Error Analysis of Direct Methods of Matrix Inversion*

J. H. Wilkinson

National Physical Laboratory, Teddington, England

Introduction

1. In order to assess the relative effectiveness of methods of inverting a matrix it is useful to have *a priori* bounds for the errors in the computed inverses. In this paper we determine such error bounds for a number of the most effective direct methods. To illustrate fully the techniques we have used, some of the analysis has been done for floating-point computation and some for fixed-point. In all cases it has been assumed that the computation has been performed using a precision of t binary places, though it should be appreciated that on a computer which has both fixed and floating-point facilities the number of permissible digits in a fixed-point number is greater than the number of digits in the mantissa of a floating-point number. The techniques used for analyzing floating-point computation are essentially those of [8], and a familiarity with that paper is assumed.

2. The error bounds are most conveniently expressed in terms of vector and matrix norms, and throughout we have used the Euclidean vector norm and the spectral matrix norm except when explicit reference is made to the contrary. For convenience the main properties of these norms are given in Section 9.

In a recent paper [7] we analyzed the effect of the rounding errors made in the solution of the equations

$$Ax = b \tag{2.1}$$

by Gaussian elimination with interchanges "pivoting for size". We showed that the computed solution was the exact solution of

$$(A + E)x \equiv b + \delta b \tag{2.2}$$

and that we could obtain bounds for E and δb *provided assumptions were made about the ratio of the maximum element of the successive reduced matrices to the maximum element of* A. All the methods we discuss in this paper depend on the successive transformation of the original matrix $A^{(1)}$ into matrices $A^{(2)}, A^{(3)}, \cdots, A^{(k)}$ such that each $A^{(s)}$ is equivalent to $A^{(1)}$ and the final $A^{(k)}$ is triangular. An essential requirement in an *a priori* analysis of any of the methods, is a bound for the quantity r, defined by

$$R = \frac{\max \mid a_{ij}^{(s)} \mid}{\max \mid a_{ij}^{(1)} \mid} \qquad (s = 1, \cdots, k) \tag{2.3}$$

* Received March, 1961.

For methods based upon Gaussian elimination we have been able to obtain a reasonable upper bound for general unsymmetric matrices only when pivoting for size is used at each stage. This bound is far from sharp, but does at least lead to a useable result, and further, the proof shows the factors which limit the size of R. For a number of classes of matrices which are important in practice, a much smaller bound is given for R, and this bound holds even when a limited form of pivoting for size is employed. Finally, we show that for positive definite matrices $R \leqq 1$, even when pivoting for size is not used.

For comparison with methods of this Gaussian type, we give an analysis of methods which are based on the reduction of A to triangular form by equivalent transformations with orthogonal matrices. For these it is easy to show that for any matrix $R < \sqrt{n}$, so that control of size is assured. Because of this limitation on the size of R, the error bounds obtained for orthogonal reductions of a general matrix are smaller than those that have been obtained for elimination methods. However, orthogonal reduction requires considerably more computation than does elimination, and since in practice the value of R falls so far short of the bound we obtained it is our opinion that the use of the orthogonal transformations is seldom, if ever, justified.

3. We do not claim that any of the *a priori* bounds we have given are in any way "best possible", and in a number of cases we know that they could have been reduced somewhat by a more elaborate argument. We feel, however, that there is a danger that the essential simplicity of the error analysis may be obscured by an excess of detail. In any case the bounds take no account of the statistical effect and, in practice, this invariably reduces the true error by a far more substantial factor than can be achieved by an elaboration of the arguments. We also adduce arguments which partly explain why it is so often true in practice that computed inverses are far more accurate than might have been expected even making allowances for the statistical effect.

All the *a priori* bounds for the error contain explicitly the factor $\| A^{-1} \|$. This means that we will hardly ever be in a position to make use of the results before obtaining the computed inverse. The nature of the analysis makes it obvious how to use the computed inverse to obtain an *a posteriori* upper bound for its error but the main value of the results is to indicate the range of applicability of a given method for a given precision in the computation, and they should be judged in this light.

Upper Bound for R in Gaussian Elimination

4. We derive first an upper bound for R when a general matrix is reduced to triangular form by Gaussian elimination, selecting as pivotal element at each stage the element of maximum modulus in the whole of the remaining square matrix. We refer to this as "complete" pivoting for size, in contrast to the selection of the maximum element in the leading column at each stage, which we call "partial" pivoting for size.

We denote the original matrix by $A^{(n)}$ and the reduced matrices by $A^{(n-1)}$,

$A^{(n-2)}, \cdots, A^{(1)}$ (so that $A^{(r)}$ is a matrix of order r) and the modulus of the pivotal element of $A^{(r)}$ by p_r so that

$$| a_{ij}^{(r)} | \leqq p_r . \tag{4.1}$$

Then we have

$$| \det A^{(r)} | = p_r p_{r-1} \cdots p_1 \qquad (r = n, n-1, \cdots, 2). \tag{4.2}$$

On the other hand, Hadamard's Theorem gives

$$| \det A^{(r)} | \leqq [rp_r]^r \qquad (r = n, \cdots, 2) \tag{4.3}$$

since the length of every column of $A^{(r)}$ is bounded by $r^{\frac{1}{2}} p_r$. We write

$$p_r p_{r-1} \cdots p_1 \leqq [r^{\frac{1}{2}} p_r]^r \qquad (r = 2, 3, \cdots, n-1) \tag{4.4}$$

but retain

$$p_n p_{n-1} \cdots p_1 = | \det A^{(n)} | . \tag{4.5}$$

Taking logarithms of relations (4.4) and (4.5) and writing

$$\log p_r = q_r \tag{4.6}$$

we have

$$\sum_1^{r-1} q_i \leqq \frac{r}{2} \log r + (r-1) q_r \qquad (r = 2, 3, \cdots, n-1) \tag{4.7}$$

$$\sum_1^n q_i = \log | \det A^{(n)} | . \tag{4.8}$$

Dividing (4.7) by $r(r-1)$ for $r = 2, 3, \cdots, n-1$ and (4.8) by $(n-1)$ and adding we have, on observing that

$$\frac{1}{r(r-1)} + \frac{1}{(r+1)r} + \cdots + \frac{1}{(n-1)(n-2)} + \frac{1}{n-1} = \frac{1}{r-1}, \tag{4.9}$$

$$\begin{aligned} \frac{q_1}{1} + \frac{q_2}{2} + \cdots + \frac{q_{n-2}}{n-2} + \frac{q_{n-1}}{n-1} + \frac{q_n}{n-1} \\ \leqq \frac{1}{2} \log 2^1 3^{\frac{1}{2}} 4^{\frac{1}{3}} \cdots (n-1)^{\frac{1}{n-2}} + \frac{1}{n-1} \log | \det A^{(n)} | \\ + \frac{q_2}{2} + \frac{q_3}{3} + \cdots + \frac{q_{n-1}}{n-1}, \end{aligned} \tag{4.10}$$

giving

$$q_1 + \frac{q_n}{n-1} \leqq \log f(n-1) + \frac{1}{n-1} \log | \det A^{(n)} | , \tag{4.11}$$

where

$$f(s) = [2^1 3^{\frac{1}{2}} \cdots s^{\frac{1}{s-1}}]^{\frac{1}{2}}. \tag{4.12}$$

Substituting for $| \det A^{(n)} |$ from (4.3) (with $r = n$) we have

$$q_1 + \frac{q_n}{n-1} \leqq \log f(n-1) + \frac{n}{2(n-1)} \log n + \frac{nq_n}{n-1}, \qquad (4.13)$$

$$q_1 - q_n \leqq \log f(n) + \tfrac{1}{2} \log n. \qquad (4.14)$$

Hence

$$p_1/p_n \leqq \sqrt{n} f(n) = g(n) \quad \text{(say)}. \qquad (4.15)$$

This gives an upper bound for the ratio of the last pivot to the first. In particular, if $p_n \leqq 1$ then $p_1 \leqq g(n)$.

On the other hand if we write

$$l = \max_j \left[\sum_{i=1}^{n} (a_{ij}^{(n)})^2\right]^{\frac{1}{2}}$$

so that all columns of $A^{(n)}$ have lengths which are bounded by l, then

$$|\det A^{(n)}| \leqq l^n \quad \text{and} \quad p_n \geqq \frac{l}{\sqrt{n}}. \qquad (4.16)$$

Hence from (4.11)

$$\begin{aligned} q_1 &\leqq \log f(n-1) + \frac{\log n}{2\,(n-1)} + \log l \\ &= \log f(n) + \log l. \end{aligned} \qquad (4.17)$$

In particular if $l \leqq 1$ we have

$$p_1 \leqq f(n). \qquad (4.18)$$

The functions $f(n)$ and $g(n)$ increase comparatively slowly and it is easy to show that

$$f(n) \sim Cn^{\frac{1}{4}\log n}. \qquad (4.19)$$

In Table 1 we show the values of $f(n)$ and $g(n)$ for representative values of n. The two forms of normalization

$$\text{(I)} \qquad \tfrac{1}{2} \leqq \max_{i,j} |a_{ij}| \leqq 1 \qquad (4.20)$$

$$\text{(II)} \qquad \frac{1}{2} \leqq \left[\sum_{i=1}^{n} a_{ij}^2\right]^{\frac{1}{2}} \leqq 1 \qquad (j = 1, \cdots, n) \quad (4.21)$$

will be used throughout this paper and will be called normalization (I) and (II) respectively. The results (4.15) and (4.18) hold for any matrix normalized in the appropriate way, and indeed their truth is in no way dependent on the left-hand inequalities in (4.20) and (4.21). However, in our opinion, pivoting for size is a reasonable strategy only if all rows and columns of the original matrix have comparable norms.

For normalization (I) this may be achieved by scaling all rows and columns so that they contain at least one element satisfying $\frac{1}{2} \leqq |a_{ij}| < 1$. For normalization (II) this should be followed by an additional scaling of the columns so that (4.21) is satisfied. F. L. Bauer has suggested that such matrices should be termed "equilibrated" matrices. In practice it is unnecessary to ensure exact

TABLE 1

n	10	20	50	100	200	1,000
$f(n)$	6.1	15	75	330	1,800	250,000
$g(n)$	19	67	530	3300	26,000	7,900,000

equilibration, but merely that there are no rows or columns consisting entirely of "small" elements. Strictly speaking, equilibration should be performed at each stage of the reduction before selecting the pivot, but these later equilibrations are of much less importance.

5. The inequality (4.15) is certainly not sharp. For the equality could hold only if the equality held in (4.4) for all r. If this were true we could show, exactly as we have for $r = n$, that

$$p_1/p_r = g(r) \qquad (r = 2, \cdots, n) \quad (5.1)$$

and hence

$$p_{r-1}/p_r = g(r)/g(r-1) = \left[\frac{r}{r-1} r^{1/r-1}\right]^{\frac{1}{2}}. \quad (5.2)$$

Except for small r, this ratio is only slightly greater than unity. This means that the pivotal sequence p_n, p_{n-1}, $\cdots$ increases quite slowly at first. Now equality in (4.4) can hold only if all elements of $A^{(r)}$ are equal to $\pm p_r$ and all columns are orthogonal. This is clearly not possible for some values of r, for example, $r = 3$. Further if $a_{ij}^{(r)} = \pm p_r$ then all $a_{ij}^{(r-1)}$ are $\pm 2p_r$ or zero. Hence if at any stage the equality sign holds in (4.4) then either $p_{r-1}/p_r = 2$ or $A^{(r-1)}$ is null, and in either case (5.2) cannot be satisfied.

These considerations suggest that the least upper bound for the pivots may be much smaller than the one we have given. In our experience the last pivot has fallen far below the given bound, and no matrix has been encountered in practice for which p_1/p_n was as large as 8. Note that we must certainly have

$$p_r \leqq p_n g(n + 1 - r) \quad (5.3)$$

since p_r is the final pivot in the reduction of a matrix of order $(n + 1 - r)$, though if the maximum is to be approached by p_1 then the initial rate of growth must be much slower than that corresponding to (5.3).

Positive Definite Matrices

6. There are a number of special types of matrices for which a better bound can be given for the largest pivot and, surprisingly, none of these require complete pivoting for size.

The most important class is that of positive definite matrices. We shall show that when a positive definite matrix is reduced by Gaussian elimination *without any pivoting for size* then no element in any reduced matrix exceeds the maximum element in the original matrix. It is convenient now to call the original matrix $A^{(1)}$ and the reduced matrices $A^{(2)}$, $A^{(3)}$, $\cdots$, $A^{(20)}$ so that $A^{(r)}$ is of order

$$(n - r + 1).$$

We first show that $A^{(2)}$ is positive definite.

The multipliers m_{i1} used in the first reduction are defined by

$$m_{i1} = \frac{a_{i1}^{(1)}}{a_{11}^{(1)}}, \tag{6.1}$$

and $a_{11}^{(1)}$ is non-zero, because $A^{(1)}$ is positive definite. The elements of $A^{(2)}$ are given by

$$\begin{aligned} a_{ij}^{(2)} &= a_{ij}^{(1)} - m_{i1} a_{1j}^{(1)} \\ &= a_{i1}^{(1)} - \frac{a_{ij}^{(1)} a_{1j}^{(1)}}{a_{11}^{(1)}} = a_{ji}^{(2)} \end{aligned} \tag{6.2}$$

from the symmetry of $A^{(1)}$. Now we have

$$\begin{aligned} \sum_{i=1}^{n} \sum_{j=1}^{n} a_{ij}^{(1)} x_i x_j - a_{11}^{(1)} \left[x_1 + \sum_{i=2}^{n} \frac{a_{i1}^{(1)}}{a_{11}^{(1)}} x_i \right]^2 &= \sum_{i=2}^{n} \sum_{j=2}^{n} \left(a_{ij}^{(1)} - \frac{a_{i1}^{(1)} a_{j1}^{(1)}}{a_{11}^{(1)}} \right) x_i x_j \\ &= \sum_{i=2}^{n} \sum_{j=2}^{n} a_{ij}^{(2)} x_i x_j . \end{aligned} \tag{6.3}$$

If $A^{(2)}$ is not positive definite there is a non-null set of $x_2, x_3, \cdots, x_n$ such that $\sum a_{ij}^{(2)} x_i x_j \leqq 0$. If we write

$$x_1 = -\sum_{i=2}^{n} \frac{a_{i1}^{(1)}}{a_{11}^{(1)}} x_i , \tag{6.4}$$

then with this $x_1, x_2, \cdots, x_n$ we have

$$\sum_{i=1}^{n} \sum_{j=1}^{n} a_{ij}^{(1)} x_i x_j \leqq 0, \tag{6.5}$$

which is impossible since $A^{(1)}$ is positive definite. We have further

$$a_{ii}^{(2)} = a_{ii}^{(1)} - \frac{a_{i1}^{(1)} a_{i1}^{(1)}}{a_{11}^{(1)}} \leqq a_{ii}^{(1)}, \tag{6.6}$$

and $a_{ii}^{(2)}$ is positive because $A^{(2)}$ is positive definite. The diagonal terms of $A^{(2)}$ are therefore not greater than those of $A^{(1)}$ and, since the maximum element of a positive definite matrix lies on the diagonal

$$\max \mid a_{ij}^{(2)} \mid \leqq \max \mid a_{ij}^{(1)} \mid . \tag{6.7}$$

In the same way it follows that $A^{(3)} \cdots A^{(n)}$ are positive definite. In particular if $\mid a_{ij}^{(1)} \mid \leqq 1$ (all i, j), then this is true of all $a_{ij}^{(r)}$.

The multipliers may be large, as for example for the matrix

$$A = \begin{bmatrix} .000{,}002 & .001{,}300 \\ .001{,}300 & .900{,}000 \end{bmatrix}, \tag{6.8}$$

so Gaussian elimination must either be performed in floating-point arithmetic, or otherwise it requires the introduction of scale-factors in a manner similar to that which is normally employed in the back-substitution.

Triangular Decomposition

7. For the symmetrical Cholesky decomposition of a positive definite symmetric matrix we may prove an even more satisfactory result. We show that if A is positive definite and $|a_{ij}| \leqq 1$, then there exists a lower triangular matrix L such that

$$LL^T = A, \tag{7.1}$$

$$|l_{ij}| \leqq 1. \tag{7.2}$$

The proof is by induction. Assume it is true for matrices of orders up to $(n-1)$. If A is a positive definite matrix of order n, we may partition it in the form

$$A = \left[\begin{array}{c|c} B & a \\ \hline a^T & a_{nn} \end{array}\right] \tag{7.3}$$

where B is a positive definite matrix of order $(n-1)$ and a is a vector of order $(n-1)$. By hypothesis there is an L satisfying

$$LL^T = B, \qquad |l_{ij}| \leqq 1. \tag{7.4}$$

If we choose l so that

$$Ll = a, \tag{7.5}$$

which can clearly be done since L is triangular and non-singular, then

$$\left[\begin{array}{c|c} L & 0 \\ \hline l^T & l_{nn} \end{array}\right] \left[\begin{array}{c|c} L^T & l \\ \hline 0 & l_{nn} \end{array}\right] = \left[\begin{array}{c|c} B & a \\ \hline a^T & a^{nn} \end{array}\right] = A, \tag{7.6}$$

where

$$l^T l + (l_{nn})^2 = a_{nn}\,. \tag{7.7}$$

Taking determinants of both sides

$$\det L \det L^T l_{nn}^2 = (\det L)^2 l_{nn}^2 = \det A. \tag{7.8}$$

Now $\det A$ is positive since A is positive-definite; hence l_{nn}^2 is positive and l_{nn} is real and may be taken to be positive. If the components of l are denoted by $l_{n1}, l_{n2}, \cdots, l_{n,n-1}$, then from equation (7.7)

$$\sum_{i=1}^{n} l_{ni}^2 = a_{nn} \leqq 1 \tag{7.9}$$

and every l_{ni} satisfies $|l_{ni}| \leqq 1$. Equation (7.6) and (7.7) now give the appropriate decomposition of A and all elements of the triangular matrix are between ±1. A consequence of this is that an LL^T decomposition may be performed in fixed-point arithmetic. Incidentally we have proved that the sum of the squares of the elements in any row of L is less than unity, and this result is used later.

Other Matrices for which $R < g(n)$

8. Matrices of Hessenberg form are of particular interest in connection with the eigenvalue problem. We consider now the reduction of an upper Hessenberg matrix with normalization I, using "partial" pivoting for size. At each stage there are only two elements in the leading column of the reduced matrix and one of these is in a row which has not yet been modified. Since even partial pivoting for size ensures that the multipliers are between ± 1 it is easy to see that the maximum element in the rth reduced matrix lies between $\pm r$, so that the maximum pivotal value is n and this can be achieved only by the last pivot. Further if we call the final triangular matrix U, and M is the matrix by which A has been multiplied to give U, so that

$$MA = U, \tag{8.1}$$

it is evident that M is lower triangular *and all its elements are less than or equal to unity in modulus.* If we write, in the usual way,

$$A = LU \tag{8.2}$$

then $L = M^{-1}$ and we have shown that all elements of L^{-1} lie between ± 1.

A special type of Hessenberg matrix is a triple-diagonal form. By much the same argument it is evident that the maximum elements of U now lie between ± 2 and again all elements of L^{-1} lie betwen ± 1. Further $\sum_j | u_{ij} | \leqq 3$.

Finally we consider matrices with dominating diagonal elements, such that

$$| a_{kk}^{(1)} | > \sum_{l \neq k} | a_{lk}^{(1)} | \qquad (k = 1, 2, \cdots, n). \tag{8.3}$$

If we perform elimination without pivoting for size, then we have typically, in the first stage

$$\sum_{i=2}^{n} | m_{i1} | = \sum_{i=2}^{n} | a_{i1}^{(1)} |/| a_{11}^{(1)} | < 1. \tag{8.4}$$

Hence

$$\begin{aligned} \sum_{i=2}^{n} | a_{ij}^{(2)} | &< \sum_{i=2}^{n} [| a_{ij}^{(1)} | + | m_{i1} || a_{1j}^{(1)} |] \\ &\leqq \sum_{i=2}^{n} | a_{ij}^{(1)} | + | a_{1j}^{(1)} | \sum_{i=2}^{n} | m_{i1} | \\ &\leqq \sum_{i=2}^{n} | a_{ij}^{(1)} | + | a_{1j}^{(1)} | = \sum_{i=1}^{n} | a_{ij}^{(1)} |. \end{aligned} \tag{8.5}$$

Further

$$\begin{aligned} | a_{ii}^{(2)} | &\geqq | a_{ii}^{(1)} | - | m_{i1} || a_{1i}^{(1)} | \\ &\geqq \sum_{j \neq i} | a_{ji}^{(1)} | - (1 - \sum_{j \neq 1, i} | m_{j1} |)| a_{1i}^{(1)} | \\ &= \sum_{j \neq 1, i} \{| a_{ji}^{(1)} | + | m_{j1} || a_{1i}^{(1)} |\} \geqq \sum_{j \neq 1, i} | a_{ji}^{(2)} |. \end{aligned} \tag{8.6}$$

The diagonal terms therefore dominate in exactly the same way in all $A^{(r)}$ and, from (8.5), the sum of the moduli of the elements of any column of $A^{(r)}$ decreases as r increases. Hence

$$\begin{aligned} \max \mid a_{ij}^{(r)} \mid &\leqq \max \mid a_{ii}^{(r)} \mid \leqq \max \sum_{j=r}^{n} \mid a_{ji}^{(r)} \mid \\ &\leqq \max \sum_{j=1}^{n} \mid a_{ji}^{(1)} \mid \leqq 2 \max \mid a_{ii}^{(1)} \mid . \end{aligned} \tag{8.7}$$

Hence the ratio, R, is certainly less than 2. By similar arguments we see that if we write the reduction in the form

$$MA = U \quad \text{or} \quad LU = A, \tag{8.8}$$

then all elements of M (or L^{-1}) are less than unity.

Even when the diagonal elements are not strictly dominant, behavior of this last type is very common. For example, the matrices of the form C_n where

$$C_4 = \begin{bmatrix} a & b & b & b \\ b & a & b & b \\ b & b & a & b \\ b & b & b & a \end{bmatrix} \qquad (a > b > 0) \tag{8.9}$$

retain this character during Gaussian elimination and the diagonals become progressively more dominant even when (8.3) is not satisfied at any stage.

Many matrices which are encountered in practice do not belong strictly to any of the special classes we have mentioned, but nevertheless resemble them in character and it is our experience that when "complete" pivoting for size is used, the pivotal sequence, even if not monotonic decreasing, usually has a general downward trend. For ill-conditioned matrices this is often very marked. This is not surprising since, as is shown by (4.11), the last pivot p_1 satisfies

$$\frac{p_1}{p_n} \leqq f(n-1) \cdot \left[\frac{\mid \det A^{(n)} \mid}{p_n} \right]^{1/n-1} \Big/ p_n \tag{8.10}$$

and $\mid \det A^{(n)} \mid$ will usually be much smaller than $p_n{}^n$ if $A^{(n)}$ is ill-conditioned.

The behavior of the special matrices may give the impression that "partial" pivoting for size is as effective as "complete" pivoting. The following example shows that it may not be adequate. Consider matrices B_n of the type:

$$B_5 = \begin{bmatrix} +1 & 0 & 0 & 0 & +1 \\ +1 & +1 & 0 & 0 & -1 \\ -1 & +1 & +1 & 0 & +1 \\ +1 & -1 & +1 & +1 & -1 \\ -1 & +1 & -1 & +1 & +1 \end{bmatrix}$$

If we use partial pivoting, then the pivot is always in the top left-hand corner and the sequence is $+1, +1, \cdots, +1, (-2)^{n-1}$. If complete pivoting is used the sequence is $+1, -2, +2, +2, \cdots, +2$. We show later that we can modify B_n in such a way as to give a very *well-conditioned* matrix for which partial pivoting for size gives a completely inaccurate inverse. Although complete pivoting is not always the best strategy, we have not been able to construct an example for which it is a very bad strategy and, in any case, the analysis of Section 16 gives an upper bound for the error.

The example also illustrates the importance of equilibration. If the last column of B_n is multiplied by 2^{-n} then partial pivoting gives the same pivotal selection as complete, with disastrous results. (See Section 29.)

Some Fundamental Properties of the Spectral Norm

9. A number of results will be required repeatedly and it is convenient to assemble them here. We give first the properties of the spectral norm which we have used [9]. We assume that A is real throughout.

(A) Definition: $\| A \| = \max [\lambda(AA^T)]^{\frac{1}{2}}$.

(B) $\| A \| = \max \| Ax \| / \| x \|, \quad \| x \| \neq 0$, and hence $\| Ax \| \leqq \| A \| \, \| x \|$.

(C) $\| A \|$ satisfies the usual relations for matrix norms.
(i) $\| kA \| = | k | \| A \|$
(ii) $\| A + B \| \leqq \| A \| + \| B \|$, and hence
(iii) $\| A - B \| \geqq | \| A \| - \| B \| |$
(iv) $\| AB \| \leqq \| A \| \, \| B \|$.

(D) If $A^{(r)}$ is any principal submatrix of A, then $\| A^{(r)} \| \leqq \| A \|$ (from (B)).

(E) $\| ab^T \| = \| a \| \, \| b \|$ (from (B)).

(F) If A is symmetric, then $\| A \| = \max | \lambda_i(A) |$.

(G) If A is positive definite, $\| A \| = \max \lambda_i(A) = \lambda_1$,
$\| A^{-1} \| = \max 1/\lambda_i(A) = 1/\lambda_n$.

(H) $\| A \| = \| A^T \|$.

(I) $\| AA^T \| = \| A \|^2$.

(J) $n \max_{i,j} | a_{ij} | \geqq \| A \| \geqq \max_{i,j} | a_{ij} |$.

(K) $\| A \| \leqq \| A \|_E = [\sum \sum a_{ij}^2]^{\frac{1}{2}}$. For $\max \lambda(AA^T) \leqq \operatorname{trace}(AA^T) = \| A \|_E^2$.

(L) $\| A \| \geqq n^{-\frac{1}{2}} \| A \|_E$. For $\max \lambda(AA^T) \geqq n \operatorname{trace}(AA^T) = \frac{1}{n} \| A \|_E^2$.

(M) $\| A \| \leqq \| | A | \|$ where $| A |$ has elements $| a_{ij} |$. This follows from (B).

(N) If $| A | \leqq B$, then $\| A \| \leqq \| | A | \| \leqq \| B \|$.

(O) $\| | A | \| \leqq n^{\frac{1}{2}} \| A \|$. For $\| A \| \geqq n^{-\frac{1}{2}} \| A \|_E$ and $\| | A | \| \leqq \| | A | \|_E = \| A \|_E$.

The following results are fundamental in the error analysis and will be used repeatedly.

(P) $\| A \| \geqq \max | \lambda_i(A) |$ (from (B)).

(Q) *If* $\| A \| < 1$, *then* $I + A$ *is non-singular.*

For $\lambda_i(I + A) = 1 + \lambda_i(A)$,

$$| \lambda_i(I + A) | \geqq 1 - | \lambda_i(A) | \geqq 1 - \| A \| > 0. \tag{9.1}$$

Hence

$$\det (I + A) = \Pi \lambda_i(I + A) \neq 0. \tag{9.2}$$

(R) If $\| A \| < 1$, then

(i) $\|(I + A)^{-1} \| \leqq 1 + \dfrac{\| A \|}{1 - \| A \|} = \dfrac{1}{1 - \| A \|}$.

(ii) $\|(I + A)^{-1} - I \| \leqq \dfrac{\| A \|}{1 - \| A \|}$.

(iii) $\|(I + A)^{-1} \| \geqq 1 - \dfrac{\| A \|}{1 - \| A \|}$.

These results follow from the identity

$$I = (I + A)^{-1}(I + A), \tag{9.3}$$

the existence of $(I + A)^{-1}$ following from (Q). Writing $(I + A)^{-1} = R$,

$$I = R + RA \tag{9.4}$$

$$1 = \| I \| \geqq \| R \| - \| RA \| \geqq \| R \| - \| R \| \| A \| \tag{9.5}$$

which gives (i) since $(1 - \| A \|)$ is positive. (9.4) then gives

$$\| I - R \| = \| RA \| \leqq \| R \| \| A \| \tag{9.6}$$

which gives (ii); from (ii) we deduce

$$\| I \| - \| R \| \leqq \frac{\| A \|}{1 - \| A \|} \tag{9.7}$$

which gives (iii).

(S) If A is non-singular and $\| A^{-1}E \| < 1$, then $(A + E)$ is non-singular and we have
(i) $\| (A + E)^{-1} - A^{-1} \| \leqq \dfrac{\| A^{-1} \| \| F \|}{1 - \| F \|} = k \| A^{-1} \|$ (say),
and hence
(ii) $(1 - k)\| A^{-1} \| \leqq \| (A + E)^{-1} \| \leqq (1 + k)\| A^{-1} \|$
where $F = A^{-1}E$, so that $\| F \| < 1$.

The proof is as follows. We may write

$$(A + E) = A(I + A^{-1}E) = A(I + F). \tag{9.8}$$

Hence

$$(A + E)^{-1} = (I + F)^{-1}A^{-1}, \tag{9.9}$$

the existence of $(I + F)^{-1}$ following from (Q). This gives

$$(A + E)^{-1} - A^{-1} = [(I + F)^{-1} - I] A^{-1} \tag{9.10}$$

$$\|(A + E)^{-1} - A^{-1} \| \leqq \frac{\| A^{-1} \| \| F \|}{(1 - \| F \|)} . \tag{9.11}$$

This gives (i), and (ii) is an immediate consequence.

In practical applications, E will be a matrix of rounding errors and the result will be of value only if $(A + E)^{-1}$ can be regarded as a reasonable approximation to A^{-1}. Often we will not know anything special about the product $A^{-1}E$ and will have to majorize F, using the inequality

$$\| F \| \leqq \| A^{-1} \| \, \| E \|. \tag{9.12}$$

The necessity for doing this may well result in our estimate for

$$\| (A + E)^{-1} - A^{-1} \|$$

being far above its true value. When the substitution (9.12) is made, (i) becomes

$$\| (A + E)^{-1} - A^{-1} \| / \| A^{-1} \| \leqq \frac{\| A^{-1} \| \, \| E \|}{1 - \| A^{-1} \| \, \| E \|}. \tag{9.13}$$

We shall refer to the expression on the left as *the relative error in* $(A + E)^{-1}$. Note that the bound we have obtained for this relative error contains $\| A^{-1} \|$ itself as a factor, while the bound for the absolute error contains $\| A^{-1} \|^2$. Whether this represents the true state of affairs depends critically upon whether (9.12) is a reasonably sharp inequality.

The example in Table 2 illustrates the last point. We give there an ill-conditioned matrix A, and show the effect on the inverse of two different perturbations E_1 and E_2, in A. These perturbations are of the same order of magnitude, but E_1 results in a perturbation in the inverse which is of order $\| A^{-1} \|^2 \| E_1 \|$ and the second, in a perturbation of order $\| A^{-1} \| \, \| E_2 \|$. Perturbations E will behave in the same way as E_2 whenever E can be expressed in the form

$$E = AG \quad \text{or} \quad E = GA \tag{9.14}$$

where G is of the same order of magnitude as E. We shall see later that such perturbations are encountered in the error analysis. In applications of (S) we will normally restrict ourselves *ab initio* by some such condition as

$$\| A^{-1} \| \, \| E \| < 0.1$$

since we will not otherwise be able to prove that the computed inverse is a reasonable approximation to the true inverse.

TABLE 2

$$A = \begin{bmatrix} 1\ 00000\ 00000 & .90000\ 00000 \\ .77777\ 77800 & .70000\ 00000 \end{bmatrix}; \qquad A^{-1} = 10^{9} \begin{bmatrix} -.35000\ 00000 & +.45000\ 00000 \\ +.38888\ 88900 & -.50000\ 00000 \end{bmatrix}$$

$$E_1 = 10^{-9} \begin{bmatrix} 0 & 0 \\ 0 & 0.1 \end{bmatrix} \qquad (A + E_1)^{-1} - A^{-1} = 10^{8} \begin{bmatrix} -\ 1842\ \ 10526 & +.2368\ \ 42115 \\ +.2046\ \ 78363 & -.2631\ \ 57895 \end{bmatrix}$$

$$E_2 = 10^{-10}[A] \qquad (A + E_2)^{-1} - A^{-1} = 10^{-1} \begin{bmatrix} +.35000\ 0000 & -.45000\ 00000 \\ -.38888\ 88900 & +.50000\ 00000 \end{bmatrix}$$

With this condition, (i) and (ii) become

$$\| (A + E)^{-1} - A^{-1} \| / \| A^{-1} \| \leqq \tfrac{10}{9} \| A^{-1}E \| \tag{9.15}$$

$$[1 - \tfrac{10}{9} \| A^{-1}E \|] \| A^{-1} \| \leqq \| (A + E)^{-1} \| \leqq [1 + \tfrac{10}{9} \| A^{-1}E \|] \| A^{-1} \|. \tag{9.16}$$

(T) The final result will be used in some form in nearly every analysis. If A is non-singular and

$$\left.\begin{array}{l} AY - I = P + Q + R \\ \| P \| \leqq a \| Y \|, \quad \| Q \| \leqq b \| A^{-1} \|, \quad \| R \| \leqq c \end{array}\right\} \tag{9.17}$$

where a, b and c are scalars, then if $a \| A^{-1} \| < 1$

(i) $\| A^{-1} \| [1 - b \| A^{-1} \| - c] / [1 + a \| A^{-1} \|]$
$\leqq \| Y \| \leqq \| A^{-1} \| [1 + b \| A^{-1} \| + c] / [1 - a \| A^{-1} \|]$,

and if the right-hand side is denoted by $k \| A^{-1} \|$

(ii) $\| AY - I \| \leqq (ak + b) \| A^{-1} \| + c$,

(iii) $\| Y - A^{-1} \| / \| A^{-1} \| \leqq \| A^{-1} \| (ak + b) + c$.

These follow by writing (9.17) in the form

$$Y - A^{-1} = A^{-1}[P + Q + R] \tag{9.18}$$

and taking norms of both sides. Provided $a\| A^{-1} \|$ and $(b\| A^{-1} \| + c)$ are appreciably less than unity, (i) and (iii) show that Y is a good approximation to A^{-1}. In applications P is always present in (9.17) but Q or R or both may be absent.

Again we stress that the result will be pessimistic if $\| A^{-1} \| \| P \|$ is a poor approximation to $\| A^{-1}P \|$ and this will happen if P may be expressed in the form AS where S is of the same order of magnitude as P (a normalized A is, of course, implicit in this remark).

Solution of Triangular Sets of Equations

10. All the methods of inversion which we analyze require the solution of sets of equations with a triangular matrix of coefficients. The solution of these sets in fixed-point arithmetic will usually demand some form of scaling, so we consider floating-point solution first. Take an upper triangular set of equations

$$Ux = b; \tag{10.1}$$

typically, x_r is calculated from the computed $x_{r+1}, \cdots, x_n$ using the relation

$$x_r \equiv \mathrm{fl}\left(\frac{-u_{r,r+1}\,x_{r+1} - u_{r,r+2}\,x_{r+2} - \cdots - u_{r,n}\,x_n + b_r}{u_{rr}}\right) \tag{10.2}$$

in the notation of [8]. From the result given there for an inner product, we have

$$\left.\begin{array}{l} u_{rr}x_r(1 \pm \theta 2^{-t}) + u_{r,r+1}x_{r+1}(1 \pm \theta(n + 1 - r)2^{-t}) \\ \quad + u_{r,r+2}x_{r+2}(1 \pm \theta(n + 1 - r)2^{-t}) + \cdots \\ \quad + u_{r,n-1}x_{n-1}(1 \pm \theta 4 \cdot 2^{-t}) + u_{rn}x_n(1 \pm \theta 3 \cdot 2^{-t}) \\ \quad \equiv b_r(1 \pm \theta 2^{-t}) \end{array}\right. \tag{10.3}$$

where each θ denotes a value, ordinarily different in each factor, satisfying $| \theta | \leqq 1$.

Hence the computed solution satisfies

$$Ux + \delta Ux = b + \delta b \tag{10.4}$$

where, typically for $n = 5$,

$$|\delta b| \leqq 2^{-t} \begin{bmatrix} |b_1| \\ |b_2| \\ |b_3| \\ |b_4| \\ 0 \end{bmatrix} \leqq 2^{-t}|b| \tag{10.5}$$

$$|\delta U| < 2^{-t} \begin{bmatrix} |u_{11}| & 5|u_{12}| & 5|u_{13}| & 4|u_{14}| & 3|u_{15}| \\ & |u_{22}| & 4|u_{23}| & 4|u_{24}| & 3|u_{25}| \\ & & |u_{33}| & 3|u_{34}| & 3|u_{35}| \\ & & & |u_{44}| & 2|u_{45}| \\ & & & & |u_{55}| \end{bmatrix} = 2^{-t}V \quad \text{(say)}, \tag{10.6}$$

we may write this in the form

$$|Ux - b| < 2^{-t}[|b| + V|x|]. \tag{10.7}$$

For the solution of the n sets of equations

$$UX = B \tag{10.8}$$

we have, for the computed X

$$|UX - B| < 2^{-t}[|B| + V|X|] \tag{10.9}$$

and hence

$$\begin{aligned} \|UX - B\| \leqq \||UX - B|\| &\leqq 2^{-t}\||B|\| + 2^{-t}\|V\|\,\||X|\| \\ &\leqq 2^{-t}n^{\frac{1}{2}}\|B\| + 2^{-t}n^{\frac{1}{2}}\|V\|\,\|X\|. \end{aligned} \tag{10.10}$$

If we merely assume that U has normalization (I) then we can obtain an upper bound for V by replacing all u_{ij} by unity. We have then

$$\begin{aligned} \|V\| \leqq \|V\|_E < [n3^2 + (n-1)4^2 + \cdots \\ + 2(n+1)^2 + (n+2)^2]^{\frac{1}{2}} \end{aligned} \tag{10.11}$$

where we have replaced all multiples in the rth column of V by $(n + 3 - r)$.

We deduce that $\|V\|_E \sim \dfrac{n^2}{\sqrt{12}}$ and in fact

$$\|V\| < 0.4\,n^2 \qquad (n \geqq 10). \tag{10.12}$$

(Here and elsewhere the result holding when $n < 10$ is not catastrophically worse than that for $n \geqq 10$ but we are not much concerned with low order matrices.) Hence

$$\|UX - B\| \leqq 2^{-t}[n^{\frac{1}{2}}\|B\| + 0.4\,n^{5/2}\|X\|]. \tag{10.13}$$

When $B = I$ we find that the computed X is exactly triangular and no term corresponding to $\| B \|$ is now present in (10.13). This is because the rth column of the inverse is the solution of $U_x = e_r$ and the error corresponding to δb in (10.5) is therefore the null vector. We do not require the whole of the majorizing matrix, V, except for the solution corresponding to e_n but we shall not attempt to take this into account. We have, therefore, for the computed solution of $UX = I$,

$$\| UX - I \| \leqq (0.4)2^{-t}n^{5/2}\| X \| \tag{10.14}$$

and from (T) of Section 9,

$$\begin{aligned} &\| U^{-1} \|/(1 + 0.4\cdot 2^{-t}n^{5/2}\| U^{-1} \|) \\ &\qquad \leqq \| X \| \leqq \| U^{-1} \|/(1 - 0.4\cdot 2^{-t}n^{5/2}\| U^{-1} \|) \end{aligned} \tag{10.15}$$

$$\| X - U^{-1} \| \leqq 0.4\cdot 2^{-t}n^{5/2}\| U^{-1} \|^2/(1 - 0.4\cdot 2^{-t}n^{5/2}\| U^{-1} \|). \tag{10.16}$$

We can *guarantee* that X is a good inverse only if $(0.4)2^{-t}n^{5/2}\| U^{-1} \|$ is appreciably less than unity.

If on the other hand we know (as we often shall) that

$$\sum_i u_{ij}^2 \leqq 1 \text{ (all } j) \quad \text{or} \quad \sum_j u_{ij}^2 \leqq 1 \text{ (all } i),$$

then we have

$$\begin{aligned} &\| V \| \leqq \| V \|_E \\ &\quad \leqq [1 + 3^2 + 4^2 + \cdots + (n + 1)^2]^{\frac{1}{2}} \leqq 0.6\, n^{3/2} \quad (n \geqq 10). \end{aligned} \tag{10.17}$$

The result for the inverse is

$$\| X - U^{-1} \| \leqq (0.6)2^{-t}n^2\| U^{-1} \|^2/(1 - (0.6)2^{-t}n^2\| U^{-1} \|). \tag{10.18}$$

11. The corresponding results for fixed-point arithmetic are rather less definite because they depend on the details of the arithmetic facilities and on whether we have *a priori* information on the size of the solution. For the sake of definiteness we assume that scalar products can be accumulated exactly in a double-precision accumulator and that division into this double accumulator is possible. These facilities are provided on all desk machines and many high speed computers.

If we know in advance that the computed solution lies between 2^{k-1} and 2^k, then we may write

$$x = 2^k y \tag{11.1}$$

and solve

$$Uy = 2^{-k}b. \tag{11.2}$$

The variable y_r is determined from the computed $y_{r+1}, \cdots, y_n$ by the relation

$$y_r \equiv \mathrm{fi}\left[\frac{2^{-k}b_r - u_{r,r+1}\, y_{r+1} - \cdots - u_{rn}y_n}{u_{rr}}\right] \tag{11.3}$$

where fi means that we take the *fixed*-point result of evaluating the expression in brackets. The numerator is accumulated exactly and the only error is the rounding error in the division. (We have tacitly assumed that intermediate values of the scalar product do not require the introduction of a larger scale factor than is necessary for the final value. On ACE this difficulty is avoided quite simply by accumulating exactly 2^{-p} times the scalar product, where 2^p is greater than the largest order matrix that can be dealt with. Overspill in the scalar product cannot therefore occur and the rounding errors introduced, 2^{p-2t}, are far below the level of the others. We shall not refer to this point again.) Hence

$$Uy \equiv 2^{-k}b + \delta b \tag{11.4}$$

$$|\,\delta b_r\,| \leqq |\,u_{rr}\,|2^{-t-1} < 2^{-t-1} \tag{11.5}$$

giving

$$Ux = b + 2^k\delta b. \tag{11.6}$$

Applied to the solution of $UX = B$, we obtain an X satisfying

$$UX \equiv B + C \tag{11.7}$$

where

$$|\,C_{ij}\,| < 2^k2^{-t-1} \leqq \max|\,X_{ij}\,|2^{-t} \leqq \|\,X\,\|2^{-t} \tag{11.8}$$

$$\|\,C\,\| \leqq n2^{-t}\|\,X\,\|. \tag{11.9}$$

When $B = I$, we have

$$UX = I + C \tag{11.10}$$

where (11.9) is satisfied, giving

$$\|\,X - U^{-1}\,\| \leqq n2^{-t}\|\,U^{-1}\,\|^2/(1 - n2^{-t}\|\,U^{-1}\,\|). \tag{11.11}$$

In this we have tacitly assumed that $k \geqq 0$. If $k < 0$ no scale factor is necessary and (11.9) becomes $\|\,C\,\| < n2^{-t-1}$. For the inversion problem k must necessarily satisfy $k \geqq 0$, because one element of X is $(1/u_{nn})$ and this is not less than unity. The strictest application of (11.5) gives

$$UX = I + C; \quad |\,C\,| \leqq 2^{-t-1}2^k \begin{bmatrix} |\,u_{11}\,| & |\,u_{11}\,| & \cdots & |\,u_{11}\,| \\ & |\,u_{22}\,| & \cdots & |\,u_{22}\,| \\ \cdots & \cdots & \cdots & \cdots \\ & & & |\,u_{nn}\,| \end{bmatrix} \tag{11.12}$$

12. When the scale factor is not known in advance, then the elements of X are scaled when necessary as the computation proceeds. If the first computed element of X requires the largest scale factor that is necessary for the whole computation then the analysis is precisely that we have just given. If complete pivoting for size has been used this will commonly be the case. Otherwise the situation will be this. Suppose that at the stage when x_r is about to be computed the current scale factor is $2^{k_{r+1}}$. We may denote the current scaled values of the variables x_i by $x_i^{(r+1)}$ $(i = r + 1, \cdots, n)$ so that

$$\left.\begin{aligned} &x_i^{(r+1)} = 2^{k_{r+1}} y_i^{(r+1)}; \quad | y_i^{(r+1)} | < 1 \qquad (\mathrm{i} = r + 1, \cdots, n) \\ &| y_i^{(r+1)} | \geqq \tfrac{1}{2} \text{ for at least one } i. \end{aligned}\right\} \tag{12.1}$$

The next step is the computation of $x_r^{(r)}$ with the introduction of an increased scale factor 2^{k_r} if necessary, and we have

$$u_{rr}x_r^{(r)} + u_{r,r+1}x_{r+1}^{(r+1)} + \cdots + u_{rn}x_n^{(r+1)} \equiv b_r + \epsilon_r, \qquad | \epsilon_r | \leqq 2^{k_r - t+1}. \tag{12.2}$$

Now it may be necessary to introduce higher scale factors at any or all of the subsequent stages so that the final values $x_r^{(1)} \cdots x_n^{(1)}$ of these variables are obtained by rounding off x_r, x_{r+1}, x_{r+2}, $\cdots$, x_n to progressively fewer figures. The worst that can have happened is that $(k_1 - k_{r+1})$ successive roundings have taken place. If then we write

$$x_i^{(r)} - x_i^{(1)} = \epsilon_{ir}, \tag{12.3}$$

then

$$| \epsilon_{ir} | < 2^{-t}[2^{k_1} - 2^{k_r}] < 2^{k_1 - t}. \tag{12.4}$$

The final values therefore satisfy

$$\begin{aligned} u_{rr}x_r^{(1)} + u_{r,r+1}x_{r+1}^{(1)} + \cdots + u_{rn}x_n^{(1)} &\equiv b_r + \epsilon_r \\ &- u_{rr}\epsilon_{rr} - u_{r,r+1}\epsilon_{r+1,r+1} - u_{r,r+2}\epsilon_{r+2,r+1} - \cdots - u_{rn}\epsilon_{n,r+1}. \end{aligned} \tag{12.5}$$

The error term on the right is bounded by $(n + 1 - r)2^{k_1 - t}$. This is worse by the factors $(n + 1 - r)$ than was obtained when the final scale factor 2^{k_1} was used throughout. In spite of this, it is our opinion that the progressive introduction of the scale factor usually gives the more accurate result, particularly for very ill-conditioned matrices. We may illustrate this by considering the inversion of the matrix A, below, working to five figures.

$$A = \begin{bmatrix} +.00311 & -.23456 & - 34567 \\ & +.00112 & -.12345 \\ & & +.00113 \end{bmatrix}$$

When calculating the last column of the inverse with progressive scale factors we obtain 884.96 for the last element in the first instance. This is subsequently rounded first to 885 and then to 9×10^2, the final column being $10^2[74552, 975, 9]$. If we introduce the correct scale factor from the start, we have for the computed final column $10^2[75818, 992, 9]$. It is obvious that the former answer is the better, though it has residuals which are about 100 times as large as those of the latter.

Matrix Multiplication

13. For matrix multiplication in floating-point we write

$$C \equiv \mathrm{fl}(AB). \tag{13.1}$$

From our result for floating-point scalar products we have

$$C_{ij} \equiv a_{i1}(1 + \epsilon_1)b_{1j} + a_{i2}(1 + \epsilon_2)b_{2j} + \cdots + a_{in}(1 + \epsilon_n)b_{nj}$$

with the usual bounds for the ϵ_i. We have therefore, typically for a matrix of order 4:

$$| C - AB | \leqq 2^{-t} \begin{bmatrix} 4|a_{11}| & 4|a_{12}| & 3|a_{13}| & 2|a_{14}| \\ 4|a_{21}| & 4|a_{22}| & 3|a_{23}| & 2|a_{24}| \\ 4|a_{31}| & 4|a_{32}| & 3|a_{33}| & 2|a_{34}| \\ 4|a_{41}| & 4|a_{42}| & 3|a_{43}| & 2|a_{44}| \end{bmatrix} \cdot | B |. \qquad (13.2)$$

Hence *a fortiori*

$$| C - AB | < n2^{-t} | A | | B | \qquad (13.3)$$

and

$$\begin{aligned} \| C - AB \| &\leqq \||| C - AB ||| \leqq n2^{-t} \||| A | | B ||| \\ &\leqq n2^{-t} ||| A ||| \, ||| B ||| \leqq n^2 2^{-t} \| A \| \, \| B \|. \end{aligned} \qquad (13.4)$$

In the only application made in this paper we have $B = A^T$ and for this

$$\| C - AA^T \| \leqq n^2 2^{-t} \| A \| \, \| A^T \| = n^2 2^{-t} \| AA^T \|. \qquad (13.5)$$

Writing this in the form

$$\frac{\| C - AA^T \|}{\| AA^T \|} \leqq n^2 2^{-t}, \qquad (13.6)$$

we see that the bound for the relative error in the computed product is independent of that product. Note that even the computed C is exactly symmetric so that we need form only the upper half of C.

For fixed-point computation of AA^T, even if the scale factors are introduced progressively, we have

$$| C - AA^T |_{ij} < 2^{k_1} 2^{-t} \qquad (13.7)$$

where 2^{k_1} is the final scale factor. If the diagonal elements are computed first, the final scale factor is determined at an early stage since the maximum element is on the diagonal. Equation (13.7) gives

$$\| C - AA^T \| \leqq 2^{k_1} n 2^{-t} \qquad (13.8)$$

and hence

$$\frac{\| C - AA^T \|}{\| AA^T \|} \leqq 2n2^{-t} \qquad (13.9)$$

since $\| AA^T \| \geqq \max |(AA^T)_{ij}| \geqq \frac{1}{2} \cdot 2^{k_1}$. This is a most satisfactory error bound.

General Error Analysis of Gaussian Elimination

14. Consider now the inversion of a general matrix, A, by reduction to triangular form by Gaussian elimination using a complete pivotal strategy. This strategy merely determines a re-ordering of the rows and columns of A and it will simplify the notation, without any loss of generality, if we consider A to be reordered in this way, so that at each stage the pivot is in the top left-hand corner. It should be emphasized that this re-ordering is not known in advance.

We first give an general analysis which is independent of the type of arithmetic that is used. We denote the computed elements of the rth matrix $A^{(r)}$, by $a_{ij}^{(r)}$ and the computed multipliers by m_{ij}. The history of the elements in the (i, j) position is as follows.

(i) $i \leqq j$.

$$\begin{aligned} a_{ij}^{(2)} &\equiv a_{ij}^{(1)} - m_{i1}a_{1j}^{(1)} + \epsilon_{ij}^{(2)} \\ a_{ij}^{(3)} &\equiv a_{ij}^{(2)} - m_{i2}a_{2j}^{(2)} + \epsilon_{ij}^{(3)} \\ &\cdots\cdots\cdots\cdots\cdots\cdots \\ a_{ij}^{(i)} &\equiv a_{ij}^{(i-1)} - m_{i,i-1}a_{i-1,j}^{(i-1)} + \epsilon_{ij}^{(i)} \end{aligned} \tag{14.1}$$

where $\epsilon_{ij}^{(k)}$ is the error made in computing $a_{ij}^{(k-1)} - m_{i,k-1}a_{k-1,j}^{(k-1)}$ from the computed $a_{ij}^{(k-1)}$, $m_{i,k-1}$ and $a_{k-1,j}^{(k-1)}$. The element $a_{ij}^{(i)}$ is an element of the ith pivotal row and undergoes no further change. Summing equations (14.1) and cancelling common elements,

$$a_{ij}^{(i)} + \sum_{k=1}^{i-1} m_{ik}a_{kj}^{(k)} \equiv a_{ij}^{(1)} + \sum_{k=2}^{k} \epsilon_{ij}^{(k)}. \tag{14.2}$$

(ii) $i > j$. We now have the same equations as (14.1) terminating with

$$a_{ij}^{(j)} \equiv a_{ij}^{(i-1)} - m_{i,j-1}a_{j-1,j}^{(j-1)} + \epsilon_{ij}^{(j)}. \tag{14.3}$$

The next value, $a_{ij}^{(j+1)}$, is taken to be exactly zero and remains zero for the rest of the reduction. The element $a_{ij}^{(j)}$ is used to calculate m_{ij} from

$$m_{ij} \equiv \frac{a_{ij}^{(j)}}{a_{jj}^{(j)}} + \eta_{ij} \tag{14.4}$$

where η_{ij} is the rounding error in the division. Hence

$$0 \equiv a_{ij}^{(j)} - m_{ij}a_{jj}^{(j)} + \epsilon_{ij}^{(j+1)} \tag{14.5}$$

where

$$\epsilon_{ij}^{(j+1)} = a_{jj}^{(j)}n_{ij}. \tag{14.6}$$

Summing equations of type (14.1) up to (14.3) and adding in (14.5),

$$\sum_{k=1}^{j} m_{ik}a_{kj}^{(k)} \equiv a_{ij}^{(1)} + \sum_{k=2}^{j+1} \epsilon_{ik}^{(k)}. \tag{14.7}$$

Note that equations (14.2) and (14.7) give a relation between elements $a_{rs}^{(r)}$ of the final triangular matrix, elements m_{rs}, and elements $a_{rs}^{(1)}$ with added perturbations. Writing L for the lower triangular matrix formed by the m_{ij} augmented by a unit diagonal, and U for the upper triangle formed by the pivotal rows, (14.2) and (14.7) give

$$LU \equiv A^{(1)} + E^{(2)} + E^{(3)} + \cdots + E^{(n)} \equiv A^{(1)} + E \qquad (14.8)$$

where $E^{(k)}$ is the matrix formed by the $\epsilon_{ij}^{(k)}$. Note that this has null rows 1 to $(k - 1)$ and null columns 1 to $(k - 2)$.

We may describe this result as follows. "The computed matrices L and U are the matrices which would have been obtained by correct computation with $A^{(1)} + E$. Further $A^{(k)}$ is the matrix which would have resulted from exact computation with

$$A^{(1)} + E^{(2)} + E^{(3)} + \cdots + E^{(k)} = A^{(1)} + F^{(k)} \quad \text{(say).''} \qquad (14.9)$$

Gaussian Elimination in Floating Point

15. We now assume that the matrix is scaled so that all elements are slightly less than $1/g(n)$ [see Sec. 4]. Strictly speaking for floating-point computation, this scaling is irrelevant, but it facilitates comparison with fixed-point computation and error analysis elsewhere. We also assume that the matrix is equilibrated although no use is made of this fact. We consider this to be necessary to justify the complete pivotal strategy. We know that for exact computation the elements of $A^{(k)}$ would be bounded by $g(k)/g(n)$ and hence every element of every $A^{(k)}$ would be less than unity. We shall assume this is also true for the computed $A^{(k)}$.

It might be felt that the result of the last paragraph guarantees this, since the final L and U *do* correspond to exact computation with $A + F^{(n)}$. Unfortunately each $A^{(k)}$ corresponds to exact computation with the corresponding $A^{(1)} + F^{(k)}$, so that although we can claim to have performed an exact Gaussian elimination with $A^{(1)} + F^{(n)}$ we cannot claim that we have selected the largest pivot at each stage corresponding to this fixed matrix. We hope to deal with this point later, but, for the moment, it remains an assumption. We assume explicitly that if

$$|a_{ij}^{(1)}| < \frac{1}{g(n)} - 2^{-t}\left[\frac{g(1) + 2g(2) + 2g(3) + \cdots + g(n)}{g(n)}\right], \qquad (15.1)$$

then

$$|a_{ij}^{(k)}| < \frac{g(k)}{g(n)} \qquad (k = 2, \cdots, n). \qquad (15.2)$$

Note that $|m_{ij}| \leqq 1$ is assured since we *do* take as pivot the largest element of computed $A^{(k)}$ at each stage.

Now we have

$$\begin{aligned} a_{ij}^{(k)} &= \mathrm{fl}(a_{ij}^{(k-1)} - m_{i,k-1}a_{k-1,j}^{(k-1)}) \\ &\equiv [a_{ij}^{(k-1)} - m_{i,k-1}a_{k-1,j}^{(k-1)}(1 + \epsilon_1)](1 + \epsilon_2); \quad |\epsilon_1|, |\epsilon_2| \leqq 2^{-t}. \end{aligned} \qquad (15.3)$$

Hence

$$\begin{aligned} |\,\epsilon_{ij}^{(k)}\,| &\leqq \frac{|\,\epsilon_2\,|\,|\,a_{ij}^{(k)}\,|}{|\,1 + \epsilon_2\,|} + |\,m_{i,k-1}\,|\,|\,a_{k-1,j}^{(k-1)}\,|\,|\,\epsilon_1\,| \\ &\leqq 2^{-t}\left[\frac{g(k) + g(k-1)}{g(n)}\right], \end{aligned} \tag{15.4}$$

showing that, under our assumption (15.1), all elements of $(A + E)$ are less than $1/g(n)$.

Since we are assuming that all $a_{ij}^{(k)}$ are less than unity, it is clear that

$$\mathrm{fl}\,[a_{ij}^{(k-1)} - m_{i,k-1}a_{k-1,j}^{(k-1)}] - [a_{ij}^{k-1} - m_{i,k-1}a_{k-1,j}^{k-1}] \leqq 2^{-t}. \tag{15.5}$$

Hence a bound for the matrix $E^{(k)}$ is, typically for $n = 5$, $k = 3$,

$$2^{-t}\begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 1 & 1 \\ 0 & 1 & 1 & 1 & 1 \\ 0 & 1 & 1 & 1 & 1 \end{bmatrix} \tag{15.6}$$

and for E is

$$2^{-t}\begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 1 & 1 \\ 1 & 2 & 2 & 2 & 2 \\ 1 & 2 & 3 & 3 & 3 \\ 1 & 2 & 3 & 4 & 4 \end{bmatrix}. \tag{15.7}$$

Using the Euclidean norm as an approximation we have

$$\|\,E\,\| \leqq (0.41)2^{-t}n^2. \tag{15.8}$$

These results all represent extreme upper bounds and cannot be attained even if the maximum possible growth takes place during the reduction. If, as is usually the case, the elements diminish somewhat in size as the process progresses then the rounding errors, which are certainly less than $(|\,a_{ij}^{(k-1)}| + |\,a_{k-1,j}^{(k-1)}\,|)2^{-t}$ at each stage, becomes progressively smaller. If, for example, the elements diminish by a factor of 2 at each stage, then the last element of the bound for E becomes $2 \times 2^{-t}/g(n)$ instead of $(n - 1) \times 2^{-t}$.

There remains the possibility that the process may break down prematurely due to the emergence of a null $A^{(k)}$. This can happen only if $A + F^{(k)}$ is exactly singular. Now $\|\,F^{(k)}\| \leqq (0.41)2^{-t}n^2$, since each element of $F^{(k)}$ is certainly bounded by the bound we have for E. Breakdown cannot occur therefore if $(0.41)2^{-t}n^2\,\|\,A^{-1}\,\| < 1$ and we shall in any case need a stronger condition if we are to guarantee a good inverse.

16. The inversion is completed by solving

$$LUX = I \tag{16.1}$$

which is performed in the two steps

$$LY = I, \qquad UX = Y. \tag{16.2}$$

Applying our results for the floating-point solution of triangular sets of equations we have finally:

$$LU \equiv A + E \qquad \text{(a)}$$

$$\text{where from (15.8):} \quad \| E \| \leqq (0.41)2^{-t}n^2 \quad \text{(b)}$$

$$LY \equiv I + F \qquad \text{(c)}$$

$$\text{where from (10.14):} \quad \| F \| \leqq (0.4)2^{-t}n^{5/2} \| Y \| \quad \text{(d)}$$

$$UX \equiv Y + G + H \qquad \text{(e)}$$

$$\text{where from (10.13):} \quad \| G \| \leqq 2^{-t}n^{\frac{1}{2}} \| Y \| ;$$

$$\| H \| \leqq (0.4)2^{-t}n^{5/2} \| X \| . \quad \text{(f)}$$

Hence

$$\begin{aligned} LUX &\equiv LY + LG + LH \\ &\equiv I + F + LG + LH \end{aligned}$$

or

$$\begin{aligned} (A + E)X &\equiv I + F + LG + LH \\ (AX - I) &\equiv -EX + F + LG + LH \end{aligned} \tag{16.3}$$

We need satisfactory bounds for the norms of each of the terms on the right. It will become apparent that we cannot guarantee a reasonable inverse unless $2^{-t}n^{7/2} \| A^{-1} \|$ is appreciably less than unity. We therefore assume

$$2^{-t}n^{7/2} \| A^{-1} \| < 0.1, \tag{16.4}$$

and this condition certainly guarantees that there is no breakdown during the reduction. We assume further $n \geqq 10$. Relation (16.4) implies that A is nonsingular. Equations (a) and (b) give

$$\| (A + E)^{-1} \| \leqq \| A^{-1} \| \left[1 + \frac{\| A^{-1} \| \, \| E \|}{1 - \| A^{-1} \| \, \| E \|} \right] \leqq 1.01 \| A^{-1} \|. \tag{16.5}$$

Now from (a),

$$\begin{aligned} \| L^{-1} \| &= \| U(A + E)^{-1} \| \\ &\leqq \| U \| \, \| (A + E)^{-1} \| \\ &\leqq \left[\frac{n(n + 1)}{2} \right]^{\frac{1}{2}} \cdot 1.01 \| A^{-1} \| \\ &\leqq (0.8)n \| A^{-1} \|, \end{aligned} \tag{16.6}$$

since all elements of U are less than unity.

From (c) and (d),

$$\begin{aligned} \| Y \| &\leqq \| L^{-1} \| / [1 - (0.4)2^{-t}n^{5/2} \| L^{-1} \|] \\ &\leqq (0.84)n \| A^{-1} \| . \end{aligned} \tag{16.7}$$

Applying these results and remembering that $\| L \| \leqq 0.8n$, since all its elements are bounded by unity, we have

$$\begin{aligned} \| EX \| &\leqq (0.4)2^{-t}n^2 \| X \| \\ \| F \| &\leqq (0.34)2^{-t}n^{7/2} \| A^{-1} \| \\ \| LG \| &\leqq (0.7)2^{-t}n^{5/2} \| A^{-1} \| \\ \| LH \| &\leqq (0.32)2^{-t}n^{7/2} \| X \| . \end{aligned} \tag{16.8}$$

We have now established a result of the type (T) of Section 9 with

$$\begin{aligned} a &= 2^{-t}[0.4n^2 + 0.32n^{7/2}] < 2^{-t}[0.34n^{7/2}] \\ b &= 2^{-t}[0.7n^{5/2} + 0.34n^{7/2}] \leqq 2^{-t}[0.41n^{7/2}]. \end{aligned} \tag{16.9}$$

Hence

$$\begin{aligned} &\| A^{-1} \| \frac{1 - 0.41n^{7/2}2^{-t} \| A^{-1} \|}{1 + 0.34n^{7/2}2^{-t} \| A^{-1} \|} \\ &\qquad \leqq \| X \| \leqq \| A^{-1} \| \frac{1 + 0.41n^{7/2}2^{-t} \| A^{-1} \|}{1 - 0.34n^{7/2}2^{-t} \| A^{-1} \|} \leqq (1.076) \| A^{-1} \| \\ &\| X - A^{-1} \| / \| A^{-1} \| \leqq (0.78)2^{-t}n^{7/2} \| A^{-1} \| \\ &\| AX - I \| \leqq (0.78)2^{-t}n^{7/2} \| A^{-1} \| . \end{aligned} \tag{16.10}$$

We can readily deduce the result for a matrix with elements bounded by unity. For if B is such a matrix then we work with $A = B/g(n)$. Normally we would divide by the smallest power of 2 greater than $g(n)$ so that no rounding was involved. If X is the computed inverse of A, then take $Z = X/g(n)$ as the inverse of B. We have then

$$BZ = g(n)AX/g(n) = AX.$$

Hence

$$\begin{aligned} \| BZ - I \| = \| AX - I \| &\leqq (0.78)2^{-t}n^{7/2} \| A^{-1} \| \\ &\leqq (0.78)2^{-t}n^{7/2}g(n) \| B^{-1} \| . \end{aligned} \tag{16.11}$$

The presence of the term $n^{7/2}$ is at first sight disappointing. However, we must remember that Goldstine and von Neumann's results for a positive definite symmetric matrix contains the factor $n^2 \| A \| \| A^{-1} \|$ and $\| A \|$ could be as large as n. In corresponding positions in our analysis we have replaced $\| L \|$ and $\| U \|$ by $0.8n$ and this contributes a factor n in our result. We have therefore only the additional factor $n^{\frac{1}{2}}$ and this result is for a general matrix.

The presence of the factor $g(n)$ is inevitable, but notice that it should be replaced by max $[1, R]$ where R is the growth ratio, in each particular case and we know in practice that R is often less than unity. For each of our special types of matrix we have an a priori upper bound for R which is much smaller than $g(n)$.

Fixed-Point Computation

17. We do not wish to repeat the analysis for all varieties of fixed-point computation and we restrict ourselves to one particular type. Before describing this we justify its selection.

The alternative methods of solving unsymmetric equations for which an error analysis has been performed are

(i) by reducing the problem to that of inverting AA^T, a positive definite matrix [3],

(ii) by orthogonal triangularizations of different types (§21–23).

None of these methods requires less than $2n^3$ multiplications, whereas the method we have just described requires n^3. If we are prepared to use $2n^3$ multiplications, then the following strategy gives high accuracy.

(I) Invert the matrix as above, thereby determining the interchanges and the required scaling factors.

(II) Repeat the computation in fixed point taking advantage of the information gained in (I) and accumulating scalar products wherever possible.

In (II) we determine L and U directly from $LU = A$, where A has its rows and columns permuted as determined by (I). We should compute $\frac{1}{2}L$ in case some of the multipliers were just greater than unity as a result of the different rounding errors. For this technique we have

$$LU \equiv A + E; \quad |e_{ij}| < 2^{-t}; \quad \|E\| \leqq 2^{-t}n \qquad (17.1)$$

since each element of $\frac{1}{2}L$ and U is determined directly by dividing into an exactly accumulated scalar product. Similarly in solving $LY = I$ and $UX = Y$ the scale factors are known in advance. Some of the bounds we give will be different according as $\max_{i,j} |x_{ij}|$ is greater or less than $\max_{i,j} |y_{ij}|$. Where this is true, the bounds in the former case are given on the left and for the latter on the right. We have

$$\left.\begin{array}{l} LY = I + F;\quad |f_{rs}| \leqq 2^{-t}\max\limits_{i,j}|y_{ij}|;\quad \|F\| \leqq n2^{-t}\|Y\| \\ \hfill \text{from (11.8)} \\ \qquad\qquad UX = Y + G \\ |g_{rs}| \leqq 2^{-t}\max\limits_{i,j}|x_{ij}| \quad \vdots \quad |g_{rs}| \leqq 2^{-t}\max\limits_{i,j}(y_{ij}) \\ \|G\| \leqq n2^{-t}\|X\| \quad \vdots \quad \|G\| \leqq n2^{-t}\|Y\| \end{array}\right\} \qquad (17.2)$$

giving

$$\left.\begin{array}{c} \| EX \| \leqq 2^{-t}n \| X \| \\ \| F \| \leqq (0.9)2^{-t}n^2 \| A^{-1} \| \\ \| LG \| \leqq (0.9)2^{-t}n^2 \| X \| \quad \vdots \quad \| LG \| \leqq (0.7)2^{-t}n^3 \| A^{-1} \| \end{array}\right\} \tag{17.3}$$

provided $2^{-t}n^2 \| A^{-1} \| \leqq 0.1$. (The constants have been taken to one decimal only.) In both cases we have results of the type (T) of Section 9 with

$$\begin{array}{ll} a = 2^{-t}[n + 0.8n^2] & \vdots \quad a = 2^{-t}n \\ b = 2^{-t}[0.9n^2] & \vdots \quad b = 2^{-t}[0.9n^2 + 0.7n^3] \end{array} \tag{17.4}$$

In practice the second of these is very uncommon since the largest element of X is usually larger than the largest element in Y, particularly if A is at all ill-conditioned. For several of the special matrices we have already seen that $\max_{i,j} | y_{ij} | = \max_{i,j} | l_{ij}^{-1} |$ is equal to unity. In any case $\max_{i,j} | y_{ij} | \leqq 2^{n-1}$, so that for small matrices $\| Y \|$ *cannot* be large.

Hence we will usually have

$$\| AX - I \| \leqq (2.0)2^{-t}n^2 \| A^{-1} \| \qquad \text{if} \quad n^2 2^{-t} \| A^{-1} \| \leqq 0.1, \quad n \geqq 10. \tag{17.5}$$

In the alternative case we will not be able to guarantee that X is even an approximate inverse unless $n^3 2^{-t} \| A^{-1} \| \leqq 0.1$, and then we have

$$\| AX - I \| \leqq (0.9)2^{-t}n^3 \| A^{-1} \| . \tag{17.6}$$

The results are for a matrix A which has been scaled sufficiently to cover the growth factor (if any) observed in (I). Hence for a matrix A with elements satisfying $| a_{ij} | \leqq 1$, we have for the double process

$$\| AX - I \| \leqq (2.0)2^{-t}n^2 R \| A^{-1} \| \; \vdots \; \| AX - I \| \leqq (0.9)2^{-t}n^3 R \| A^{-1} \| . \tag{17.7}$$

Error Analysis of the Symmetric Cholesky Decomposition

18. We omit the analysis of Gaussian elimination of positive-definite matrices which can now be done in floating-point arithmetic with no pivotal strategy. The analysis is similar to that of the last section except that we know there is no growth in size of pivotal elements, and the norms of L and U are more specially related to those of A. We pass instead to the LL^T decomposition which is important in other contexts.

We consider fixed-point computation of L for a positive definite matrix A. We show that if

$$(n + 1)2^{-t} \| A^{-1} \| < 1; \quad | a_{ij} | < (1 - 2^{-t}), \tag{18.1}$$

then all computed elements of L satisfy $| l_{ij} | < 1$ and

$$LL^T \equiv A + E; \quad | e_{ij} | \leqq \begin{cases} 2^{-t-1} | l_{ii} |, & j > i \\ 2^{-t-1} | l_{jj} |, & i > j \\ 2^{-t} | l_{ii} |, & i = j. \end{cases} \tag{18.2}$$

The proof is by a double induction. Assume that when we have computed the first s rows of L and the first r elements of its $(s + 1)$th row, the computed elements are all bounded by unity and are those that would have resulted from the exact decomposition of the matrix $(A + E_{sr})$. Here E_{sr} is a symmetric matrix having zero elements except in its leading principal matrix of order s and in the first r elements of its $(s + 1)$th row and column where it has elements which are bounded as in (18.2). For $n = 5$, $s = 3$, $r = 1$, we have typically:

$$| E_{3,1} | \leqq 2^{-t-1} \begin{bmatrix} 2| l_{11} | & | l_{11} | & | l_{11} | & | l_{11} | & 0 \\ | l_{11} | & 2| l_{22} | & | l_{22} | & 0 & 0 \\ | l_{11} | & | l_{22} | & 2| l_{33} | & 0 & 0 \\ | l_{11} | & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix} \tag{18.3}$$

From (18.1) and our assumptions, $A + E_{rs}$ is a positive definite matrix with elements bounded by unity.

Consider now the determination of $l_{s+1,r+1}$. The equation from which it is computed is

$$l_{s+1,r+1} = \left[\frac{a_{s+1,r+1} - l_{s+1,1}\, l_{r+1,1} - l_{s+1,2}\, l_{r+1,2} - \cdots - 1_{s+1,r}\, l_{r+1,r}}{l_{r+1,r+1}} \right] \tag{18.4}$$
$$= q \quad \text{(say)}.$$

If this quotient were computed exactly it would be the $(s + 1, r + 1)$-element corresponding to the exact decomposition of $(A + E_{rs})$, and since this is a positive definite matrix with its elements bounded by unity, the exact quotient is bounded by unity. Hence the computed $l_{s+1,r+1}$ is also bounded by unity and

$$l_{s+1,r+1} \equiv q + \epsilon; \quad | \epsilon | \leqq 2^{-t-1}. \tag{18.5}$$

Hence

$$l_{s+1,1}l_{r+1,1} + l_{s+1,2}l_{r+1,2} + \cdots + l_{s+1,r+1}l_{r+1,r+1} \equiv a_{s+1,r+1} + l_{r+1,r+1}\epsilon \tag{18.6}$$

This completes our induction apart from the element $l_{s+1,s+1}$. This is computed from the relation

$$l_{s+1,s+1} = \sqrt{a_{s+1,s+1} - l^2_{s+1,1} - l^2_{s+1,2} - \cdots - l^2_{s+1,s}} = \sqrt{q} \quad \text{(say)}. \tag{18.7}$$

Again if the square root were computed exactly it would be less than unity, since it is an element in the exact decomposition of $A + E_{s+1,s}$ which is positive definite with elements bounded by unity.

We do not wish to enter into a detailed study of square root subroutines.

On *ACE*, on which the square root is calculated by the Newton process using division into the double-precision accumulator, special care is taken with round off, and the computed value is the correctly rounded value

$$l_{s+1,s+1} \equiv \sqrt{q} + \epsilon; \quad |\,\epsilon\,| \leqq 2^{-t-1}. \tag{18.8}$$

We shall assume this in our analysis, though if we had

$$|\,l_{s+1,s+1} - \sqrt{q}\,| < k \times 2^{-t-1}$$

for any reasonable k, the conditions on A would not need to be much changed from (18.1). Since the computed value is the correctly rounded value,

$$|\,l_{s+1,s+1}\,| \leqq 1.$$

The result is clearly true for a 1 × 1 matrix so the induction is complete. Note that if any l_{ii} is small (and if A is ill-conditioned this is almost certain to be true) the corresponding elements of the error matrix $(LL^T - A)$ are far smaller than 2^{-t-1}. In other words *the errors arising from the decomposition may be far less than those from the initial rounding of the elements of A, in the case when they are not exactly representable by numbers of t binary digits.* The LL^T decomposition is fundamental in the L-R technique of Rutishauser [5] as applied to symmetric matrices, and in the calculation of the zeros of det $(B - \lambda A)$ and det $(AB - \lambda I)$ when A is positive definite. The closeness of LL^T to A is therefore of great practical importance.

Completion of the Inverse

19. The inversion can now be completed by solving

$$L^T X = I \tag{19.1}$$

and computing XX^T. We can no longer avoid scaling in this part of the work and we give the analysis for floating-point computation since it will be seen that the errors made at this stage are less important. Using the result for the inversion of a triangular matrix and taking advantage of the fact that the sum of the squares of the elements in any row of L is not greater than unity (Section 7) we have from (10.18):

$$\begin{aligned} L^T X = I + F; \quad &\|\,F\,\| \leqq (0.6)2^{-t}n^2\,\|\,X\,\|\,; \\ &\|\,X\,\| \leqq \|(L^T)^{-1}\|/[1 - (0.6)2^{-t}n^2\,\|(L^T)^{-1}\|]. \end{aligned} \tag{19.2}$$

The computed inverse Y satisfies

$$Y \equiv XX^T + G; \quad \|\,G\,\| \leqq n^2 2^{-t}\,\|\,X\,\|^2 \tag{19.3}$$

from (13.5). We need compute only the upper half of Y, since even the computed Y is exactly symmetric. The total number of multiplications involved in computing Y from A is $\frac{1}{2}n^3$, so that full advantage is taken of symmetry in this method.

Equations (19.2) and (19.3) together with

$$LL^T \equiv A + E; \quad \| E \| \leqq n2^{-t} \tag{19.4}$$

enable us to assess the accuracy of Y. Because of the close relationship between the norms of L and L^{-1} and those of A and A^{-1} it is simpler to proceed to a direct computation of the relative error rather than by assessing $\| AY - I \|$; we comment on this in Section 20. It will simplify the assessment if we assume

$$n2^{-t} \| A^{-1} \| < 0.1; \quad n^2 2^{-t} \| A^{-1} \|^{\frac{1}{2}} < 0.1; \quad n \geqq 10. \tag{19.5}$$

Whether the first or the second of these is the more restrictive depends on the size of $\| A^{-1} \|$. In any case the first is more than adequate to insure against breakdown when computing L, and breakdown cannot occur anywhere else.

We assemble in equations (19.6) the deductions which can be made from the symmetry of A and relations (19.2) to (19.5):

$$\left.\begin{array}{lll}
\text{(a)} & \| L^{-1} \|^2 = \| (L^T)^{-1} \|^2 = \| (A + E)^{-1} \| \leqq 1.12 \| A^{-1} \| & \text{from S(ii)} \\
\text{(b)} & \| (A + E)^{-1} - A^{-1} \| \leqq 1.12 n\, 2^{-t} \| A^{-1} \|^2 & \text{from (9.13)} \\
\text{(c)} & 2^{-t} n^2 \| L^{-1} \| < 2^{-t} n^2 \times 1.06 \| A^{-1} \|^{\frac{1}{2}} < 0.106 & \text{from (a)} \\
\text{(d)} & \| X \| \leqq 1.07 \| L^{-1} \| < 1.14 \| A^{-1} \|^{\frac{1}{2}} & \\
\text{(e)} & \| F \| \leqq (0.7) 2^{-t} n^2 \| A^{-1} \|^{\frac{1}{2}}. &
\end{array}\right\} \tag{19.6}$$

Hence

$$\begin{aligned} XX^T &= [(L^T)^{-1} + (L^T)^{-1} F][L^{-1} + F^T L^{-1}] \\ &= (LL^T)^{-1} + H \quad \text{(say)} \end{aligned} \tag{19.7}$$

$$\begin{aligned} \| XX^T - (A + E)^{-1} \| = \| H \| &\leqq 2 \| F \| \, \| L^{-1} \|^2 + \| L^{-1} \|^2 \| F \|^2 \\ &< (1.6) 2^{-t} n^2 \| A^{-1} \|^{3/2} + (0.055) 2^{-t} n^2 \| A^{-1} \|^{3/2} \\ &< (1.66) 2^{-t} n^2 \| A^{-1} \|^{3/2}. \end{aligned} \tag{19.8}$$

The conditions (19.5) allow us to express the terms in $\| F \|^2$ in a variety of ways but it is convenient to have it in the same form as the main component of H. Finally we have

$$\begin{aligned} \| Y - XX^T \| = \| G \| &\leqq n^2 2^{-t} \| X \|^2 \\ &< 1.3 n^2 2^{-t} \| A^{-1} \| . \end{aligned} \tag{19.9}$$

Combining (19.6b), (19.8) and (19.9) we have for the relative error in Y,

$$\| Y - A^{-1} \| / \| A^{-1} \| < 1.12 n 2^{-t} \| A^{-1} \| + 1.66 n^2 2^{-t} \| A^{-1} \|^{\frac{1}{2}} + 1.3 n^2 2^{-t} \tag{19.10}$$

where the three terms on the right come from the calculation of L, X and Y respectively. By using fixed-point arithmetic in the calculation of X and Y we can reduce the n^2 in the last two terms to n, but if A is very ill-conditioned, so that $\| A^{-1} \|$ is large, the errors made in the triangularization are in any case the most important, and those made in the calculation of XX^T are the least.

Instead of solving $L^TX = I$ we could solve $LX = I$ and then calculate X^TX. The bound obtained for the relative error is identical in the two cases, though, of course, the computed inverses are not identical.

The Residual Matrix

20. In most of the error analysis we have worked with the residual matrix $(AY - I)$. It is worth noting that we can obtain a far better bound for this matrix if we solve $L^TX = I$ than if we solve $LX = I$. Further, the contributions made to the residual matrix from the separate parts of the computation are not in the same ratio as the contributions to the relative error.

Working with $L^TX = (I + F)$ we have

$$\begin{aligned} AY - I &= A[XX^T + G] - I \\ &= (A + E)XX^T - EXX^T + AG - I \\ &= LL^TXX^T - EXX^T + AG - I \\ &= L[I + F]X^T - EXX^T + AG - I \\ &= L[I + F][L^{-1} + F^TL^{-1}] - EXX^T + AG - I \\ &= LFL^{-1} + LF^TL^{-1} + (LFF^TL^{-1}) - EXX^T + AG. \end{aligned} \tag{20.1}$$

The order of magnitudes of the bounds for the norms of the components arising from E, F, G respectively are

$$1.3n2^{-t} \| A^{-1} \|, \quad 1.5n^22^{-t} \| A \|^{\frac{1}{2}} \| A^{-1} \| \quad \text{and} \quad 1.3 \| A \| n^22^{-t} \| A^{-1} \| . \tag{20.2}$$

The terms containing F and F^T make the same contributions, and that of the FF^T term is negligible. Note that the contribution of G to the residual may easily be the largest even when it has the smallest effect on the inverse itself.

On the other hand if we use

$$LX = I + F, \qquad Y = X^TX + G,$$

then F and G are bounded as before but

$$\begin{aligned} AY - I = LL^TF^T(L^{-1})^TL^{-1} &+ F \\ &+ LL^TF^T(L^{-1})^TL^{-1}F - EXX^T + AG. \end{aligned} \tag{20.3}$$

The orders of magnitude of the terms containing F^T and F are now

$$n^22^{-t} \| A \| \; \| A^{-1} \|^{3/2} \qquad \text{and} \qquad n^22^{-t} \| A^{-1} \|^{\frac{1}{2}}$$

respectively, so that if A is ill-conditioned the first of these may be serious. Except for specially constructed matrices, however, this disparity between the two terms does not occur in practice. This is because of the nature of the inverses of triangular matrices and is discussed in Sections 27–28.

Orthogonal Reduction to Triangular Form

21. We now consider the reduction of a matrix to triangular form by premultiplication with orthogonal matrices. In methods of this type the original matrix is transformed successively to $A_1, A_2, \cdots, A_s$, where each member of the sequence has at least one extra zero below the diagonal in addition to those in the previous matrix. These zeros may be produced one at a time (Givens [2]), or a whole subcolumn at a time (Householder [4]). Whatever the details of the method, the error analysis has certain features in common as we shall now show.

We may define the mathematical process as follows. Starting with

$$A_1X = I = B_1, \tag{21.1}$$

we produce

$$A_rX = B_r \qquad (r = 2, \cdots, s) \tag{21.2}$$

where

$$R_rA_r = A_{r+1} \qquad R_rB_r = B_{r+1} \qquad R_r^TR_r = I, \tag{21.3}$$

the matrices R_r being the successive orthogonal matrices.

We now consider the practical procedure, and we let A_r denote the *r-th computed matrix*. Corresponding to this A_r (N.B. this is not the matrix which would have been obtained by an exact computation in the previous stages), the technique defines an exact orthogonal matrix R_r. The computed matrix $\bar{R}_r$ differs from this, and we may write

$$\bar{R}_r \equiv R_r + E_r\,. \tag{21.4}$$

We then premultiply A_r by $\bar{R}_r$ and make further errors so that the computed A_{r+1} and B_{r+1} satisfy:

$$\left.\begin{aligned} A_{r+1} &\equiv \bar{R}_rA_r + F_r \\ B_{r+1} &\equiv \bar{R}_rB_r + G_r \end{aligned}\right\} \tag{21.5}$$

Note that we take those elements which should have been made zero to be exactly equal to zero, and F_r must include the errors implicit in this assumption. *If F_r is to be small it is therefore essential that $\bar{R}_rA_r$ should at least have small elements in place of those which should be made zero.*

We can now pursue the analysis in two different ways.

(A) We may write

$$\begin{aligned} A_{r+1} &\equiv R_rA_r + (E_rA_r + F_r) \equiv R_rA_r + H_r \\ B_{r+1} &\equiv R_rB_r + (E_rB_r + G_r) \equiv R_rB_r + K_r \end{aligned} \tag{21.6}$$

and assume that we can find bounds for $\| H_r \|$ and $\| K_r \|$ of the form

$$\| H_r \| \leqq h(n, r)2^{-t} \qquad \| K_r \| \leqq k(n, r)2^{-t}, \tag{21.7}$$

where h and k are simple functions of n and r.

The final set of equations is

$$A_s X = B_s, \tag{21.8}$$

a triangular set, and the computed X satisfies

$$A_s X \equiv B_s + M \quad \text{(say)}. \tag{21.9}$$

Premultiplying by $R_1^T R_2^T \cdots R_{s-1}^T$, we have

$$\left[A_1 + \sum_1^{s-1} S_i H_i\right] X \equiv I + \sum_1^{s-1} S_i K_i + S_{s-1} M \tag{21.10}$$

where

$$S_i = R_1^T R_2^T \cdots R_i^T. \tag{21.11}$$

Since all the R_i, and hence all the S_i, are orthogonal, we have

$$\left.\begin{aligned} &[A_1 + N]X \equiv I + P + Q \\ &\| N \| \leqq \sum_1^{s-1} \| H_i \| \\ &\| P \| \leqq \sum_1^{s-1} \| K_i \| \\ &\| Q \| = \| M \| . \end{aligned}\right\} \tag{21.12}$$

(B) We may work instead with $\bar{R}_r$ and assume that we have bounds for F_r and G_r of the form

$$\| F_r \| \leqq f(n, r) 2^{-t}; \qquad \| G_r \| \leqq g(n, r) 2^{-t}. \tag{21.13}$$

Premultiplying (21.9) now by $\bar{R}_1^{-1} \bar{R}_2^{-1} \cdots \bar{R}_{s-1}^{-1}$ we have

$$\left[A_1 + \sum_1^{s-1} S_i F_i\right] X \equiv I + \sum_1^{s-1} S_i G_i + S_{s-1} M \tag{21.14}$$

where

$$S_i = \bar{R}_1^{-1} \bar{R}_2^{-1} \cdots \bar{R}_i^{-1}. \tag{21.15}$$

Now the S_i are no longer exactly orthogonal, but suppose we can show that $\| E_r \| \leqq a < 1$. Then we have

$$\| \bar{R}_r^{-1} \| \leqq \frac{\| R_r^{-1} \|}{1 - \| E_r R_r^{-1} \|} \leqq \frac{1}{1 - a} \tag{21.16}$$

since R_r^{-1} is exactly orthogonal. Hence we have a relation of the same form as (21.12) in which:

$$\begin{aligned} &\| N \| \leqq \sum_1^{s-1} \frac{1}{(1 - a)^i} \| F_i \| \\ &\| P \| \leqq \sum_1^{s-1} \frac{1}{(1 - a)^i} \| G_i \| \\ &\| Q \| \leqq \frac{1}{(1 - a)^{s-1}} \| M \| \end{aligned} \tag{21.17}$$

Observe now that we can permit a value of a which is considerably greater than 2^{-t} without losing very much. Suppose for example that $a = 10^{-4}$. Now in the Householder method $s = n$, so that even for $n = 10^4$, the factor $1/(1 - a)^{s-1}$ is merely equal to e. We therefore see that although the errors made in pre-multiplying A^r by $\bar{R}_r$ should be kept as low as possible (of order 2^{-t}) and also that the *exact* $\bar{R}_r A_r$ should have elements of order 2^{-t} in the positions in which zeros should be produced, we can tolerate a matrix $\bar{R}_r$ which is by no means orthogonal to working accuracy. All we are really demanding of $\bar{R}_r$ is that both its norm and that of $\bar{R}_r^{-1}$ should be reasonably close to unity. Analysis (B) is clearly the more penetrating.

This contrasts with the situation when orthogonal transformations are used in connection with the symmetric eigenproblem. There we pre-multiply with $\bar{R}_r$ and post-multiply with $\bar{R}_r^T$ and we assume that the resulting matrix is still symmetric. Unless $\bar{R}_r$ is accurately orthogonal this will not be an accurate similarity transformation.

22. If the analysis is to give satisfactory results, it is essential that we should be able to obtain satisfactory bounds for the norms of H and K or for F and G, according as we use (A) or (B). The former will imply finding very good bounds for E_r, but the latter is not very demanding in this respect. Now in Gaussian elimination the main obstacle to finding bounds was the possibility of a progressive increase in size of the elements of the reduced matrices, since in general the current rounding error is proportional to the current size of the elements. With orthogonal (or nearly orthogonal) matrices this difficulty is no longer present. The Euclidean norm of each column of A_1 is preserved by exact orthogonal transformations. Hence, if originally the sum of the squares of the elements in each column is less than unity, this remains true in all later matrices.

In practice when working in fixed-point we must scale A_1 so that the accumulation of rounding errors cannot cause the norm of any column of any A_r to exceed unity and we must take $B_1 = \frac{1}{2}I$ for the same reason. In floating-point we need not take these extra precautions, but will need to deal with the possible slight increase in norms when we are estimating bounds for H_r and K_r or F_r and G_r.

Detailed Analysis for Givens' Reduction

23. We now give a brief analysis of the Givens' reduction. We work in floating-point since the gain involved in working in fixed-point with accumulation of scalar products is not substantial. We take A_1 to have normalization (II) originally.

Consider a typical stage in the reduction at which r columns have been made triangular and elements $(r + 2, r + 1); \cdots ; (k, r + 1)$ in the $(r + 1)$th colum have been made zero so that typically for $n = 5,\ r = 1,\ k = 3$ the current matrix is of the form

$$\begin{bmatrix} a_{11} & a_{12} & a_{13} & a_{14} & a_{15} \\ 0 & a_{22} & a_{23} & a_{24} & a_{25} \\ 0 & 0 & a_{33} & a_{34} & a_{35} \\ 0 & a_{42} & a_{43} & a_{44} & a_{45} \\ 0 & a_{52} & a_{53} & a_{54} & a_{55} \end{bmatrix} \tag{23.1}$$

The next step is to make the $(r + 1, k + 1)$-element zero by a rotation in the $(r + 1, k + 1)$-plane. This modifies only the elements in rows $(r + 1)$ and $(k + 1)$ and leaves the earlier zeros undisturbed. We shall assume that the Euclidean norms of all columns of A and B at this stage are bounded by

$$(1 \pm 9.2^{-t})^p,$$

where p is the number of transformations completed so far, and justify this assumption in our analysis of the next stage.

The components of the rotation matrix are computed from the relations

$$\begin{aligned} \cos\theta &\equiv \text{fl}[a_{r+1,r+1}/(a^2_{r+1,r+1} + a^2_{k+1,r+1})^{\frac{1}{2}}] \\ &\equiv \text{fl}\,[b/(b + c^2)^{\frac{1}{2}}] \quad \text{(say)} \end{aligned} \qquad (23.2)$$

$$\sin\theta \equiv \text{fl}[c/(b^2 + c^2)^{\frac{1}{2}}] \qquad (23.3)$$

$$\begin{aligned} \text{fl}(b^2 + c^2) &\equiv [b^2(1 + \epsilon_1) + c^2(1 + \epsilon_2)](1 + \epsilon_3) \\ &\equiv (b^2 + c^2)(1 + \epsilon_4)(1 + \epsilon_3) \end{aligned} \qquad (23.4)$$

$$\begin{aligned} \text{fl}(b^2 + c^2)^{\frac{1}{2}} &\equiv (b^2 + c^2)^{\frac{1}{2}}(1 + \epsilon_4)^{\frac{1}{2}}(1 + \epsilon_3)^{\frac{1}{2}}(1 + 2\epsilon_5) \\ &\equiv (b^2 + c^2)^{\frac{1}{2}}(1 + 3\epsilon_6) \end{aligned} \qquad (23.5)$$

where all $| \epsilon_i | \leqq 2^{-t}$ and we have assumed a square root subroutine which gives errors of less than 2 parts in 2^t. The calculated $\cos\theta$ and $\sin\theta$ therefore satisfy

$$\begin{aligned} \cos\theta &\equiv \text{fl}[b/(b^2 + c^2]^{\frac{1}{2}} \equiv \frac{b}{(b^2 + c^2)^{\frac{1}{2}}}\frac{1 + \epsilon_7}{(1 + 3\epsilon_6)} \\ \sin\theta &\equiv \text{fl}[c/(b^2 + c^2)^{\frac{1}{2}}] \equiv \frac{c}{(b^2 + c^2)^{\frac{1}{2}}}\frac{(1 + \epsilon_8)}{(1 + 3\epsilon_6)} \end{aligned} \qquad (23.6)$$

In the notation of the last section we have, by quite crude inequalities,

$$\| E \| = \| \bar{R} - R \| < 6.2^{-t} \qquad (23.7)$$

independent of the size of b and c. Hence in the notation of Section 21 we may take $a = 6.2^{-t}$, and it is evident that the departure from orthogonality is quite unimportant.

We now turn to the determination of the F and G matrices. These are null except in rows $(r + 1)$ and $(k + 1)$, and for F both rows are also null in positions 1 to r.

The element $(r + 1, k + 1)$ of F must be treated specially but bounds for each of the other non-zero elements of F and G are given either by

$$\delta_1 = |(x \cos\theta + y \sin\theta) - \text{fl}(x \cos\theta + y \sin\theta)| \qquad (23.8)$$

or by

$$\delta_2 = |(-x \sin\theta + y \cos\theta) - \text{fl}(-x \sin\theta + y \cos\theta)| \qquad (23.9)$$

where x and y are the existing elements in the $(r + 1)$ and $(k + 1)$ positions of a column of the current A or B. For δ_1 we have, ignoring 2^{-2t},

$$\begin{aligned}\delta_1 &\leqq |\, x \cos\theta + y \sin\theta \,|\, \epsilon_1 + |\, x\cos\theta\, \epsilon_2 + y \sin\theta\, \epsilon_3 \,| \,; \quad |\, \epsilon_i \,| \leqq 2^{-t} \\ &< 2^{-t+1} \sqrt{x^2 + y^2} < 2^{-t+1} \times \text{(Euclidean norm of column)}\end{aligned} \tag{23.10}$$

and the same result holds for δ_2. This enables us to justify the original assumption made about the norms of the columns of the transformed A and B. For the ratio of the norms of any column before and after this transformation certainly lies between $(1 \pm 9 \times 2^{-t})$. This is ample to cover the error in $\bar{R}$ and the errors δ_1 and δ_2 made in multiplying by it. All norms of all transformed columns therefore lie between $(1 \pm 9.2^{-t})^{\frac{n(n-1)}{2}}$. These norms will be of order unity since we will in any case need to have $n^2 \times 2^{-t}$ appreciably less than unity for a useful inverse. The exact value of the modulus of the $(k + 1, r + 1)$ element of $\bar{R}A$ is

$$\frac{|\, bc \,|}{(b^2 + c^2)^{\frac{1}{2}}} \, \frac{|\, \epsilon_7 - \epsilon_8 \,|}{1 + 3\epsilon_6}$$

which is certainly less than 2^{-t}. In replacing this by zero we are committing an error with a bound which is smaller than those we have obtained for other elements of rows $(r + 1)$ and $(k + 1)$. Combining these results we have for the current F and G,

$$\begin{aligned}\|\, F \,\| &< 2\sqrt{2(n - r)}\; 2^{-t}\, (1 + 9.2^{-t})^{\frac{n(n-1)}{2}} \\ \|\, G \,\| &< 2\sqrt{2n}\; 2^{-t}\, (1 + 9.2^{-t})^{\frac{n(n-1)}{2}}.\end{aligned} \tag{23.11}$$

Summing these results and taking $\sum_1^r s^{\frac{1}{2}} \doteq \frac{2}{3} r^{3/2}$, $\sum_1^r s^{3/2} \doteq \frac{2}{5} r^{5/2}$, we have from (21.12) and (21.17),

$$(A_1 + N)X \equiv I + P + Q \tag{23.12}$$

$$\left.\begin{aligned}\|\, N \,\| &< \frac{1}{(1 - 6.2^{-t})^{\frac{n(n-1)}{2}}} \, \frac{4\sqrt{2}}{5}\, n^{5/2}\, 2^{-t}\, (1 + 9.2^{-t})^{\frac{n(n-1)}{2}} \\ \|\, P \,\| &< \frac{1}{(1 - 6.2^{-t})^{\frac{n(n-1)}{2}}} \, \frac{n(n - 1)}{2}\, 2\sqrt{2n}\, 2^{-t}\, (1 + 9.2^{-t})^{\frac{n(n-1)}{2}} \\ \|\, Q \,\| &< \frac{1}{(1 - 6.2^{-t})^{\frac{n(n-1)}{2}}} \, \|\, M \,\|\end{aligned}\right\} \tag{23.13}$$

where we have replaced all $1/(1 - a)^s$ by $1/(1 - a)^{\frac{n(n-1)}{2}}$. We write for the moment

$$(1 + 9.2^{-t})^{\frac{n(n-1)}{2}} = 1 + d; \qquad \frac{1}{(1 - 6.2^{-t})^{\frac{n(n-1)}{2}}} = 1 + e. \tag{23.14}$$

Now M is the error made in solving the final triangular set of equations

$$A_s X = B_s\,, \tag{23.15}$$

so that we know from (10.10) and (10.17) that the computed X satisfies

$$A_s X \equiv B_s + M \tag{23.16}$$

where certainly

$$\begin{aligned} \| M \| &\leqq 2^{-t}[n^{\frac{1}{2}} \| B_s \| + 0.6(1 + d)n^2 \| X \|] \\ &< 2^{-t}(1 + d)[n + 0.6n^2 \| X \|] \end{aligned} \tag{23.17}$$

since the norm of each column of A_s and B_s is less than $(1 + d)$. We have therefore

$$\| AX - I \| < 2^{-t}(1 + d)\ (1 + e) \left[\frac{4\sqrt{2}}{5} n^{5/2} \| X \| \right. \\ \left. + \sqrt{2} n^{5/2} + n + 0.6n^2 \| X \| \right] \tag{23.18}$$

which is a result of the standard form. For a reasonably good inverse we shall require

$$2^{-t} n^{5/2} < 0.05 \quad \text{(say)}$$

and assuming $n \geqq 10$ this certainly gives

$$\left. \begin{aligned} &(1 + d) < 1.1 \\ &(1 + e) < 1.07 \\ &(1 + d)(1 + e) < 1.2. \end{aligned} \right\} . \tag{23.19}$$

We shall require further

$$2^{-t} n^{5/2} \| A^{-1} \| < 0.1$$

and if we assume this we certainly have when $n > 10$

$$\| AX - I \| < 2^{-t}[1.61 n^{5/2} \| X \| + 1.71 n^{5/2}]. \tag{23.20}$$

From (T) of Section 9 we obtain

$$\frac{\| X - A^{-1} \|}{\| A^{-1} \|} < 2^{-t} n^{5/2}\, [2.1\ \| A^{-1} \| + 1.8]. \tag{23.21}$$

The result is for a matrix with normalization (II). For normalization (I) we have immediately, in the usual way,

$$\frac{\| X - A^{-1} \|}{\| A^{-1} \|} < 2^{-t} n^{5/2}\, [2.1\, n^{\frac{1}{2}}\, \| A^{-1} \| + 1.8]. \tag{23.22}$$

The more dangerous term is the $(2.1)2^{-t} n^3 \| A^{-1} \|$ and this comes primarily from the rounding errors made in computing $\bar{R}A$ at each stage.

We cannot gain very much by working in fixed-point because there is no stage at which we can accumulate a scalar product of any magnitude. Care must be exercized in computing $\cos\theta$ and $\sin\theta$. The most satisfactory technique is to accumulate $(b^2 + c^2)$ exactly and then scale it by the largest 2^{2k} for which $2^{2k}(b^2 + c^2) < 1$. We then calculate $\cos\theta$ and $\sin\theta$ from $2^k b$ and $2^k c$. The computed matrix is rather closer to an orthogonal matrix than for floating-point computation and further, only one rounding error is made in computing each of $(x \cos\theta + y \sin\theta)$ and $(-x \sin\theta + y \cos\theta)$. The gain in accuracy is at best equivalent to a small constant factor and we cannot gain by a factor of $\sqrt{n}$.

Householder's Reduction

24. Only $(n - 1)$ transformations are involved in the Householder reduction. At the rth step we multiply by a matrix of the form $(I - 2w_r w_r^T)$ where w_r^T is of the form

$$(0\ 0 \cdots 0\ x_r\ x_{r+1} \cdots x_n); \quad \sum x_i^2 = 1 \tag{24.1}$$

and this produces zeros below the diagonal in the rth column. It leaves unchanged the first $(r - 1)$ rows of A and B and the first $(r - 1)$ columns of A, so that $(n - r + 1)^2$ elements of A are modified and $n(n - r + 1)$ elements of B. We are forced to round these elements before going on to the next stage so that we have no hope of having sharper bounds for F_r and G_r than

$$\| F_r \| \leqq (n - r + 1)2^{-t-1}; \quad \| G_r \| \leqq \sqrt{n(n - r + 1)}2^{-t-1}. \tag{24.2}$$

This gives for N and P,

$$\| N \| \leqq \tfrac{1}{2}n^2 2^{-t-1}; \quad \| P \| \leqq \tfrac{2}{3}n^2 2^{-t-1} \tag{24.3}$$

even if we take a, defined on p. 311, to be zero. The most we can hope to gain over Givens therefore is a factor of $\sqrt{n}$, and perhaps a small constant factor, if we compute $(I - 2w_r w_r^T)A_r$ and $(I - 2w_r w_r^T)B_r$ with sufficient care.

If we work in conventional floating-point arithmetic, we do not gain even the $\sqrt{n}$ factor. As before it is easy to see that the departure from orthogonality is slight and therefore unimportant and we may concentrate on the errors made in computing A_{r+1} itself.

We assume that we have an exact w_r and estimate the errors made in computing $A_r - 2w_r w_r^T A_r$. The first stage is typical so we deal with $A_1 - 2w_1 w_1^T A_1$. We have

$$p_1 \equiv \mathrm{fl}(w_1^T A_1) \equiv w_1^T A_1 + q_1 \quad \text{(say)} \tag{24.4}$$

and q_{1k} may be written in the form

$$q_{1k} = x_1 a_{1k} \epsilon_1 + x_2 a_{2k} \epsilon_2 + \cdots + x_n a_{nk} \epsilon_n \tag{24.5}$$

where

$$| \epsilon_r | \leqq (n + 1 - r)2^{-t}.$$

Now $\sum x_i^2 = 1$ and $\sum_i a_{ik}^2 = 1$ (the latter persisting approximately at later stages), but we cannot conclude anything sharper than

$$| q_{1k} | < n\, 2^{-t} \| q_1 \| < n^{3/2} 2^{-t}. \tag{24.6}$$

When computing $\mathrm{fl}(w_1 w_1^T A_1)$, the contribution from q_1 alone is $\| w_1 \| \, \| q_1 \|$ and the only bound we have for this is $n^{3/2} 2^{-t}$. Similarly in the later stages we have contributions of magnitude $r^{3/2} 2^{-t}$ and together they give the term $n^{5/2} 2^{-t}$. In practice [1], it is convenient, in floating-point computation, to express the equations in the form exhibited in equations (24.7) (the first step being typical).

$$\left.\begin{aligned} A_2 &= A_1 - K u_1 u_1^T A_1 \\ u_1^T &= (a_{11} \pm S, a_{21}, \cdots, a_{n1}); \quad S^2 = a_{11}^2 + a_{21}^2 + \cdots + a_{n1}^2 \\ K &= 1/(S^2 \pm a_{11} S) \quad \text{where } \pm S \text{ has the sign of } a_{11}. \end{aligned}\right\} \qquad (24.7)$$

This form avoids unnecessary square roots, but does not lead to a lower error bound.

In fixed-point arithmetic we can indeed realize the gain of the factor $n^{\frac{1}{2}}$, but only on a computer with good arithmetic facilities would this be worthwhile. To ensure an accurate determination of w, S^2 should be accumulated exactly and then multiplied by 2^{2k} where k is the greatest integer for which $2^{2k} S^2$ is not greater than zero. The vector w can then be calculated using $2^k a_{i1}$ instead of a_{i1}. This gives a very accurate determination of w and we need only concern ourselves with the errors made in computing $(A_1 - 2w_1 w_1^T A_1)$. In forming $p_1 = \mathrm{fi}(w_1^T A_1)$ we can accumulate scalar products, and since the columns of A_1 have norms bounded by unity, the elements of $w_1^T A_1$ are less than unity. If we write

$$p_1 \equiv \mathrm{fi}(w_1^T A) \equiv w_1^T A + q_1, \qquad (24.8)$$

then all elements of q_1 are less than 2^{-t-1} and hence $\| q_1 \| \leqq \sqrt{n} 2^{-t-1}$. Although the elements of $2w_1 w_1^T A_1$ may be greater than unity, we know that those of $A_1 - 2w_1 w_1^T A_1$ are not, and therefore no harm is done by the overspill of the former. To be safe throughout, we must scale A originally so that its columns are sufficiently less than unity for those of A_r to be less than unity throughout. With sufficient care we can obtain a computed inverse X satisfying

$$\frac{\| X - A^{-1} \|}{\| A^{-1} \|} \leqq 2^{-t} n^2 [a \| A^{-1} \| + b] \qquad (24.9)$$

where a and b are of order unity. For a matrix with normalization (I) this gives

$$\frac{\| X - A^{-1} \|}{\| A^{-1} \|} < 2^{-t} n^2 [a n^{\frac{1}{2}} \| A^{-1} \| + b]. \qquad (24.10)$$

Comparison of Methods for Inverting Unsymmetric Matrices

25. The method of Givens requires four times as many multiplications as Gaussian inversion and therefore a fair comparison is between single-precision Givens inversion and double-precision Gaussian. On this basis the Gaussian solution is much the better. The method of Householder on the other hand requires only twice as many multiplications as Gaussian inversion. However, we have already seen [(17.7) and Section 9, (T)] that if we are prepared to perform Gaussian elimination twice we can expect an inverse satisfying

$$\| X - A^{-1} \| / \| A^{-1} \| < (2.0)2^{-t}n^2R\| A^{-1} \|.$$

This is more favorable than that for Householder, even when the computation is carried out with extreme care, provided "R", the pivot ratio, is less than $\sqrt{n}$. In our experience this is almost invariably true and there are many important types of matrix for which we know in advance that it will be true.

Comparing Gaussian elimination with Householder when both are performed in floating-point, we have a relative error of $2^{-t}n^{7/2}r\| A^{-1} \|$ for the former and $2^{-t}n^3\| A^{-1} \|$ for the latter. In our view the presence of the factor $n^{7/2}$ springs rather from the exigencies of the analysis than from any true shortcoming, and the case for the Householder is slight.

The other method for which we have an error analysis is that of Goldstine and von Neumann based on the inversion of the positive definite matrix, AA^T. However, this requires $2n^3$ multiplications and has $\| A^{-1} \|^2$ in place of $\| A^{-1} \|$ in error terms. If A is at all ill-conditioned, this method has nothing to recommend it.

General Comments on Results

26. The bounds we have obtained are in all cases strict upper bounds. In general, the statistical distribution of the rounding errors will reduce considerably the function of n occurring in the relative errors. We might expect in each case that this function should be replaced by something which is no bigger than its square root and is usually appreciably smaller. For example, in forming a scalar product in floating-point we have used the result

$$\mathrm{fl}(\sum_1^n a_ib_i) \equiv \sum a_ib_i(1 + \epsilon_i) \tag{26.1}$$

and have then replaced the ϵ_i by their upper bounds, $(n - i)2^{-t}$. Now for these bounds to be attained not only must all the individual rounding errors have their greatest values but the a_ib_i must have a special distribution.

On the other hand the term $\| A^{-1} \|$ which invariably occurs in the relative error is not usually subject to any such considerations. It is useful to have some simple standard by which to judge the relative error and this is provided by the following considerations. Consider the effect on the inverse of perturbations of order 2^{-t} in the elements of A. We have

$$\|(A + E)^{-1} - A^{-1} \| \leqq \frac{\| A^{-1} EA^{-1} \|}{1 - \| EA^{-1} \|}. \tag{26.2}$$

If these perturbations are not specially correlated, then we may write

$$\| EA^{-1} \| \leqq \| E \| \| A^{-1} \| \tag{26.3}$$

and expect that the right-hand side is a reasonable approximation to the left.

We have therefore

$$\|(A + E)^{-1} - A^{-1}\|/\|A^{-1}\| \leqq \frac{\|E\|\|A^{-1}\|}{1 - \|E\|\|A^{-1}\|} < \frac{n2^{-t}\|A^{-1}\|}{1 - n2^{-t}\|A^{-1}\|} \tag{26.4}$$

and we would expect to be able to choose some E for which the right-hand side was not a severe over-estimate. If then a method of inversion gives an inverse X for which

$$\|X - A^{-1}\|/\|A^{-1}\| \leqq f(n)2^{-t}\|A^{-1}\|, \tag{26.5}$$

we could say that the errors in the inverse are of the order of magnitude of those which could be caused by perturbations of order $(1/n)f(n)2^{-t}$ in the individual elements of A. *It might be thought that the computed X would be the exact inverse of some matrix $(A + F)$ where the elements of F were of order $(1/n)f(n)2^{-t}$ but this is seldom true. Indeed the elements of the inverse of X usually differ from those of A by quantities which are of the order of $(1/n)f(n)2^{-t}\|A^{-1}\|$*. Since $\|A^{-1}\|$ will be very large if A is ill-conditioned, this last result is a little surprising. However it *is*, in general, true that the rth column of X is the rth column of the exact inverse of $(A + E_r)$ where E_r has elements of order $(1/n)f(n)2^{-t}$, but it requires a different E_r for each column. The reader may readily verify these comments for an ill-conditioned 2 × 2 matrix. They are true equally of symmetric and unsymmetric matrices.

High Accuracy of Computed Inverses

27. It is a common experience that even when the most generous estimate is made for the reduction of the error by statistical effects, computed inverses are much more accurate than might be expected. It is unlikely that there is a single simple explanation of this phenomenon, but the following example in which the effect is very marked does give considerable insight into some of the main causes. We consider the inversion, working with 8 decimals, of a symmetric segment of order five of the Hilbert matrix. The matrix is positive definite and $\|A^{-1}\|$ is approximately 10^6. It is so ill-conditioned that truncation of the elements of the original matrix to 8 decimals already affects the second figure of the inverse. This is to be expected since $\|A^{-1}\|\|E\|\|A^{-1}\|$ is of order

$$(10^6)^2 \times 10^{-8}, \quad \text{i.e. } 10^4.$$

Naturally we compare our computed inverse with the exact inverse of the truncated matrix and not with that of the exact Hilbert segment! We denote the exact Hilbert segment by H, the matrix with the truncated elements by $\bar{H}$, and the computed inverse of $\bar{H}$ by X. Now we see from Table 3 that

$$\max|H^{-1} - \bar{H}^{-1}|_{ij} \doteqdot 6{,}000.0. \tag{27.1}$$

On the other hand, for X computed by the techniques of Section 19, we see that

$$\max|\bar{H}^{-1} - X| \doteqdot 1.0. \tag{27.2}$$

The total effect of all the rounding errors made in the process of solution is far less than those which come from the initial truncation.

There are two main phenomena which account for the remarkable accuracy of the solution.

(i) We saw in Section 18 that $LL^T = \bar{H} + E$ and that a typical element of E is bounded by $2^{-t-1}| l_{ii} |$. If some of the elements $| l_{ii} |$ are much less than unity, the bounds for the corresponding elements of E are far smaller than 2^{-t}. Now because $\bar{H}$ is so ill-conditioned, the l_{ii} are small and become progressively smaller with increasing i. The matrix E is displayed and it will be seen that its elements also become progressively smaller as we move to the bottom right-hand corner. Here they are far smaller than $\frac{1}{2}\cdot 10^{-8}$ (the optimum value which we might expect for 8 decimal computation). Note that LL^T is so close to $\bar{H}$ precisely because $\bar{H}$ is so ill-conditioned!

We now have to consider the effect of the errors E, on the inverse. The perturbation resulting from a change ϵ in the (i, j)-element is

$$-\epsilon \text{ (}i\text{th column of } \bar{H}^{-1}) \text{ (}j\text{th row of } \bar{H}^{-1})/(1 + \epsilon \bar{H}_{ji}^{-1});$$

now the denominator is approximately unity and the ith row and jth column of $\bar{H}^{-1}$ increase generally in size with increasing i and j. This means that the smallest elements of E are in precisely those positions which have the most effect on the inverse. The result is that we do not get the full effect of the $\| \bar{H}^{-1} \|$ in the computed inverse. The effect we have just noticed is very common in ill-conditioned matrices with elements which are of a systematic nature. It becomes even more marked if we consider Hilbert segments of higher order. The presence of very small elements in E is also quite common in ill-conditioned matrices which are less obviously of this type.

(ii) The second phenomenon is of a more general nature and is associated with the inversion of triangular matrices. We have seen that for a matrix A of general form, the effect of perturbations in the end figures is to introduce a relative error which is usually proportional to $\| A^{-1} \|$, and an absolute error which is proportional to $\| A^{-1} \|^2$. Such an effect is avoided only if the rounding errors are peculiarly correlated. Now whereas for general matrices such a correlation is very uncommon, it is so common for triangular matrices as to constitute the rule rather than the exception (Sect. 28). In formal terms we have, quite commonly, for the computed inverse of a triangular matrix,

$$\frac{\| X - A^{-1} \|}{\| A^{-1} \|} < a(n)2^{-t} \tag{27.3}$$

in contrast to the more common result for a general matrix,

$$\frac{\| X - A^{-1} \|}{\| A^{-1} \|} < b(n)2^{-t} \| A^{-1} \| . \tag{27.4}$$

This means that for triangular matrices it is common for the relative error to be unaffected by the condition of the matrix. The triangular matrix in Table 3 is one such example. *None of the elements in the computed inverse X has an error*

TABLE 3

$\bar{H}$					L				
.5000 0000	.3333 3333	.2500 0000	.2000 0000	.1666 6667	.7071 0678	.4714 0452	.3535 5339	.2828 4271	.2357 0227
	.2500 0000	2000 0000	.1666 6667	.1428 5714		.1666 6667	.2000 0000	.2000 0002	.1904 7614
		.1666 6667	.1428 5714	.1250 0000			.0408 2488	.0699 8519	.0874 8183
			.1250 0000	.1111 1111				.0101 0280	.0224 4694
				.1000 0000					0025 1879
$10^8 E = 10^8(LL^T - \bar{H})$					*Computed solution of* $L^T X = I$				
−.1678 0316	+.2214 6456	−.0839 0158	−.2085 4262	+.3178 3906	+1.4141	−4.00000	+7 34846	−11.31024	+15.73580
	+.0365 3193	+.0107 3228	−.0723 6174	−.0579 4858		+6.00000	−29.39384	+84.84116	−188.92188
		+.0407 5065	−.0059 1859	−.0198 8743			+24.49487	−169.68344	+661.43599
			−.0012 6794	+.0049 0042				+98.98246	−882 11140
				−.0008 9709					+397.01603
Computed inverse from XX^T					*True inverse of* $\bar{H}$				
+447 58	−4172.56	+12507.69	−15000.42	+6247 36	+447.58	−4172.61	+12507.82	−15000.57	+6247.43
	+41789 50	−140075.87	+175047.93	−75005.01		+43789.81	−140076.57	+175048.57	−75005.21
		+466890.04	−600255.91	+262600.69			+466891 08	−600256 18	+262600.48
			+787918 05	−350212.37				+787916 87	−350211.37
				+157621.73					+157621.07
Inverse of H									
+450	−4200	+12600	−15120	+6300					
	+44100	−141120	+176400	−75600					
		+470400	−604800	+264600					
			+793800	−352800					
				+158760					

which is greater than one in the last figure retained in the computation. This means that the errors made in the triangular inversion are negligible and we already know that the errors involved in computing XX^T are negligible. The combined effect is to produce an inverse of quite astonishing accuracy. The total effect of all the rounding errors is far smaller than that corresponding to a single perturbation of $\frac{1}{2}\cdot 10^{-8}$ in the (5, 5) element of $\bar{H}$.

The Inversion of Triangular Matrices

28. The remarkable property of triangular matrices mentioned in the last section is of such widespread significance in practice that it is perhaps worth indicating the types of triangular matrix which possess it.

Perhaps the most important class is of those having positive diagonal elements and negative off-diagonal elements. We consider, without loss of generality, lower triangular matrices and show that if $\bar{x}_{ij}$ is the computed (i, j)-element and x_{ij}, the exact, then

$$\frac{\bar{x}_{ij}}{x_{ij}} \equiv (1 + \epsilon_{ij}); \quad |\epsilon_{ij}| \leqq 3(i + 1 - j)2^{-t} \tag{28.1}$$

however ill-conditioned L may be. The proof is by induction. We need consider only the first column of the inverse, since in general the rth column is derived from the triangular matrix obtained by setting the first $(r - 1)$ columns of L equal to zero and this has the same properties. Assuming that (28.1) is true with $j = 1$ for $i = 1, \cdots, r - 1$, we have, with an obvious abbreviated notation:

$$\begin{aligned} \bar{x}_{r1} &\equiv -\mathrm{fl}[(l_{r1}\bar{x}_{11} + l_{r2}\bar{x}_{21} + \cdots + l_{r,r-1}\bar{x}_{r-1,1})/l_{rr}] \\ &\equiv -[\{l_{r1}\bar{x}_{11}(1 \pm \theta\overline{r+1}\cdot 2^{-t}) + l_{r2}\bar{x}_{21}(1 \pm \theta r\cdot 2^{-t}) + \cdots \\ &\qquad + l_{r,r-1}\bar{x}_{r-1,1}(1 \pm \theta 3\cdot 2^{-t})\}/l_{rr}] \qquad\qquad (28.2) \\ &\equiv -\ [\{l_{r1}x_{11}(1 \pm \theta\overline{r+4}\cdot 2^{-t}) + l_{r2}x_{21}(1 \pm \theta\overline{r+6}\cdot 2^{-t}) \\ &\qquad + \cdots + l_{r,r-1}x_{r-1,1}(1 \pm \theta 3r\cdot 2^{-t})\}/l_{rr}]. \end{aligned}$$

Now from our assumption about the signs of the $l_{r,i}$, all $\bar{x}_{ij}$ and x_{ij} are positive. Further we have

$$x_{r1} \equiv -[l_{r1}x_{11} + l_{r2}x_{21} \cdots l_{r,r-1}x_{r-1,1}/l_{rr}], \tag{28.3}$$

and hence $\bar{x}_{r1}/x_{r1}$ lies between $(1 + 3r\cdot 2^{-t})$ and $(1 - 3r\cdot 2^{-t})$; this completes the proof.

The nature of the proof shows that the result is not sharp and we would expect something appreciably better than

$$\frac{\bar{x}_{ij}}{x_{ij}} \equiv (1 \pm \sqrt{3(i + 1 - j)}\, 2^{-t}). \tag{28.4}$$

Even for very high order matrices this gives a very good result. It shows that if we use floating-point arithmetic, *then even the small elements have a low relative* *error*. Matrices of this kind are produced when Gaussian elimination is performed on the matrices derived from finite-difference approximations to elliptic partial differential equations.

When complete pivoting for size is used then we can prove a much weaker result which is, however, independent of the sign distribution of the elements of L. We show that if $| l_{ii} | \geqq | l_{ij} |$, then for fixed-point computation with a fixed scale factor,

$$| \bar{x}_{ij} - x_{ij} | \leqq 2^{i+1-j} \max_k | x_{kj} | 2^{-t}. \tag{28.5}$$

The proof of this is immediate by induction, since we have, typically for the first column,

$$\begin{aligned} \bar{x}_{r1} &\equiv -\mathrm{fi}[(l_{r1}\bar{x}_{11} + l_{r2}\bar{x}_{21} + \cdots + l_{r,r-1}\bar{x}_{r-1,1})/l_{rr}] \\ &\equiv -[(l_{r1}\bar{x}_{11} + l_{r2}\bar{x}_{21} + \cdots + l_{r,r-1}\bar{x}_{r-1,1})/l_{rr}] \pm 2^{p-t-1} \end{aligned} \tag{28.6}$$

where 2^p is the scale factor.

Hence

$$\begin{aligned} | \bar{x}_{r1} - x_{r1} | &\leqq \left|\frac{l_{r1}}{l_{rr}}\right| | \bar{x}_{11} - x_{11} | + \cdots + \left|\frac{l_{r,r-1}}{l_{rr}}\right| | \bar{x}_{r-1,1} - x_{r-1,1} | + 2^{p-t-1} \\ &\leqq (2^1 + 2^2 + \cdots + 2^{r-1}) \max | x_{k1} | 2^{-t} + 2^{p-t-1} \\ &\leqq 2^r \max | x_{k1} | 2^{-t} \end{aligned} \tag{28.7}$$

since

$$2^p < 2 \max | x_{i1} |.$$

This result is not very spectacular but, for small order matrices, 2^n may be far smaller than $\| L^{-1} \|$. Hence if we have used complete pivoting on a matrix of low order we are certain to get a "comparatively good" result when the matrix is ill-conditioned.

When Gaussian elimination with complete pivoting is used on a matrix which has one isolated very small eigenvalue then it is common for "all the ill-condition" to be concentrated in the final element. This happens, for example when $(A - \lambda I)$ is reduced to triangular form for a value of λ which is very close to an isolated eigenvalue. The triangular matrix is then commonly of the form typified in (28.8):

$$\begin{bmatrix} .31265 & .12321 & .21623 \\ & .41357 & .41632 \\ & & .00001 \end{bmatrix} \tag{28.8}$$

When this set of equations is solved with any right-hand side, the computed solution has a low relative error in all components. It is commonly believed that it is a poor strategy to use a method of reduction that concentrates all the smallness in the last pivot since it means that this last pivot has a high relative error. In our opinion there is no substance in this belief. The emergence of a small last pivot shows that the matrix is ill-conditioned, since whatever pivoting may have been done, the reciprocal of the last pivot is an element of the inverse matrix. On a computer on which scalar products can be accumulated, it is usually better to use a pivotal strategy which results in a very small last pivot than one which does not. In the example of the last section for instance it is better to work with the matrix as it was, rather than working with the pivots in positions 4, 5, 3, 2 and 1 in that order, in spite of the fact that the former leads to the smaller value for the last pivot. If scalar products cannot be accumulated, then there is a very slight advantage in a strategy which does not make the last pivot very small. It should be realized that the harm that comes from a small last pivot is primarily a property of the matrix and is not something induced by the method of solution. When scalar products are not accumulated exactly, the perturbation in the original element of A which corresponds to the last pivot is the sum of $(n - 1)$ rounding errors, whereas that in the element corresponding to the rth pivot is the sum of $(r - 1)$ rounding errors. Ideally we should try to make the last pivot correspond to the element i, j for which the product of the norms of the ith column and the jth row of A^{-1} is as small as possible. There does not seem to be any practical strategy which assures this.

The triangular matrix of Section 27 does not belong to any of the classes we nave mentioned so far. We now describe a much larger class of matrices for which our result is true. Consider the solution of a lower triangular set of equations in fixed-point arithmetic, using a constant scale factor 2^k. Let $\bar{X}$ be the computed inverse and X the true inverse. We have

$$L\bar{X} \equiv I + E \tag{28.9}$$

$$\begin{aligned} |E| &< 2^{k-t-1} \begin{bmatrix} |l_{11}| & & & \\ |l_{22}| & |l_{22}| & & \\ \cdots & \cdots & \cdots & \cdots \\ |l_{nn}| & |l_{nn}| & \cdots & |l_{nn}| \end{bmatrix} \\ &= 2^{k-t-1} \operatorname{diag} |l_{ii}| \begin{bmatrix} 1 & & & \\ 1 & 1 & & \\ \cdots & \cdots & \cdots & \cdots \\ 1 & 1 & \cdots & 1 \end{bmatrix} \end{aligned} \tag{28.10}$$

Suppose X is such that $|x_{ji}/x_{ii}| < C$ for $j > i$. Now the exact solution is $\bar{X} - XE$ so that the error matrix is bounded by $|XE|$ and we have

$$|XE| \leqq |X|\,|E| < C \begin{bmatrix} |x_{11}| & & & \\ |x_{11}| & |x_{22}| & & \\ \cdots & \cdots & \cdots & \cdots \\ |x_{11}| & |x_{22}| & \cdots & |x_{nn}| \end{bmatrix} \cdot |E|. \tag{28.11}$$

Hence, since $x_{ii} = 1/l_{ii}$,

$$| XE | < C2^{k-t-1} \begin{bmatrix} 1 & & & \\ 1 & 1 & & \\ \cdots & \cdots & \cdots & \cdots \\ 1 & 1 & \cdots & 1 \end{bmatrix}^2 \tag{28.12}$$

giving

$$\max | (XE)_{ij} | < Cn2^{k-t-1} < Cn2^{-t} \max | \bar{X}_{ij} |. \tag{28.13}$$

This shows that the error in any element of X is a small multiple of its maximum element. In the example of the last section, C is less than $2\frac{1}{2}$ and the ratio $| x_{ji}/x_{ii} |$ is less than one for many of the components.

However, the result we have just proved should not mislead us into thinking that it is essential that the components of X should not increase rapidly as we move along any column in a direction away from the diagonal. On the contrary this type of behavior is usually extremely favorable. Consider, for example, the matrix of order n with elements

$$\begin{bmatrix} \text{`0.1'} & \text{`1.0'} & & & \\ & \text{`0.1'} & \text{`1.0'} & & \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ & & & \text{`0.1'} & \text{`1.0'} \\ & & & & \text{`0.1'} \end{bmatrix} \tag{28.14}$$

where we use 'a' to denote a number of the order of magnitude of a, but with end figures which will lead to rounding errors. Now the inverse of this matrix contains an element of order 10^n. If n is at all large it is therefore very ill-conditioned. However, it is obvious in the light of our previous analysis that the elements of the computed solution have a very low relative error.

An attempt to construct a matrix for which the result is not true reveals how widespread the phenomenon is likely to be. An example is the matrix

$$\begin{bmatrix} 10^{-10} & .9 & -.4 \\ & .9 & -.4 \\ & & 10^{-10} \end{bmatrix} \tag{28.15}$$

The last column of the exact inverse is $[0, \frac{4}{9}10^{10}, 10^{10}]$. The last column of the computed inverse is $[4 \times 10^9;\ 44444\,44444;\ 10^{10}]$ so that we really do have a computed inverse with a relative error of order $\| A^{-1} \| \times 10^{-10}$. A matrix of low order must be very specially designed if it is to give such a result. If we have used complete pivoting, then we cannot have such examples in matrices of low order.

Selecting three matrices at random of order 18, 25 and 43 from moderately ill-conditioned unsymmetric matrices inverted on DEUCE, we found in all cases that the inversion of computed L and computed U and the computation of $L^{-1}U^{-1}$ made practically no contribution to the errors in the inverse. This shows the great importance of keeping the difference $(LU - A)$ as small as possible since this is usually the main source of error. If partial pivoting for size is used it is easy to ensure that $|(LU - A)_{ij}| < R2^{-t-1}$ where R is the pivotal growth factor [7]. It does not appear to be possible with complete pivoting, without using twice as much storage as in the ordinary reduction or alternatively doing the computation twice.

It is natural to ask why this accuracy of computed inverses of triangular matrices was not revealed by our earlier analysis. The reason is that the analysis is based on an assessment of $\| X - L^{-1} \|$ derived from a bound for $\| LX - I \|$. Now the norm of $(LX - I)$ is not significantly smaller in the cases when X has the exceptional accuracy than in cases when it has not. The size of the residual gives a very crude estimate of the accuracy of X and the range covered by approximate inverses having residual matrices of a prescribed size is remarkably wide when the original matrix is ill-conditioned. We are able to put the residual matrix to such good use when analyzing computed inverses of general matrices, only when the methods used are such as to bias the residual in such a way as to be almost as small as possible having regard to the accuracy of the computed inverse.

The really important norm is not that of $(LX - I) = E$ (say) but that of $L^{-1}E$. Now it happens that for triangular matrices, E can usually be expressed in the form LF where F has much the same norm as E. This gives $L^{-1}E$ much the same norm as E itself. We can now see why it is that the progressive introduction of scale factors in the inversion of a triangular matrix usually gives a better result than that in which the same scale factor is used throughout, although the latter gives the smaller residual. We may illustrate it by taking an extreme case. Suppose no scale factor is required until the very last stage, and this requires quite a large scale factor, 2^k. The computed values $x_n, x_{n-1}, \cdots, x_2$ will then have to be rounded to k less figures to give $x_n', x_{n-1}', \cdots, x_2'$ with $|x_i - x_i'| \leqq 2^{k-t-1}$. The contributions to the residuals corresponding to this change are precisely $L(x - x')$.

Now although this may constitute the major part of the residual as far as magnitude is concerned, it is obvious that the solution of $Ly = L(x - x')$ is $(x - x')$, whereas if the other part of the residual is E_1 the solution of $Ly = E_1$ is $L^{-1}E_1$ and may be very large. The increased size of the residual is entirely misleading. On the other hand, because the early stages have been done without the scale factor, the residuals just before computing the last element x_1 are bounded by $2^{-t-1}| l_{ii} |$ instead of $2^{k-t-1}| l_{ii} |$.

When the residual matrix R, defined by $(AX - I)$, is used to assess the error in an approximate inverse X, then we can, in general, make no claim unless $\| R \| < 1$. Thus Goldstine and von Neumann [3, p. 1080] say that unless $\| R \| < 1$, even the null matrix gives a better residual. Although this is true it is reasonable to regard a matrix X, for which $\| X - A^{-1} \| / \| A^{-1} \| \ll 1$, as a much better inverse than the null matrix. Indeed if a normalized matrix A has an inverse such that $\| A^{-1} \|$ is of order 10^{20}, then if X is the matrix obtained by rounding the elements of A^{-1} to 10 significant figures, the absolute errors in some elements of X may be as large as $\frac{1}{2}\ 10^{10}$ and the residual matrix may have components as large $(n/2)\ 10^{10}$. Yet it would be reasonable to regard X as a "good" inverse.

Failure of Partial Pivoting Strategy

29. That the partial pivoting strategy can fail completely on a well-conditioned matrix is well illustrated by the following example. Consider matrices of the form illustrated in (29.1):

$$A_6 = \begin{bmatrix} +1 & 0 & 0 & 0 & 0 & +1 \\ +1 & +1 & 0 & 0 & 0 & -1 \\ -1 & +1 & +1 & 0 & 0 & +1 \\ +1 & -1 & +1 & +1 & 0 & -1 \\ -1 & +1 & -1 & +1 & +1 & +1 \\ +1 & -1 & +1 & -1 & +1 & -1 \end{bmatrix} \tag{29.1}$$

The inverses are of the form illustrated in (29.2):

$$A_6^{-1} = \begin{bmatrix} +2^{-1} & +2^{-2} & -2^{-3} & +2^{-4} & -2^{-5} & +2^{-5} \\ 0 & +2^{-1} & +2^{-2} & -2^{-3} & +2^{-4} & -2^{-4} \\ 0 & 0 & +2^{-1} & +2^{-2} & -2^{-3} & +2^{-3} \\ 0 & 0 & 0 & +2^{-1} & +2^{-2} & -2^{-2} \\ 0 & 0 & 0 & 0 & 2^{-1} & +2^{-1} \\ +2^{-1} & -2^{-2} & +2^{-3} & -2^{-4} & +2^{-5} & -2^{-5} \end{bmatrix} \tag{29.2}$$

Consider now a matrix A_{31}, of this form. It is clearly a very well-conditioned matrix. Let B_{31} be the matrix obtained by replacing its (31, 31)-element by $\frac{1}{2}$. This matrix is also well-conditioned and its inverse differs from that of A_{31} by $+\frac{1}{2}$ (last column of A_{31}^{-1}) (last row of A_{31}^{-1})$/(1 - 2^{-31})$. Some of the alterations are therefore as big as 2^{-3} so that the inverses differ in the first decimal. Now if we invert these matrices by Gaussian elimination with the partial pivotal strategy using floating-point computation with a 30 binary digit mantissa, then in the reductions all multipliers are identical and all reduced matrices differ by exactly $\frac{1}{2}$ in the (31, 31)-element until we reach the final step. In the very last move A_{31} should give the value 2^{30} and B_{31} should give the value $2^{30} - \frac{1}{2}$ for the final element of the triangle. The latter is rounded to 2^{30} so that both A_{31} and B_{31} give identical upper triangular matrices. They therefore produce identical L and U matrices and therefore identical inverses! That of the B_{31} matrix has errors of norm greater than $(0.1)B_{31}^{-1}$. The use of exact numbers plays no essential part but merely aids the demonstration. If we replace all elements of A_{31} by elements differing only in the last five binary digits but such that the pivotal decisions are unaltered, then we will obtain a poor inverse in nearly all cases. We know from our *a priori* analysis that the complete pivotal strategy gives a good inverse for A_{31}, B_{31} and all the perturbations of A_{31}.

The example also exposes a common fallacy. In the computation with B_{31} *all the pivotal values have a very low relative error.* In fact all of them are correct except the last which has an error of only one part in 2^{31}. In spite of this the results are useless. Significant figure arguments are much too superficial in general to give reliable indications, and reflection back into the original matrix is a much more reliable guide.

Iterative Methods of Solution

30. There is still a widely held belief that iterative methods of solving linear equations are less affected by rounding errors than direct methods because in the former one continues to work with the original matrix instead of modifying it. Our analysis has shown that the total effect of the rounding errors made in good methods of solution corresponds to only very small changes in the elements of the original matrix. It is interesting to compare these perturbations with those which "effectively" occur when iteration is performed with the original matrix. Consider a typical step in a Gauss-Seidel iteration performed in floating point. Typically an improved x_1 is computed from the previous $x_2 x_3 \cdots x_n$ by the relation

$$\begin{aligned} x_1 &\equiv \mathrm{fl}\left[\frac{b_1 - a_{12}x_2 - a_{13}x_3 - \cdots - a_{1n}x_n}{a_{11}}\right] \\ &\equiv [b_1(1+\epsilon_1) - a_{12}x_2(1+\epsilon_2) \\ &\qquad - a_{13}x_3(1+\epsilon_3) - \cdots - a_{1n}x_n(1+\epsilon_n)]/a_{11} \end{aligned} \tag{30.1}$$

with the usual bounds for the ϵ_i. We have therefore performed an exact Gauss-Seidel step with a matrix with elements $b_1(1+\epsilon_1)$ and $a_{1i}(1+\epsilon_i)$. The bounds for the $a_{1i}\epsilon_i$ are no better than the perturbations we obtained in Gassian elimination and further we obtain a different set of ϵ_i for each iteration. The idea that we have worked with the original matrix is an illusion. Even if we accumulate scalar products exactly we still do not do an exact iteration with A and b and ill-conditioning of A still affects the computation adversely. This is well-illustrated by the equations:

$$\left.\begin{aligned} .96326x_1 + .81321x_2 &= .88824 \\ .81321x_1 + .68654x_2 &= .74988 \end{aligned}\right\} \tag{30.2}$$

These equations are positive definite and therefore Gauss-Seidel iteration should be convergent, though very slowly. However, if we work with five decimals and accumulate scalar products exactly, then starting with $x_1 = .33116$, $x_2 = .70000$, for example, no progress is made towards the true solution

$$x_1 = .39473\cdots, \quad x_2 = .62470\cdots.$$

Convergence is so slow that the modifications to be made in the first step are less than $\frac{1}{2}10^{-5}$. We may express this in similar terms to those used earlier, if we observe that x_1 is computed from the relation

$$x_1 \equiv \mathrm{fi}\left[\frac{.88824 - (.81321).7}{.96326}\right] \equiv \mathrm{fi}\left[\frac{.318993}{.96326}\right] = .33116. \tag{30.3}$$

This would be an exact result if the denominator were .318993/.33116 = .963259 $\cdots$. The computed x_1 corresponds to an exact computation with a perturbed a_{11} element. Similarly the computed x_2 corresponds to an exact computation with a perturbed a_{22} element. Starting with almost any x_2 in the range .4 to .8 the iterated vectors remain constant.

Relative Effectiveness of Fixed-point and Floating-point Computation

31. At most stages in this paper the results obtained for fixed-point computation have been better than those for floating-point computation. Strictly speaking the term fixed-point is something of a misnomer since what we have called fixed-point computation has usually involved the introduction of scale factors. We are therefore comparing a technique in which the scale factors are introduced only as and when they are necessary (so-called fixed-point) and that in which scale-factors are used automatically at every stage (floating-point). The superiority of the fixed-point scheme springs entirely from the ability to accumulate exact scalar products. Now the inability to do this conveniently in floating-point is not fundamental, it is merely a result of the design of existing computers. Further there are some *fixed-point* computers on which it is inconvenient to accumulate exact scalar products. On ACE, where floating-point operations are performed by subroutines, it was found that the routines for addition and subtraction could, without any appreciable loss of speed, deal with double-precision floating numbers. It has therefore been possible to make a set of floating-point subroutines which add double-precision numbers, divide a single-precision number into a double-precision number but multiply only single-precision numbers, all operations being essentially at the speed associated with single-precision working. With these subroutines we have all the convenience of floating-point working with the full advantage of accurate accumulation of scalar products. The provision of automatic floating-point facilities of this nature would be of great value in matrix work.

Acknowledgments

The work described here has been carried out as part of the Research Programme of the National Physical Laboratory and this paper is published by permission of the Director of the Laboratory. The author wishes to thank G. E. Forsythe and E. T. Goodwin for reading the manuscript and making many valuable suggestions.

REFERENCES

1. Bauer, F. L. Sequential reduction to tridiagonal form *J. Soc. Indust. Appl. Math. 7* (1959), 107.

2. Givens, W. The linear equations problem Technical Report no. 3. Applied Mathematics and Statistics Series, Nonr 225 (37). Stanford University, 1959.

3. Goldstine, H. H.; and von Neumann, J. Numerical inverting of matrices of high order. *Bull. Amer. Math. Soc. 53* (1947), 1021.

4. HOUSEHOLDER, A. S Unitary triangularization of a nonsymmetric matrix *J. ACM 5* (1958), 339.

5 RUTISHAUSER, H. Solution of eigenvalue problems with the L-R transformation. In National Bureau of Standards Applied Math Series 49 (1958).

6. TURING, A. M. Rounding-off errors in matrix processes. *Quart. J. Mech. Appl. Math. 1* (1948), 287.

7 WILKINSON, J. H. Rounding errors in algebraic processes. Proceedings of International Conference on Information Processing, UNESCO, 1959.

8. WILKINSON, J. H. Error analysis of floating-point computation. *Num Math 2* (1960), 319

9. FADDEEVA, V. N. Computational methods of linear algebra. (Translated by C. D. Benster.) New York: Dover; London. Constable, 1959.