a_{i_0k} - \sum_{j<i_0} l_{i_0j}d_ju_{jk} \quad (k \geqslant i_0),$$

$$l_{i_0k}d_ku_{kk} = a_{i_0k} - \sum_{j<k} l_{i_0j}d_ju_{jk} \quad (k < i_0).$$

The right-hand sides of these equations are entirely in terms of quantities already determined. When $k = i_0$ the first equation is satisfied and can

only be satisfied by putting d_{i_0} = right-hand side, determining d_{i_0}. The equations for $k > i_0$ can then be satisfied by one and only one set of values of $u_{i_0 k}$, provided $d_{i_0} \neq 0$. The equations for $k < i_0$ can also be satisfied by one and only one set of values of $l_{i_0 k}$, since each d_k is different from 0 The new diagonal element d_{i_0} is not 0 because the i_0th principal minor of A is equal to the product of the first i_0 diagonal elements d_k.

4. The elimination method

Suppose that we wish to solve the equations $\mathbf{Ax} = \mathbf{b}$ by the elimination method. The procedure is as follows. We first add such multiples of the first equation to the others that the coefficient of x_1 is reduced to zero in all of them (excepting the first). We then add multiples of the second equation to the later ones until the coefficient of x_2 is reduced to zero. After $n-1$ steps of this nature we shall be left with a set of equations of the form $\sum_{i \leqslant j} v_{ij} x_j = c_i$. From the equation $v_{nn} x_n = c_n$ the unknown x_n can then be found immediately, and by substituting it in the equation $v_{n-1,n-1} x_{n-1} + v_{n-1,n} x_n = c_{n-1}$ we then find x_{n-1}, and so on until by repeated back-substitution we have found all the coefficients of the (originally) unknown vector $\mathbf{x}$. This description of the elimination process is all that is required in order to apply it. We shall find it instructive, however, to look at it further from a number of points of view.

(1) The process of replacing the rows of a matrix by linear combinations of other rows may be regarded as left-multiplication of the matrix by another matrix, this second matrix having coefficients which describe the linear combinations required. Each stage of the above-described elimination process is of this nature, so that we first convert the equations $\mathbf{Ax} = \mathbf{b}$ into $\mathbf{J}_1\mathbf{Ax} = \mathbf{J}_1\mathbf{b}$ and record $\mathbf{J}_1\mathbf{A}$ and $\mathbf{J}_1\mathbf{b}$. We then convert them into $\mathbf{J}_2\mathbf{J}_1\mathbf{Ax} = \mathbf{J}_2\mathbf{J}_1\mathbf{b}$, and so on, until we finally have $\mathbf{J}_{n-1}\ldots\mathbf{J}_1\mathbf{Ax} = \mathbf{J}_{n-1}\ldots\mathbf{J}_1\mathbf{b}$. In accordance with the theorem on triangular resolution we may write $\mathbf{J}_{n-1}\ldots\mathbf{J}_1 = \mathbf{L}^{-1}$ and $\mathbf{J}_{n-1}\ldots\mathbf{J}_1\mathbf{A} = \mathbf{DU}$. The matrix $\mathbf{DU}$ is upper triangular, that is, it has no coefficients other than zeros below the diagonal. The matrix $\mathbf{L}^{-1}$ and its inverse $\mathbf{L}$ are lower triangular.

(2) The matrix $\mathbf{L}$ can be very easily obtained from the matrices $\mathbf{J}_1,\ldots,\mathbf{J}_{n-1}$. We have in fact $\mathbf{L} = 1 + \sum_{r=1}^{n-1}(1-\mathbf{J}_r)$. The proof of this will be left to the reader.

(3) There is no need for us to take either the equations or the unknowns in the order in which they are given. In other words, if $\mathbf{P}$, $\mathbf{Q}$ represent permutations we may solve instead $\mathbf{A}'\mathbf{x}' = \mathbf{b}'$, where $\mathbf{A}' = \mathbf{PAQ}$, $\mathbf{b}' = \mathbf{Pb}$, $\mathbf{x} = \mathbf{Qx}'$. The permutations $\mathbf{P}$, $\mathbf{Q}$ may be chosen bit by bit as we carry the process through. One popular method is to let $\mathbf{Q}$ be the identity, that is,

to take the variables in the order given, and to choose $\mathbf{P}$ so that the coefficients in the matrices $\mathbf{J}_r$ do not exceed unity in absolute magnitude. This is always possible, and for almost all matrices gives a unique $\mathbf{P}$. Alternatively, this variation of the method may be described by saying that $\mathbf{P}$ is chosen so that d_1 shall have the largest possible value, and subject to this, d_2 to be as large as possible, and so on. This procedure is called 'taking the largest coefficient in the column as pivot'. The diagonal elements $d_1, d_2, \ldots, d_n$ are known as the first, second,..., last pivots. There seems to be a definite advantage in using the largest pivot in the column as it is likely to have smaller proportionate errors than other possible pivots, and saves us from the embarrassment of getting a pivot which is little different from zero. It is possible that there is also a further advantage in choosing the largest coefficient in the matrix as pivot.

(4) The leading terms of the work involved in solving a set of n equations by the elimination method are as follows: $\frac{1}{3}n^3+O(n^2)$ multiplications and recordings of which $\frac{1}{2}n^2+O(n)$ recordings involve the vector $\mathbf{b}$.

(5) If, after we have solved one set of equations $\mathbf{Ax} = \mathbf{b}$, we are asked to solve a second set $\mathbf{Ax}' = \mathbf{b}'$ with the same matrix $\mathbf{A}$, we have only to operate on $\mathbf{b}$ with the matrices $\mathbf{J}_1,\ldots,\mathbf{J}_{n-1}$ the values of which may be supposed to have been kept for reference, and then solve $\mathbf{DUx} = \mathbf{J}_{n-1}\ldots\mathbf{J}_1\mathbf{b}$. In other words, if the matrices $\mathbf{J}_1,\ldots,\mathbf{J}_{n-1}$ have been kept (amounting to $\frac{1}{2}n(n-1)$ numbers) the work involved in solving a second set with the same $\mathbf{A}$ is that part of the original work which involved $\mathbf{b}$, namely, $\frac{1}{2}n^2+O(n)$ multiplications and n recordings.

This process may also be expressed in another form, which appears to be quite different, but actually is an identical calculation. As mentioned in (2), the triangle $\mathbf{L}$ in the resolution $\mathbf{A} = \mathbf{LDU}$ may be obtained immediately from the matrices $\mathbf{J}_1,\ldots,\mathbf{J}_{n-1}$. If we put $\mathbf{DUx}' = \mathbf{y}'$ we shall then have $\mathbf{Ly}' = \mathbf{b}'$. The equations $\mathbf{Ly}' = \mathbf{b}'$ may be solved for $\mathbf{y}'$ by one back-substitution process and then the equation $\mathbf{DUx}' = \mathbf{y}'$ solved by a second back-substitution.

(6) As we have described it, the matrices $\mathbf{J}_1\mathbf{A}, \mathbf{J}_2\mathbf{J}_1\mathbf{A},\ldots$, are all written down in full. Actually, however, we are not really interested in all the coefficients of all these matrices. All we need in the end are $\mathbf{J}_1,\ldots,\mathbf{J}_{n-1}$ and $\mathbf{J}_{n-1}\ldots\mathbf{J}_1\mathbf{A}$. It is sufficient, therefore, to calculate all coefficients of $\mathbf{J}_{n-1}\ldots\mathbf{J}_1\mathbf{A}$, and those coefficients of $\mathbf{J}_r\ldots\mathbf{J}_1\mathbf{A}$ which are required for the determination of $\mathbf{J}_{r+1}$. If we write $\mathbf{A}^{(r)}$ for $\mathbf{J}_r\ldots\mathbf{J}_1\mathbf{A}$ we have

$$\mathbf{A}_{ij}^{(r)} = \mathbf{A}_{ij}^{(r-1)}+(\mathbf{J}_r)_{ir}\,\mathbf{A}_{rj}^{(r-1)} \quad (i > r),$$

where

$$(\mathbf{J}_r)_{ir} = -\frac{\mathbf{A}_{ir}^{(r-1)}}{\mathbf{A}_{rr}^{(r-1)}},$$

and by addition

$$\mathbf{A}_{ij}^{(r)} = \mathbf{A}_{ij} + \sum_{s=1}^{r} (\mathbf{J}_s)_{is} \mathbf{A}_{sj}^{(s-1)}.$$

If $i \leqslant r$ we have $\mathbf{A}_{ij}^{(r)} = \mathbf{A}_{ij}^{(r-1)}$ and so

$$\mathbf{A}_{ij}^{(n)} = \mathbf{A}_{ij} + \sum_{s=1}^{n-1} (\mathbf{J}_s)_{is} \mathbf{A}_{sj}^{(n)},$$

$$(\mathbf{J}_r)_{ir} = -\frac{\mathbf{A}_{ir} + \sum_{s=1}^{r-1} (\mathbf{J}_s)_{is} \mathbf{A}_{sr}^{(s-1)}}{\mathbf{A}_{rr} + \sum_{s=1}^{r-1} (\mathbf{J}_s)_{rs} \mathbf{A}_{sr}^{(s-1)}}.$$

Thus we can obtain the numbers actually required ($\mathbf{A}_{ij}^{(n)}$, $(\mathbf{J}_r)_{ir}$) without recording intermediate quantities. This variation of the elimination method will be seen to be identical with the method (1) of § 6 (the 'unsymmetrical Choleski method').

This form of the elimination method is to be preferred to the original form in every way. The recording involved in the work on the matrix is reduced from $\frac{1}{3}n^3 + O(n^2)$ to $n^2 + O(n)$, and the rounding off is at the same time made correspondingly less frequent.

(7) The elimination method may be used to invert a matrix. One method is to solve a succession of sets of equations $\mathbf{A}\mathbf{x}^{(r)} = \mathbf{b}^{(r)}$, where $\mathbf{b}^{(r)} = \delta_{ir}$. The total work involved in the inversion is then $n^3 + O(n^2)$ multiplications. Alternatively, we may invert the matrices **L** and **DU** separately by back-substitution and then multiply them together. The work is still $n^3 + O(n^2)$ multiplications.

(8) When the matrix **A** is symmetric, the matrices **L** and **U** are transposes, and it is therefore unnecessary to calculate both of them. The best arrangement is probably to proceed as with an unsymmetrical matrix, but to ignore all the coefficients below the diagonal in the matrices $\mathbf{A}^{(r)}$. These coefficients are all either zero or equal to the corresponding elements of the transpose. This fact enables us to find the appropriate matrices $\mathbf{J}_r$ at each stage.

(9) The elimination method can be described in another, superficially quite unrelated form. We may combine multiplication of rows and addition to other rows with multiplication of columns and adding to other columns. In other words, we may form a product $\mathbf{J}_{n-1} \ldots \mathbf{J}_1 \mathbf{A} \mathbf{K}_1 \ldots \mathbf{K}_{n-1}$, and try to arrange that it shall be diagonal. The matrix $\mathbf{J}_r$ is to differ from unity only in the rth column below the diagonal, and $\mathbf{K}_r$ is to differ from unity only in the rth row above the diagonal. If we carry out the multiplications by $\mathbf{J}_1, \ldots, \mathbf{J}_{n-1}$ before the multiplications by $\mathbf{K}_1, \ldots, \mathbf{K}_{n-1}$, then it is clear that we have only the elimination method, for in either case we form $\mathbf{J}_1 \mathbf{A}$,

$\mathbf{J}_2\mathbf{J}_1\mathbf{A}$,... and the multiplications by $\mathbf{K}_1$,..., $\mathbf{K}_{n-1}$ which come after actually involve no computation; they merely result in replacing certain coefficients in the matrix $\mathbf{J}_{n-1}...\mathbf{J}_1\mathbf{A}$ by zeros (compare note (2)). It is not quite so clear in the case where the order of calculation is $\mathbf{A}$, $\mathbf{J}_1\mathbf{A}$, $\mathbf{J}_1\mathbf{A}\mathbf{K}_1$, $\mathbf{J}_2\mathbf{J}_1\mathbf{A}\mathbf{K}_1$,.... In this case, however, the right-multiplications do not alter that part of the matrix which will be required later; in fact, they again do nothing but replace certain coefficients by zeros. So far as the subsequent work is concerned, we may consider that these right-multiplications were omitted, and that we formed $\mathbf{J}_{n-1}...\mathbf{J}_1\mathbf{A}$ as in the elimination method.

When this method is used and we choose the largest pivot in the matrix, it is clear that all the coefficients of $\mathbf{J}_r$ and of $\mathbf{K}_r$ do not exceed unity. This provides one proof that when the largest pivot in the matrix is chosen the coefficients of $\mathbf{L}$, $\mathbf{U}$ do not exceed unity (in absolute magnitude).

5. Jordan's method for inversion

In § 4 (1) we mentioned that the elimination process could be regarded as the reduction of a matrix to triangular form by left-multiplication of it by a sequence of matrices $\mathbf{J}_1$,..., $\mathbf{J}_{n-1}$. In the Jordan method we left-multiply the matrix $\mathbf{A}$ by a similar sequence of matrices. The difference is that with the Jordan method we aim at reducing $\mathbf{A}$ to a diagonal,† or preferably to the unit matrix, instead of merely to a triangle.†

The process consists in forming the successive matrices $\mathbf{J}_1\mathbf{A}$, $\mathbf{J}_2\mathbf{J}_1\mathbf{A}$,..., where $\mathbf{J}_r$ differs from the unit matrix only in the rth column, and where $\mathbf{J}_r...\mathbf{J}_1\mathbf{A}$ differs from a diagonal matrix only in the columns after the rth. Let us put

$$\mathbf{A}^{(r)} = \mathbf{J}_r...\mathbf{J}_1\mathbf{A}, \qquad \mathbf{X}^{(r)} = \mathbf{J}_r...\mathbf{J}_1,$$

we shall then have

$$\mathbf{A}^{(r)}_{ij} = \mathbf{A}^{(r-1)}_{ij} + (\mathbf{J}_r)_{ir}\,\mathbf{A}^{(r-1)}_{rj} \quad (i \neq r),$$

$$(\mathbf{J}_r)_{ir} = -\frac{\mathbf{A}^{(r-1)}_{ir}}{\mathbf{A}^{(r-1)}_{rr}} \quad (i \neq r)$$

(so that

$$\mathbf{A}^{(r)}_{ir} = 0 \quad \text{if} \quad i \neq r),$$

$$\mathbf{A}^{(r)}_{rj} = (\mathbf{J}_r)_{rr}\,\mathbf{A}^{(r-1)}_{rj},$$

$$\mathbf{X}^{(0)}_{ij} = \delta_{ij},$$

$$\mathbf{X}^{(r)}_{ij} = \mathbf{X}^{(r-1)}_{ij} + (\mathbf{J}_r)_{ir}\,\mathbf{X}^{(r-1)}_{rj}.$$

The particular diagonal to which $\mathbf{A}$ is reduced is at our disposal. Possible choices include the following. The diagonal may be the unit matrix. Or we may arrange that the diagonal elements of the $\mathbf{J}_r$ are all unity and tolerate the non-unit diagonal elements in $\mathbf{J}_n...\mathbf{J}_1\mathbf{A}$. A third alternative is to arrange that the diagonal elements in $\mathbf{J}_n...\mathbf{J}_1\mathbf{A}$ shall be between 0·1 and 1 and that the diagonal elements in $\mathbf{J}_r$ shall be powers of 10.

† Hereafter 'triangle' and 'diagonal' will be written for 'triangular matrix' and 'diagonal matrix'.

Jordan's method is probably the most straightforward one for inversion. Although it can be used for the solution of equations, it is not very economical for that purpose. For hand work it has the serious disadvantage that the recording is very heavy and cannot be avoided by methods such as that suggested in connexion with the elimination method. It may be the best method for use with electronic computing machinery.

6. Other methods involving the triangular resolution

There are several ways of obtaining the triangular resolution. When it has been obtained, it can be used for the solution of sets of equations, or for the inversion of the matrix as has been described under the elimination method. Possible methods of resolution are described below.

(1) We may use the formulae given in the proof of the theorem on triangular resolution. This involves $\frac{1}{3}n^3+O(n^2)$ multiplications, $n^2+O(n)$ recordings. This method is closely related to Choleski's method for symmetrical matrices ((7) below), and we may therefore describe it as the 'unsymmetrical Choleski method'.

(2) We may apply the elimination method, regarded as a means of obtaining the triangular resolution; see notes (1), (2), (6) on the elimination method.

(3) We may obtain simultaneously, and bit by bit, the four triangles $\mathbf{L}$, $\mathbf{L}^{-1}$, $\mathbf{U}$, $\mathbf{U}^{-1}$ and the diagonal $\mathbf{D}$. The method makes use of the following simple facts about triangles:

(*a*) If we wish to invert a triangle, but only know the values in a subtriangle, we can obtain the coefficients of the inverse in the corresponding subtriangle: for example, if we know the first 5 rows of a lower triangle $\mathbf{L}$, then we can obtain the first 5 rows of $\mathbf{L}^{-1}$.

(*b*) If we know the first r columns of a unit lower triangle then we know its first $r+1$ rows: likewise, if we know the first r rows of a unit upper triangle we know also its first $r+1$ columns.

Let us suppose that we have carried the process to the point of knowing the first r rows of $\mathbf{L}$, the first $r-2$ of $\mathbf{L}^{-1}$ and $r-1$ of $\mathbf{U}$ and $\mathbf{U}^{-1}$. We carry on the inversion of $\mathbf{L}$ to obtain the $(r-1)$th and rth rows of $\mathbf{L}^{-1}$, and then multiply these rows into $\mathbf{A}$ to obtain the rth and $(r-1)$th rows of $\mathbf{L}^{-1}\mathbf{A}$, i.e. of $\mathbf{DU}$. From this we obtain at once the rth and $(r-1)$th rows of $\mathbf{D}$, and dividing obtain the rth and $(r-1)$th rows of $\mathbf{U}$. By (*b*) we have the rth and $(r+1)$th columns of $\mathbf{U}$ and by (*a*) obtain those of $\mathbf{U}^{-1}$. Multiplying we obtain the rth and $(r+1)$th columns of $\mathbf{AU}^{-1}$, i.e. of $\mathbf{LD}$, and from this the rth and $(r+1)$th elements of $\mathbf{D}$ and columns of $\mathbf{L}$. By (*b*) we have the $(r+1)$th and $(r+2)$th rows of $\mathbf{L}$.

We can, of course, arrange to increase r by 1 instead of 2 at each stage.

This is essentially Morris's escalator method (ref. 4), so called because by breaking off the work at any stage we obtain the solution for one of the principal minors of **A**; the order of the minor increases in steps. Morris's method differs in one small point. The diagonal elements **D** are not obtained as the diagonal of $\mathbf{L}^{-1}\mathbf{A}$ or of $\mathbf{AU}^{-1}$, but by using the identity $d_k = a_{kk} - \sum_{i<k} (\mathbf{AU}^{-1})_{ki} d_i^{-1} (\mathbf{L}^{-1}\mathbf{A})_{ik}$, which follows from the (kk) coefficient of the matrix equation $\mathbf{A} = (\mathbf{AU}^{-1})\mathbf{D}^{-1}(\mathbf{L}^{-1}\mathbf{A})$.

If Morris's method is used for the inversion of a matrix the work involved consists of $\frac{5}{3}n^3+O(n^2)$ multiplications (two triangle inversions each $\frac{1}{6}n^3+O(n^2)$, two multiplications of a triangle by **A**, each $\frac{1}{2}n^3+O(n^2)$, and one multiplication of two triangles of opposite type, $\frac{1}{3}n^3+O(n^2)$), and $3n^3+O(n^2)$ recordings (this can be slightly reduced). It does not appear to be especially satisfactory in either respect.

To relate the above account to Morris's put

$$q_k = d_k, \quad x_i = (U^{-1})_{1i}, \quad y_i = (U^{-1})_{2i}, \ldots, \quad x_i' = (L^{-1})_{i1}, \quad y_i' = (L^{-1})_{i2}, \ldots .$$

(4) We may look for an upper triangular matrix **M** such that

$$\mathbf{M^*A^*AM} = 1,$$

that is, so that **AM** is orthogonal. From the first r rows of **M** (which are also the first r columns of **M***) we can obtain the first r rows of **M*** because of its triangular character, and hence the corresponding rows of **M*A*** and **M*A*A**. The equation **M*A*A**.**M** = 1 is then applied, using the first r columns in the $(r+1)$th row of the product. This determines the ratios of the coefficients of **M** in the $(r+1)$th row. The $(r+1)$th diagonal element of the equation then determines the multiplying factor. Having found **M** and **AM** we obtain the inverse as **M**(**AM**)*, or we may solve $\mathbf{Ax} = \mathbf{b}$ by forming (**AM**)***b** and then **M**(**AM**)***b**. In the terminology of orthogonal vectors, as described below, the formation of (**AM**)***b** would be 'expressing **b** in terms of the base of orthogonal vectors'.

This method is the orthogonalization process described in ref. (3), p. 9. It is closely related to the Morris method for symmetrical matrices (see (5) below). We may apply Morris's method by forming **A*A** and then looking for the upper triangular matrix **M** to satisfy **M*A*AM** = 1. This would only involve **A** through the formation of **A*A** and hence of **MA*A**. Thus Morris's method applied to the normalized matrix **A*A** differs from the orthogonalization process only in that **M*A*A** is obtained as **M***(**A*A**) instead of as (**M*A***)**A**.

We now come to methods for symmetrical matrices. These can all be made to provide methods for unsymmetrical matrices by normalizing the given matrix, that is, forming **AA*** from **A**. For instance, if we wish to solve $\mathbf{Ax} = \mathbf{b}$, we may form **A*A** and **A*b**, and then solve $\mathbf{A^*Ax} = \mathbf{A^*b}$ by

one of these methods. This normalizing technique is, however, of doubtful value. The formation of $\mathbf{A}^*\mathbf{A}$ involves $\frac{1}{2}n^3+O(n^2)$ multiplications, so that the work involved is greater with normalization than without, in the case of solving equations, and is no less for the case of inversion. Moreover, normalizing tends to make equations more 'ill-conditioned' (see § 8 below).

(5) A scheme mentioned in note (8) under the elimination method.

(6) We may apply the method (1), but we shall only need to find $\mathbf{L}$ and $\mathbf{D}$, since $\mathbf{U} = \mathbf{L}^*$. As a slight variation we may find $\mathbf{LD}$.

(7) Another variation on (6) is to find $\mathbf{LD}^{\frac{1}{2}}$. This method is due to Choleski (ref. 1). The matrix $\mathbf{LD}^{\frac{1}{2}}$ may involve some pure imaginary numbers, but no strictly complex ones.

(8) Morris's method simplifies considerably for symmetric matrices. From the first r rows of $\mathbf{L}$ we can obtain the first r columns of $\mathbf{L}^{*-1}$, i.e. $\mathbf{U}^{-1}$, by inverting. Left-multiplication by $\mathbf{A}$ gives the first r columns of $\mathbf{AU}^{-1}$, i.e. of $\mathbf{LD}$, and from this we obtain the first $(r+1)$ rows of $\mathbf{L}$. Again Morris obtains $\mathbf{D}$ differently.

This method is identical with a variation of the orthogonalization method, applicable to symmetric matrices and due to L. Fox (ref. 2). Fox regards two vectors $\mathbf{b}$ and $\mathbf{c}$ as 'orthogonal' relative to $\mathbf{A}$ if $(\mathbf{c}, \mathbf{Ab}) = 0$ (scalar product). Fox finds a set of vectors $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n$ which are orthogonal in this sense. The vectors $\mathbf{Av}_r$ may be used as a base for other vectors: we have in fact

$$\mathbf{b} = \sum_r \frac{(\mathbf{b}, \mathbf{v}_r)}{(\mathbf{v}_r, \mathbf{Av}_r)} \mathbf{Av}_r.$$

The solution of equations is effected by means of the formula

$$\mathbf{A}^{-1}\mathbf{b} = \sum_r \frac{(\mathbf{b}, \mathbf{v}_r)}{(\mathbf{v}_r, \mathbf{Av}_r)} \mathbf{v}_r.$$

It is best to obtain $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n$ by orthogonalizing the unit coordinate-axis vectors, that is, besides the vectors being orthogonal, $\mathbf{v}_r$ is restricted to be a linear combination of $\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_n$, or in other words, to have all coefficients after the rth equal to 0. In this case the vectors $\mathbf{v}_r$ are the rows of $\mathbf{L}^{-1}$, and the orthogonality relation is $\mathbf{L}^{-1}(\mathbf{AL}^{-1*}) = \mathbf{D}$. The orthogonalization process by which $\mathbf{L}^{-1}$ is found is identical with the inversion of $\mathbf{AL}^{-1*}\mathbf{D}^{-1}$.

7. Measure of the magnitude of a matrix

There are a number of ways in which the magnitude of a matrix may be measured by a real number. They include:

The norm. The norm $N(\mathbf{A})$ of the matrix $\mathbf{A}$ is given by

$$N(\mathbf{A}) = (\text{trace } \mathbf{A}^*\mathbf{A})^{\frac{1}{2}} = \Big(\sum_{i,j} a_{ij}^2\Big)^{\frac{1}{2}}.$$

The maximum expansion $B(\mathbf{A})$. This is given by

$$B(\mathbf{A}) = \max_{\mathbf{x}} \frac{|\mathbf{Ax}|}{|\mathbf{x}|} = \max_{\mathbf{x}} \frac{(\mathbf{Ax}, \mathbf{Ax})^{\frac{1}{2}}}{(\mathbf{x}, \mathbf{x})^{\frac{1}{2}}}.$$

The maximum coefficient $M(\mathbf{A})$. This is the largest coefficient in the matrix:

$$M(\mathbf{A}) = \max_{i,j} |a_{ij}|.$$

Of these measures one of the first two above is probably of greatest theoretical significance. In this paper we deal chiefly with the maximum coefficient, since it is the most easily computed.

A number of inequalities relating these are listed below.

$$M(\mathbf{X}+\mathbf{Y}) \leqslant M(\mathbf{X})+M(\mathbf{Y}) \tag{7.1}$$
$$M(\mathbf{XY}) \leqslant nM(\mathbf{X})M(\mathbf{Y}) \tag{7.2}$$
$$B(\mathbf{X}+\mathbf{Y}) \leqslant B(\mathbf{X})+B(\mathbf{Y}) \tag{7.3}$$
$$B(\mathbf{XY}) \leqslant B(\mathbf{X})B(\mathbf{Y}) \tag{7.4}$$
$$N(\mathbf{X}+\mathbf{Y}) \leqslant N(\mathbf{X})+N(\mathbf{Y}) \tag{7.5}$$
$$N(\mathbf{XY}) \leqslant N(\mathbf{X})N(\mathbf{Y}) \tag{7.6}$$
$$N(\mathbf{X}) \leqslant nM(\mathbf{X}) \tag{7.7}$$
$$M(\mathbf{X}) \leqslant N(\mathbf{X}) \tag{7.8}$$
$$M(\mathbf{X}) \leqslant B(\mathbf{X}) \tag{7.9}$$
$$B(\mathbf{X}) \leqslant n^{\frac{1}{2}}M(\mathbf{X}) \tag{7.10}$$
$$B(\mathbf{X}) \leqslant N(\mathbf{X}) \tag{7.11}$$
$$N(\mathbf{X}) \leqslant n^{\frac{1}{2}}B(\mathbf{X}) \tag{7.12}$$

8. Ill-conditioned matrices and equations

When we come to make estimates of errors in matrix processes we shall find that the chief factor limiting the accuracy that can be obtained is 'ill-conditioning' of the matrices involved. The expression 'ill-conditioned' is sometimes used merely as a term of abuse applicable to matrices or equations, but it seems most often to carry a meaning somewhat similar to that defined below.

Consider the equations

$$\left.\begin{aligned} 1{\cdot}4x+0{\cdot}9y &= 2{\cdot}7 \\ -0{\cdot}8x+1{\cdot}7y &= -1{\cdot}2 \end{aligned}\right\} \tag{8.1}$$

and form from them another set by adding one-hundredth of the first to the second, to give a new equation replacing the first

$$\left.\begin{aligned} -0{\cdot}786x+1{\cdot}709y &= -1{\cdot}173 \\ -0{\cdot}800x+1{\cdot}700y &= -1{\cdot}200 \end{aligned}\right\}. \tag{8.2}$$

The set of equations (8.2) is fully equivalent to (8.1), but clearly if we attempt to solve (8.2) by numerical methods involving rounding-off errors

we are almost certain to get much less accuracy than if we worked with equations (8.1). We should describe the equations (8.2) as an *ill-conditioned* set, or, at any rate, as ill-conditioned compared with (8.1). It is characteristic of ill-conditioned sets of equations that small percentage errors in the coefficients given may lead to large percentage errors in the solution. If we are required to solve the equations $\mathbf{Ax} = \mathbf{b}$, but the coefficients used are those of $\mathbf{A}-\mathbf{S}$ instead of those of $\mathbf{A}$, $\mathbf{S}$ being a small matrix, then, to first order in $\mathbf{S}$, the solution obtained will be $\mathbf{x}_0+\mathbf{A}^{-1}\mathbf{S}\mathbf{x}_0$, where $\mathbf{x}_0$ is the correct solution. We may average the effect of this over a random population of matrices $\mathbf{S}$, and over the coefficients in the solution and matrix, and we shall find the

$$\frac{\text{R.M.S. error of coefficients of solution}}{\text{R.M.S. coefficient of solution}} = \frac{1}{n}N(\mathbf{A})N(\mathbf{A}^{-1})\frac{\text{R.M.S. error of coefficients of A}}{\text{R.M.S. coefficient of A}}.$$

This equation suggests that we might take either $N(\mathbf{A})N(\mathbf{A}^{-1})$ or $\frac{1}{n}N(\mathbf{A})N(\mathbf{A}^{-1})$ as a measure of the degree of ill-conditioning in a matrix. We will adopt the latter and call $\frac{1}{n}N(\mathbf{A})N(\mathbf{A}^{-1})$ *the N-condition number of* $\mathbf{A}$. We will also use $nM(\mathbf{A})M(\mathbf{A}^{-1})$ as another measure of ill-conditioning and call it *the M-condition number of* $\mathbf{A}$. There is substantial agreement between the two measures, though the M-number tends to be the larger, especially with diagonal or nearly diagonal matrices.

It should be noted that if all the coefficients of a matrix are multiplied by the same factor the condition numbers are unaltered, but that if a row or column is multiplied by a very large or a very small number the condition numbers are usually increased. For instance, the matrices

$$\begin{pmatrix} 0{\cdot}8 & 0{\cdot}6 \\ -0{\cdot}6 & 0{\cdot}8 \end{pmatrix} \quad (8.3) \qquad \text{and} \qquad \begin{pmatrix} 0{\cdot}008 & 0{\cdot}006 \\ -0{\cdot}6 & 0{\cdot}8 \end{pmatrix} \quad (8.4)$$

have the M-condition numbers 1·28 and 128 respectively and N-condition numbers 1 and 50·005. This may be considered quite a satisfactory example of the application of the definition. In practice one will tend to work with the same number of figures throughout a matrix, and the small values in the first row of (8.4) will prejudice the accuracy obtainable, because of the number of significant figures available. It is certainly true that a trivial modification improves the conditioning, but we should consider that until the possibility of this modification has been observed and action taken, the matrix remains ill-conditioned.

It is often stated that ill-conditioned matrices are ones which have small determinants, that is, small considering the magnitudes of the coefficients. This statement contains a certain amount of truth. It is certainly the case that bad conditioning and small determinants tend to go together. However, the determinant may differ very greatly from the above-defined condition numbers as a measure of conditioning. This may be illustrated by the cases of the matrices

$$\begin{pmatrix} 1 & 0 & 0 \\ 0 & 0{\cdot}1 & 0 \\ 0 & 0 & 0{\cdot}1 \end{pmatrix}; \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 1 \\ 0 & 0 & 0{\cdot}01 \end{pmatrix}; \begin{pmatrix} 1 & 1 & 1 \\ 1 & 1{\cdot}1 & 1 \\ 1 & 1 & 1{\cdot}1 \end{pmatrix}; \begin{pmatrix} 1 & 1 & 1 \\ 1 & 2 & 1 \\ 1 & 1 & 1{\cdot}01 \end{pmatrix}$$

all of which have the determinant 0·01, and which have the M-condition numbers 30, 300, 69·3, 612, respectively, and N-condition numbers 4·77, 47·1, 33·0, 232.

The best conditioned matrices are the orthogonal ones, which have N-condition numbers of 1. Their M-condition numbers are mostly of the order of magnitude of $\log n$ (for large order n). If the coefficients of a matrix are chosen at random from a normal population we shall get N-condition numbers of the order of $n^{\frac{1}{2}}$ and M-condition numbers about $\log n$ times greater. Thus random matrices are only slightly ill-conditioned.

The matrices which occur in practical problems are by no means random in this sense. There is a very large class of problems which naturally give rise to highly ill-conditioned equations. Suppose, for example, that we have reason to believe that some function of position in two dimensions can be represented by a polynomial of the fourth degree and that we wish to determine the coefficients. To this end we measure the values of the function at 25 points, and so obtain 25 linear equations for the desired coefficients. It may well happen that we are only able to make the measurements within a small region, and this will certainly mean that the equations are ill-conditioned. In such a case the equations might be improved by a differencing procedure, but this will not necessarily be the case with all problems. Preconditioning of equations in this way will always require considerable liaison between the experimenter and the computer, and this will limit its applicability.

9. The classical iterative method

Suppose that $\mathbf{B}$ is an approximate inverse of $\mathbf{A}$. Then we can obtain from it a better inverse $\mathbf{B}_2$ by the formula $\mathbf{B}_2 = 2\mathbf{B}-\mathbf{BAB}$. If we write $\mathbf{E} = 1-\mathbf{AB}$, $\mathbf{E}_2 = 1-\mathbf{AB}_2$, so that $\mathbf{E}$ and $\mathbf{E}_2$ give a measure of the incorrectness of the two inverses: we have $\mathbf{E}_2 = \mathbf{E}^2$, so that at each application of this process the error is essentially squared.

The work involved in applying this method is considerable, since it involves $2n^3$ multiplications at each stage. It may be useful in cases where a good approximate inverse is already available, and $1-\mathbf{AB}$ has already been calculated, but found to be a little larger than can be tolerated. We may then calculate $\mathbf{B}_2$ but carry the process no farther. This involves n^3 multiplications, but since we may write $\mathbf{B}_2 = \mathbf{B}+\mathbf{BE}$, the number of figures in one of the factors (viz. in $\mathbf{E}$) may be kept small.

A somewhat similar type of method applies for the improvement of solutions of sets of equations. Suppose, for example, we have to solve the equations $\mathbf{Ax} = \mathbf{b}$ and that we have obtained a resolution $\mathbf{A} = \mathbf{L}.\mathbf{DU}$ (say), somewhat inaccurately. By double back-substitution we obtain a solution $\mathbf{x}_1$ of $\mathbf{L}.\mathbf{DUx} = \mathbf{b}$, which is an inaccurate solution of $\mathbf{Ax} = \mathbf{b}$. We may further test this solution by forming the 'residual' vector $\mathbf{b}_1 = \mathbf{b}-\mathbf{Ax}_1$, and if this is too large we solve $\mathbf{Ax} = \mathbf{b}_1$ to obtain a correction. In this process we do not obtain 'quadratic convergence' but only convergence in geometric progression. On the other hand, the method is very practical because the work involved per stage is only $2n^2$ multiplications.

10. General remarks on error estimates. The error in a reputed inverse

Error estimates can be of two kinds. We may wish to know how accurate a certain result is, and be willing to do some additional computation to find out. A different kind of estimate is required if we are planning calculations and wish to know whether a given method will lead to accurate results. In the former case we do not care what quantities the error is expressed in terms of, provided they are reasonably easily computed. With these estimates we wish to be absolutely sure that the error is within the range stated, but at the same time not to state a range which is very much larger than necessary. With the second type of estimate, the error is preferably expressed in terms of quantities whose meaning is sufficiently familiar that the general run of values involved may at least be guessed at. We are also as much interested in the statistical behaviour of the errors as in the maximum possible value.

This paper is mainly concerned with estimates of the second kind, since those of the first kind can be quickly dismissed. Let $\mathbf{B}$ be a reputed inverse of $\mathbf{A}$. To determine its accuracy we form $\mathbf{E} = 1-\mathbf{AB}$. Then in view of the inequalities (7.1), (7.2), and the equation

$$\mathbf{A}^{-1}-\mathbf{B} = \mathbf{B}(\mathbf{E}+\mathbf{E}^2+\ldots)$$

we have

$$M(\mathbf{B}-\mathbf{A}^{-1}) \leqslant \sum_{r=1}^{\infty} M(\mathbf{B}\mathbf{E}^r) \leqslant \sum_{r=1}^{\infty} n^r M(\mathbf{B})\{M(\mathbf{E})\}^r = \frac{nM(\mathbf{B})M(\mathbf{E})}{1-nM(\mathbf{E})},$$

which is the required error estimate. In order to apply this inequality it is necessary to carry out the matrix multiplication $\mathbf{BA}$, involving n^3 multiplications. However, if it is intended to apply the classical iteration method for improving the inverse at least once, we shall have to calculate $\mathbf{E}$ in doing so, and we shall have $1-\mathbf{AB}_2 = \mathbf{E}_2 = \mathbf{E}^2$ and therefore

$$M(\mathbf{B}_2-\mathbf{A}^{-1}) \leqslant \frac{nM(\mathbf{B}_2)M(\mathbf{E}_2)}{1-nM(\mathbf{E}_2)} \leqslant \frac{n^2M(\mathbf{B}_2)\{M(\mathbf{E})\}^2}{1-n^2\{M(\mathbf{E})\}^2}.$$

It should be observed that this inequality is only applicable to the inversion of a matrix, and not to the solution of equations. It is difficult to determine the accuracy of the solution of a set of equations without inverting the matrix. This is another reason why it is preferable to treat inversion rather than solution of equations as a standard process.

When making estimates of the effects of rounding-off errors we need the process under examination to be rather minutely described. If, for instance, a product abc is to be formed, we need to know whether it is obtained as $ab.c$ or as $a.bc$. If it is obtained as $ab.c$ we shall need to know how many figures are kept in ab. This may be either a definite number of decimal or binary places, or a definite number of significant figures, or the number of figures kept may be made to depend on the results of previous calculations. Usually, however, by a trivial modification of the quantity recorded, these latter cases can be reduced to one of the former.

The variety of possible detailed calculation procedures is, of course, vastly greater than the list of methods which we have considered, for these can be subdivided into numerous alternatives which appear only trivially different at first sight, but which may differ very seriously from the point of view of error estimates. We cannot here carry out the analysis for more than a very few of the procedures. These have been chosen so as to give bounds of error which are both reasonably small and also fairly simple in their analytical form. We have concentrated particularly on error estimates which can be expressed in terms of the matrix $\mathbf{A}$ and its inverse. In practical work the details of the procedure must be determined by other considerations. With any particular procedure it will usually be found possible to obtain some estimate of the type proved in this paper, but usually quantities such as $M(\mathbf{L})$, $M(\mathbf{D}^{-1})$, etc., will be involved. These can be obtained conveniently as a by-product in the calculation. Alternatively, one may find bounds of error by calculating $1-\mathbf{AB}$ as above. In this case the importance of the analysis which follows is to show that

it is probable that the error obtained will be reasonably small if a process is used which is somewhat similar to one of those here considered, and that these methods are therefore reasonable ones to use. Our main purpose in this paper is to establish that the exponential build-up of errors need not occur, and this will be proved when we have found one method of inversion where it is absent.

11. Rounding-off errors in Jordan's method

The Jordan method was described in § 5, but we have now to specify the details of the rounding-off and the diagonal. We shall consider the case where A is reduced to a unit matrix. We assume that in the calculation of each quantity

$$A_{ij}^{(r-1)} - \frac{A_{rj}^{(r-1)} A_{ir}^{(r-1)}}{A_{rr}^{(r-1)}},$$

an error of at most ϵ is made. How this is to be secured need not be specified, but it is clear that the number of figures to be retained in $A_{ir}^{(r-1)}/A_{rr}^{(r-1)}$ will have to depend on the values of the $A_{rj}^{(r-1)}$. Likewise, we assume that in the calculation of

$$X_{ij}^{(r-1)} - \frac{X_{rj}^{(r-1)} A_{ir}^{(r-1)}}{A_{rr}^{(r-1)}}$$

an error of at most ϵ' is made. It is convenient to think of these errors as quantities deliberately added after the accurate calculation has been made. If the quantities added after the calculation of $A^{(r)}$, $X^{(r)}$ are the matrices S_r, S'_r we shall have

$$\begin{aligned} J_n[\ldots\{J_2(J_1 A+S_1)+S_2\}\ldots]+S_n &= 1, \\ J_n[\ldots\{J_2(J_1+S'_1)+S'_2\}\ldots]+S'_n &= \Xi, \end{aligned} \tag{11.1}$$

where Ξ represents the actual matrix obtained at the end of the calculation as the value of A^{-1}.

The equations (11.1) give us

$$\begin{aligned} A+\sum X_r^{-1} S_r &= X_n^{-1}, \\ 1+\sum X_r^{-1} S'_r &= X_n^{-1}\Xi \end{aligned} \tag{11.2}$$

and hence

$$\Xi = \Big(1+A^{-1}\sum_r X_r^{-1} S_r\Big) A^{-1} \Big(1+\sum_r X_r^{-1} S'_r\Big). \tag{11.3}$$

The matrix $X_r A$ is the result of the first r stages of the reduction of A and agrees with D in the first r columns. This fact may be expressed in the equation

$$(X_r A-1)I_r = 0, \tag{11.4}$$

where I_r is that matrix which agrees with the unit matrix in the first r columns and with the zero matrix elsewhere. It is also clear that X_r differs

from the unit matrix only in the first r columns; this fact may be expressed in the equation

$$(\mathbf{X}_r-1)(1-\mathbf{I}_r) = 0. \tag{11.5}$$

From (11.4) and (11.5) we now find $\mathbf{X}_r^{-1}$;

$$\mathbf{X}_r^{-1} = \mathbf{A}\mathbf{I}_r+1-\mathbf{I}_r. \tag{11.6}$$

When we ignore the second-order terms in the rounding-off errors (11.3), (11.6) give us

$$\begin{aligned}\Xi-\mathbf{A}^{-1} &= -\mathbf{A}^{-1}\Big(\sum_r \mathbf{X}_r^{-1}\mathbf{S}_r\Big)\mathbf{A}^{-1}+\mathbf{A}^{-1}\sum_r \mathbf{X}_r^{-1}\mathbf{S}'_r \\ &= \sum_r \{\mathbf{I}_r+\mathbf{A}^{-1}(1-\mathbf{I}_r)\}(\mathbf{S}_r\mathbf{A}^{-1}-\mathbf{S}'_r).\end{aligned} \tag{11.7}$$

Let us now assume that each coefficient $\mathbf{S}_r$ is at most ϵ and each coefficient of $\mathbf{S}'_r$ at most ϵ'. From (11.7) we can estimate the error in M-measure

$$\begin{aligned}M(\Xi-\mathbf{A}^{-1}) &\leqslant \sum_r n\{1+M(\mathbf{A}^{-1})\}M(\mathbf{S}_r\mathbf{A}^{-1}-\mathbf{S}'_r) \\ &\leqslant \sum_r n\{1+M(\mathbf{A}^{-1})\}\{n\epsilon M(\mathbf{A}^{-1})+\epsilon'\} \\ &\leqslant n^2\{1+M(\mathbf{A}^{-1})\}\{\epsilon'+n\epsilon M(\mathbf{A}^{-1})\},\end{aligned} \tag{11.8}$$

or in B-measure,

$$\begin{aligned}B(\Xi-\mathbf{A}^{-1}) &\leqslant \sum_r B\{\mathbf{I}_r+\mathbf{A}^{-1}(1-\mathbf{I}_r)\}\{B(\mathbf{S}_r\mathbf{A}^{-1})+B(\mathbf{S}'_r)\} \\ &\leqslant \sum_r \{1+B(\mathbf{A}^{-1})\}\{\epsilon n^{\frac{1}{2}}B(\mathbf{A}^{-1})+\epsilon' n^{\frac{1}{2}}\} \\ &\leqslant n^{\frac{1}{2}}\{1+B(\mathbf{A}^{-1})\}\{\epsilon B(\mathbf{A}^{-1})+\epsilon'\},\end{aligned} \tag{11.9}$$

or in N-measure,

$$\begin{aligned}N(\Xi-\mathbf{A}^{-1}) &\leqslant \sum_r N\{\mathbf{I}_r+\mathbf{A}^{-1}(1-\mathbf{I}_r)\}\{N(\mathbf{S}_r\mathbf{A}^{-1})+N(\mathbf{S}'_r)\} \\ &\leqslant \sum_r \{r^{\frac{1}{2}}+(1-r)^{\frac{1}{2}}N(\mathbf{A}^{-1})\}\{n\epsilon N(\mathbf{A}^{-1})+n\epsilon'\} \\ &\leqslant \tfrac{2}{3}(n+1)^{\frac{1}{2}}\{1+N(\mathbf{A}^{-1})\}\{\epsilon'+\epsilon N(\mathbf{A}^{-1})\}.\end{aligned} \tag{11.10}$$

If we use the relations $\mathbf{S}_r\mathbf{I}_r = \mathbf{S}'_r(1-\mathbf{I}_r) = 0$, which follow from the restrictions on the coefficients which can suffer rounding-off errors, (11.8) may be improved to

$$M(\Xi-\mathbf{A}^{-1}) \leqslant n\epsilon'+\frac{n(n-1)}{2}M(\mathbf{A}^{-1})\Big\{\epsilon+\epsilon'+\frac{2n-1}{3}\epsilon M(\mathbf{A}^{-1})\Big\}. \tag{11.11}$$

This result is best possible in the sense that given ϵ, ϵ', M we can find $\mathbf{S}_r$, $\mathbf{S}'_r$, $\mathbf{A}$ so that $M(\mathbf{S}_r) \leqslant \epsilon$, $M(\mathbf{S}'_r) \leqslant \epsilon'$, $M(\mathbf{A}^{-1}) = M$ and the error $M(\Xi-\mathbf{A}^{-1})$, still ignoring second-order terms, is exactly

$$n\epsilon'+\frac{n(n-1)}{2}M\Big(\epsilon+\epsilon'+\frac{2n-1}{3}\epsilon M\Big).$$

We may also use (11.7) to give us an estimate of the statistical error. Let the coefficients of the matrices $\mathbf{S}_1,\ldots,\mathbf{S}_n$ which are not obliged to be

0 be $s_1,\ldots, s_K$ in some order, and likewise let the coefficients of $\mathbf{S}'_1,\ldots, \mathbf{S}'_n$ which are not necessarily zero be $s_{K+1},\ldots, s_P$. The equation (11.7) may then be put in the form

$$(\Xi-\mathbf{A}^{-1})_{ij} = \sum_{u=1}^{P} c_{iju}\, s_u,$$

where c_{iju} depends only on the coefficients of $\mathbf{A}^{-1}$. Suppose that the rounding-off errors s_u are independent and have standard deviation σ_u and zero mean, then the mean square value of $(\Xi-\mathbf{A}^{-1})_{ij}$ is $\sum_{u=1}^{P} c_{iju}^2\, \sigma_u^2$. Let us put $\sigma_u = \eta$ for $u \leqslant K$, $\sigma_u = \eta'$ for $u > K$ and the mean square error in $\mathbf{A}_{ij}^{-1}$ becomes $\eta^2 \sum_{u=1}^{K} c_{iju}^2 + \eta'^2 \sum_{u=K+1}^{P} c_{iju}^2$. When we substitute in the correct values for c_{iju} we obtain:

mean square error in $(\mathbf{A}^{-1})_{ij}$

$$= \eta^2 \sum_{m,K} (\mathbf{A}^{-1})_{im}^2 (\mathbf{A}^{-1})_{Kj}^2 \min(K, i-1) + \eta^2 \sum_{K>i} (\mathbf{A}^{-1})_{Kj}^2 (K-i) + \\ + \eta'^2 \left[\sum_{m} (\mathbf{A}^{-1})_{im}^2 \min(j, m-1) + \frac{(n-1)(n-i+1)}{2} \right],$$

where η is the standard deviation and zero the mean of each coefficient of $\mathbf{S}_r$, and η' is the standard deviation and zero the mean of each coefficient of $\mathbf{S}'_r$.

Also

mean square error in $(\mathbf{A}^{-1})_{ij}$

$$\leqslant \eta^2 \left[\{M(\mathbf{A}^{-1})\}^4 \frac{n(n+1)(n-\frac{1}{2})}{3} + \{M(\mathbf{A}^{-1})\}^2 \frac{(n-1)(n-i+1)}{2} \right] + \\ + \eta'^2 \left[\{M(\mathbf{A}^{-1})\}^2 (n-\tfrac{1}{2}-\tfrac{1}{2}j) + \frac{(n-i)(n-i+1)}{2} \right].$$

The leading term in the R.M.S. error in $(\mathbf{A}^{-1})_{ij}$ is therefore at most

$$\eta \{M(\mathbf{A}^{-1})\}^2 \frac{n^{\frac{3}{2}}}{\sqrt{3}}.$$

The assumptions $M(\mathbf{S}_r) < \epsilon$, $M(\mathbf{S}'_r) < \epsilon'$ in the above analysis state in effect that we are working to a fixed number of decimal places both in the reduction of the original matrix to unity and in the building up of the inverse. It is not easy to obtain corresponding results for the case where a definite number of *significant* figures are kept, but we may make some qualitative suggestions.

The error when working with a fixed number of decimal places arose almost entirely from the reduction of the original matrix, and very little from the building up of the inverse. This, at any rate, applies for the inversion of ill-conditioned matrices with coefficients of moderate size.

However, the coefficients of the inverse are larger than those of the original matrix, so that if we work to the same number of significant figures in both we may expect the discrepancy to disappear. The general idea of this may be expressed by putting

$$M(\mathbf{S}_r) < \delta M(\mathbf{A}), \qquad M(\mathbf{S}'_r) < \delta' M(\mathbf{A}^{-1}),$$

so that $$\frac{M(\Xi - \mathbf{A}^{-1})}{M(\mathbf{A}^{-1})} < n^3 M(\mathbf{A}) M(\mathbf{A}^{-1}) \left(1 + \frac{1}{M(\mathbf{A}^{-1})}\right)\left(\delta + \frac{\delta'}{M(\mathbf{A})}\right).$$

There still remains the factor $\frac{1}{M(\mathbf{A})}$ multiplying δ'. This could be removed by arranging to reduce A, not to the unit matrix, **1**, but to $M(\mathbf{A}).\mathbf{1}$. This would be a reasonable procedure in any case, though it would be more convenient to choose the nearest power of 10 to take the place of $M(\mathbf{A})$. We see now that it is the M-condition number $nM(\mathbf{A})M(\mathbf{A}^{-1})$ which determines the magnitude of the errors when we work to a definite number of figures.

In the case of positive definite, symmetric matrices it is possible to give more definite estimates for the case where calculation is limited to a specific number of significant figures. Results of this nature have been obtained by J. v. Neumann and H. H. Goldstine (ref. 6).

It is instructive to compare the estimates of error given above with the errors liable to arise from the inaccuracy of the original matrix. If we desire the inverse of A, but the figures given to us are not those of A but of $\mathbf{A}-\mathbf{S}$, then if we invert perfectly correctly we shall get $(\mathbf{A}-\mathbf{S})^{-1}$ instead of $\mathbf{A}^{-1}$, that is, we shall make an error of $(\mathbf{A}-\mathbf{S})^{-1}-\mathbf{A}^{-1}$, i.e. of

$$(1-\mathbf{A}^{-1}\mathbf{S})^{-1}\mathbf{A}^{-1}\mathbf{S}\mathbf{A}^{-1}.$$

If we ignore the second-order terms this is $\mathbf{A}^{-1}\mathbf{S}\mathbf{A}^{-1}$. The leading terms in the error in the Jordan method were $\mathbf{A}^{-1}\left(\sum_r (1-\mathbf{I}_r)\mathbf{S}_r\right)\mathbf{A}^{-1}$ so that we might say that the greater part of the error is equal to that error which would have been produced by an original error in the matrix of $\sum_r (1-\mathbf{I}_r)\mathbf{S}_r$.

It is possible to give error estimates also for several others amongst the methods suggested elsewhere in this paper. This is, for instance, the case for the elimination method.

The elimination method in its first phase proceeds similarly to the Jordan process, but we only attempt to reduce A to a triangle and not to a diagonal: also the matrix representing the complete operation in this first phase is triangular.

12. Errors in the Gauss elimination process

We will consider the errors in the Gauss elimination process as consisting of two parts, one arising from the reduction of the matrix to the

triangular form, and the other from the back-substitution. Of these we are mainly interested in the error arising from the reduction, since this is the part of the process which has been most criticized. We adopt the description of the process given in § 4, note (1), and observe that apart from a slight difference in the form of the matrices $\mathbf{J}_r$, the reduction is similar to the Jordan process. As in the Jordan process, we shall assume that we make matrix errors $\mathbf{S}_1$, $\mathbf{S}_2$,..., $\mathbf{S}_n$ in the various stages of the reduction of $\mathbf{A}$, and vector errors $\mathbf{s}_1$, $\mathbf{s}_2$,..., $\mathbf{s}_n$ in the operations on $\mathbf{b}$. Assuming there are no back-substitution errors, and ignoring the second-order terms in the errors we should have:

$$\text{error in } \mathbf{x} = \mathbf{U}^{-1}\mathbf{X}_n \sum_{r=1}^{n} \mathbf{X}_r^{-1}(\mathbf{s}_r' - \mathbf{S}_r \mathbf{U}^{-1}\mathbf{X}_n \mathbf{b}),$$

where $\mathbf{X}_r = \mathbf{J}_r \ldots \mathbf{J}_1$. Now, assuming that the process has been done with the largest pivot chosen from each column, we shall have $M(\mathbf{X}_r^{-1}) = 1$, for $\mathbf{X}_r^{-1} = 1 + \sum_{s \leqslant r} (1 - \mathbf{J}_s)$ as mentioned in § 4 (2). Then

$$\begin{aligned} |\text{error in } \mathbf{x}_m| &= |(\mathbf{A}^{-1} \textstyle\sum \mathbf{X}_r^{-1}(\mathbf{s}_r' - \mathbf{S}_r \mathbf{A}^{-1}\mathbf{b}))_m| \\ &= \left| \sum_{\substack{j,k,r \\ j \geqslant k}} (\mathbf{A}^{-1})_{mj}(\mathbf{X}_r^{-1})_{jk}(\mathbf{s}_r')_k - \sum_{l,p} (\mathbf{S}_r)_{kl}(\mathbf{A}^{-1})_{lp}\mathbf{b}_p \right| \\ &\leqslant \frac{n^2(n+1)}{2} M(\mathbf{A}^{-1})\epsilon' + \frac{n^4(n+1)}{2} \{M(\mathbf{A}^{-1})\}^2 M(\mathbf{b})\epsilon, \end{aligned}$$

where $M(\mathbf{s}_r') \leqslant \epsilon'$, $M(\mathbf{S}_r) \leqslant \epsilon$.

To these errors we have to add those which arise from the back-substitution. This consists in solving the equations $\mathbf{DUx} = \mathbf{L}^{-1}\mathbf{b}$, where $\mathbf{U}$ is unit upper triangular and $\mathbf{D}$ diagonal. We obtain x_n first and then x_r in order of decreasing r by means of the formula $x_r = \mathbf{d}_r^{-1}(\mathbf{L}^{-1}\mathbf{b})_r - \sum_{i>r} (\mathbf{DU})_{ri} x_i$. Now if we make an error of t_r in the calculation of x_r from the previously obtained coefficients of $\mathbf{x}$, then we shall have solved accurately the equations $\mathbf{DUx} = \mathbf{L}^{-1}\mathbf{b} + \mathbf{Dt}$, that is, we shall have introduced an error of $\mathbf{U}^{-1}\mathbf{t}$, or, since $\mathbf{A} = \mathbf{LDU}$, of $\mathbf{A}^{-1}\mathbf{LDt}$. If we arrange that $M|t_r| \leqslant \epsilon d_r^{-1}$, the greatest error in any coefficient from this source is $n^2 M(\mathbf{A}^{-1})\epsilon$, and normally much smaller than the error arising from the first part of the process. Furthermore, d_r will normally tend to be less than 1.

It is interesting to note the value of the error in the last pivot, that is, the error in the (nn) coefficient of $\mathbf{J}_n \ldots \mathbf{J}_1 \mathbf{A}$. The matrix error in $\mathbf{J}_n \ldots \mathbf{J}_1 \mathbf{A}$ is $\mathbf{X}_n \sum_r \mathbf{X}_r^{-1}\mathbf{S}_r$, that is, since $\mathbf{X}_n \mathbf{L}^{-1} = \mathbf{DUA}^{-1}$, it is $\mathbf{DUA}^{-1} \sum_r \mathbf{X}_r^{-1}\mathbf{S}_r$. The (nn) coefficient is $d_n \sum_r a_{nj}(\mathbf{X}_r^{-1}\mathbf{S}_r)_{jn}$ and since $M(\mathbf{X}_r^{-1}) = 1$, $M(\mathbf{S}_r) \leqslant \epsilon$ it does not exceed $n^2 d_n M(\mathbf{A}^{-1})\epsilon$ in absolute magnitude, that is, the proportionate error in the last pivot is at most $n^2 M(\mathbf{A}^{-1})\epsilon$. This cannot be very large

unless the matrix is ill-conditioned. With worst possible conditioning we find an error somewhat similar to Hotelling's estimate. The matrix error in $\mathbf{J}_n \ldots \mathbf{J}_1 \mathbf{A}$ may be written $\mathbf{L}^{-1} \sum_r \mathbf{X}_r^{-1} \mathbf{S}_r$, from which we find that the error in the last pivot cannot exceed $n^2 \epsilon M(\mathbf{L}^{-1})$. But since $M(\mathbf{L}) = 1$ we find $M(\mathbf{L}^{-1}) \leqslant 2^{n-1}$ (and equality can be attained): that this error may actually be as great as $2^{n-2}\epsilon$ may be seen by considering the inversion of a matrix differing only slightly from

$$\begin{bmatrix} 1 & 0 & 0 & . & . & . & 0 \\ -1 & 1 & 0 & . & . & . & 0 \\ -1 & -1 & 1 & . & . & . & 0 \\ . & . & . & . & . & . & . \\ -1 & -1 & -1 & . & . & . & 1 \end{bmatrix}$$

It appears then that the error in the last pivot can only be large if $\mathbf{L}^{-1}$ is large, and that this can only happen with ill-conditioned equations. Actually even then we may consider ourselves very unlucky if $\mathbf{L}^{-1}$ is large. Normally, even with ill-conditioned equations we may expect the off-diagonal coefficients of $\mathbf{L}$ to be distributed fairly uniformly between -1 and 1, possibly with a tendency to be near 0. Only when there is a strong tendency for negative values will we find a large $\mathbf{L}^{-1}$.

13. Errors in the unsymmetrical Choleski method

When obtaining the triangular resolution of a matrix by the method of the theorem (§ 3) it is convenient to think of the process as follows. We are given a matrix $\mathbf{A}$ and the matrices $\mathbf{L}$ and $\mathbf{DU}$ ($= \mathbf{W}$, say). We form the product $\mathbf{LW}$ coefficient by coefficient. When calculating any one of the coefficients of $\mathbf{LW}$, we always find that the data are incomplete to the extent of one number, and we therefore choose this number so as to give the required coefficient in $\mathbf{A}$. The unknown quantity when forming a_{ij} is always either l_{ij} or w_{ij}. Regarding the process in this way suggests the following rule for deciding the number of figures to be retained. We always retain sufficient figures to give us an error of not more than ϵ in the coefficient of $\mathbf{A}$ under consideration. In actual hand computation this rule is extremely simple to apply. Suppose, for example, that ϵ is $\frac{1}{2}10^{-7}$ and that we are forming the product $(\mathbf{LW})_{94}$, i.e. $\sum_{j=1}^{4} l_{9j} w_{j4}$. We first form $\sum_{j=1}^{3} l_{9j} w_{j4}$ accumulating the products in the machine. All the relevant quantities should be available at this stage. We then set up the multiplicand w_{44} which should also be known and 'turn the handle' until the quantity in the product register, rounded off to seven figures, first agrees

with the given value of a_{94} (which is assumed to have zeros in the eighth and later figures). All the figures in the multiplier register are then written down as the value of l_{94}.

The theory of the errors in this method is peculiarly simple. The triangular resolution obtained is an exact resolution of a matrix $\mathbf{A}-\mathbf{S}$, where $M(\mathbf{S}) < \epsilon$, and the resultant error in the inverse is $\mathbf{A}^{-1}\mathbf{S}\mathbf{A}^{-1}$, and in any coefficient at most $n^2\{M(\mathbf{A}^{-1})\}^2\epsilon$. A similar procedure is appropriate in the inversion of the triangles $\mathbf{L}$ and $\mathbf{W}$. When inverting $\mathbf{W}$ (say) we can arrange, by an exactly similar computing procedure, that its product with its reputed inverse differs from unity by at most ϵ' in each coefficient, i.e. $\mathbf{LK} = 1-\mathbf{S}'$, where $M(\mathbf{S}') < \epsilon$ and $\mathbf{K}$ is the reputed inverse. Note the order in the product which is significant. Likewise we find a reputed inverse $\mathbf{V}$ for $\mathbf{DU}$ such that $\mathbf{V}.\mathbf{DU} = 1-\mathbf{S}''$ and $M(\mathbf{S}'') < \epsilon'$. The error arising from using these reputed inverses is $-(1-\mathbf{S}'')^{-1}\mathbf{VK}(1-\mathbf{S}')+\mathbf{VK}$, or neglecting second-order terms, $\mathbf{S}''\mathbf{A}^{-1}+\mathbf{A}^{-1}\mathbf{S}'$. Finally, there is a possible source of error due to rounding off in the actual formation of the product $\mathbf{VK}$. If this does not exceed ϵ'' in any coefficient, the error in any coefficient of the reputed inverse of $\mathbf{A}$ is in all at most

$$n^2\epsilon\{M(\mathbf{A}^{-1})\}^2+2n\epsilon' M(\mathbf{A}^{-1})+\epsilon''.$$

This paper is published with the permission of the Director of the National Physical Laboratory.

REFERENCES

1. Commandant BÉNOIT, 'Note sur une méthode, etc.' (Procédé du Commandant CHOLESKY), *Bull. Géod.* (Toulouse), 1924, No. 2, 5–77.
2. L. FOX, H. D. HUSKEY, and J. H. WILKINSON, 'Notes on the solution of algebraic linear simultaneous equations', see above, pp. 149–73.
3. J. v. NEUMANN, V. BARGMANN, and D. MONTGOMERY, *Solution of Linear Systems of High Order*, lithographed, Princeton (1946).
4. J. MORRIS, 'An escalator method for the solution of linear simultaneous equations', *Phil. Mag.*, series 7, **37** (1946), 106.
5. H. HOTELLING, 'Some new methods in matrix calculation', *Ann. Math. Stat.* **14** (1943), 34.
6. J. v. NEUMANN and H. H. GOLDSTINE, 'Numerical inverting of matrices of high order', *Bull. Amer. Math. Soc.* **53** (1947), 1021–99.