ERROR CONTROL IN MATRIX CALCULATION

F. E. Satterthwaite

Aetna Life Insurance Company

1. Introduction. The solutions of large sets of simultaneous equations and the inversion of matrices are often complicated by the fact that errors, such as those introduced by rounding, become magnified in the course of the calculations to such an extent that the results are useless. In this paper we shall show that if the norm of the matrix $A - I$ is less than 0.35, operations involving the inversion A or the multiplication by A^{-1} will be in a state of error control for "Doolittle" methods of calculation. Thus such calculations may be carried through with assurance that the errors in the results will be limited to two or three significant figures. We also point out that as soon as an approximation to A^{-1} is available, most problems may be restated to bring them within the requirements for error control. Therefore the solution can be immediately completed to the desired degree of accuracy in one step instead of requiring multiple steps as do the iterative methods.

2. The inversion of special matrices. Consider the problem of inverting the matrix $(I + F)$ where I is the identity matrix and $(I + F)$ is a non-singular square matrix. Let

$$G = (I + F)^{-1}. \tag{2.1}$$

Then

$$(I + F)G = I \tag{2.2}$$

or

$$G = I - FG. \tag{2.3}$$

In ordinary algebra this would not be a practical formula for the calculation of G. However in matrix algebra the situation may be different. Examine the expanded form of G:

$$g_{ij} = \delta_{ij} - \Sigma f_{ik}g_{kj}. \tag{2.4}$$

The summation is over all values of k from 1 to n. Next examine the affect of imposing certain restrictions on F. For example, let $f_{ij} = 0$ if $j \geq i$. This is equivalent to making the summation in (2.4) over the range 1 to $i - 1$. The first row of (2.4) then becomes

$$g_{ij} = \delta_{ij}$$

and no g's appear on the right. For the second row

$$g_{2j} = \delta_{2j} - f_{21}g_{1j}.$$

The only g's on the right are those on the first row which have already been calculated. For the third row

$$g_{3j} = \delta_{3j} - f_{31}g_{1j} - f_{32}g_{2j}\,.$$

The only g's on the right are in the first and second rows and have already been calculated. Similarly for the fourth and later rows.

Thus it is seen that if F is a "pre-diagonal" matrix, (2.3) is a very simple and practical formula for the numerical calculation of the inverse of $(I + F)$. Also if F is a post-diagonal matrix, (2.3) may be used by working up from the bottom row.

Similarly, if a matrix H is to be multiplied by the inverse of $(I + F)$, let

$$G = (I + F)^{-1}H \tag{2.5}$$

and the working equation becomes

$$G = H - FG. \tag{2.6}$$

The inversion of a diagonal matrix is accomplished by inverting each of its diagonal elements. That is if

$$F = \| \delta_{ij}s_{ii} \| \tag{2.7}$$

then

$$F^{-1} = \| \delta_{ij}s_{ii}^{-1} \|. \tag{2.8}$$

3. The inversion of general matrices. The general inversion problem will be solved if a general matrix can be factored into matrices of the special types treated in the last section. For the moment assume that such a factorization is possible and let the factors of the general matrix, A, be

$$A = (R_1 + I)S_1(I + T_1) \tag{3.1}$$

where R_1 is a prediagonal matrix, S_1 a diagonal matrix, and T_1 a postdiagonal matrix. Then

$$A = S_1 + R_1S_1 + S_1T_1 + R_1S_1T_1\,. \tag{3.2}$$

A slight change in form now appears desirable so let

$$\begin{aligned} A &= (RS^{-1} + I)S(I + S^{-1}T) \\ &= R + S + T + RS^{-1}SS^{-1}T \end{aligned} \tag{3.3}$$

$$= R + S + T + RS^{-1}T. \tag{3.4}$$

For convenience let

$$B = R + S + T \tag{3.5}$$

and remember that R, S, and T have no non-zero elements in common. Therefore the non-zero elements of R, S, and T are equal to the corresponding elements of B. Rearranging (3.4) gives

$$B = A - RS^{-1}T \tag{3.6}$$

and the elements of B are determined by

$$b_{ij} = a_{ij} - \Sigma \frac{r_{ik} t_{kj}}{s_{kk}}. \tag{3.7}$$

Since $r_{ik} = 0$ for $k \geq i$, there is no point in making the summation beyond $k = i - 1$. Also since $t_{kj} = 0$ for $k \geq j$, there is no point in making the summation beyond $j - 1$. Therefore the summation in (3.7) is to be considered to be over the range 1 to the smaller of $i - 1$ and $j - 1$. The r's, s's and t's on the right of (3.7) can now be replaced by the corresponding b's:

$$b_{ij} = a_{ij} - \Sigma \frac{b_{ik} b_{kj}}{b_{kk}}. \tag{3.8}$$

Since the first row (column) of b's equal the first row (column) of a's, the second row (column) of b's is a function of only those b's in the first row and the first column, etc., any calculation routine which works down from the top and from the left to the right will lead to a ready determination of all the b's by (3.8).

Thus we see that the assumed factorization (3.3) of A is always possible (unless some of the diagonal elements, b_{kk}, of B are zero) and moreover the elements of the factors are readily calculated by the simple equations, (3.8).

Therefore, to invert the general non-singular square matrix A, calculate the elements of an intermediate matrix $B = R + S + T$ by equations (3.5) and (3.7). Then from (3.3) we have

$$A^{-1} = (I + S^{-1}T)^{-1} S^{-1}(I + RS^{-1})^{-1} \tag{3.9}$$

which can be readily calculated by the methods of (2.3) and (2.6).

4. The Doolittle method. The Doolittle method of matrix calculation can now be expressed in terms of the matrices R, S, and T studied above. To illustrate we shall use the set of equations:

$$\begin{aligned} a_{11}x_1 + a_{12}x_2 + a_{13}x_3 &= c_{11}y_1 + c_{12}y_2 + c_{13}y_3 = d_1, \\ a_{21}x_1 + a_{22}x_2 + a_{23}x_3 &= c_{21}y_1 + c_{22}y_2 + c_{23}y_3 = d_2, \\ a_{31}x_1 + a_{32}x_2 + a_{33}x_3 &= c_{31}y_1 + c_{32}y_2 + c_{33}y_3 = d_3. \end{aligned} \tag{4.1}$$

This set of equations will be represented in the form of a three element matrix,

$$|| AX : CY : D ||. \tag{4.1.a}$$

The essential feature of the Doolittle method of solution is that we replace (4.1) by an equivalent set of equations for which the prediagonal coefficients of the X's are all zero and the diagonal coefficients are all unity. Therefore consider the set formed as follows:

$$(4.2)\quad \| A_3X : C_3Y : D_3 \| = S^{-1}\{\| AX : CY : D \| - R \| A_3Y : C_3Y : D_3 \|\}.$$

Then

$$(4.3)\qquad A_3 = S^{-1}\{A - RA_3\}$$

or

$$(4.4)\qquad (S + R)A_3 = A$$

or

$$(4.5)\qquad \begin{aligned} A_3 &= (S + R)^{-1}(RS^{-1} + I)S(I + S^{-1}T) && \text{by (3.3)} \\ &= (I + S^{-1}T). \end{aligned}$$

Since $S^{-1}T$ is a postdiagonal matrix, $\| A_3X : C_3Y : D_3 \|$ are the required intermediate equations for a Doolittle type of solution.

The final solution is now easily obtained. Consider

$$(4.6)\quad \| A_4X : C_4Y : D_4 \| = \| A_3X : C_3Y : D_3 \| - (S^{-1}T) \| A_4X : C_4Y : D_4 \|.$$

We have

$$(4.7)\qquad A_4 = A_3 - (S^{-1}T)A_4$$

or

$$(4.8)\qquad \begin{aligned} (I + S^{-1}T)A_4 &= A_3 \\ &= I + S^{-1}T && \text{by (4.5).} \end{aligned}$$

Therefore A_4 is in fact the identity matrix and (4.6) can be rewritten

$$(4.9)\quad \| X : C_4Y : D_4 \| = \| A_3X : C_3Y : D_3 \| - (S^{-1}T) \| X : C_4Y : D_4 \|$$

5. The non-symmetric case. In actual practice, the work has to be so arranged that the elements of the matrices R, S, and $(S^{-1}T)$ are set out so as to be readily available for use as multipliers in forming the intermediate and final sets of equations. Table I gives such a practical layout for the non-symmetric case.

The elements of $(S^{-1}T)$ are set out as the postdiagonal elements of A_3 so that they do not need further attention. To determine the elements of R and S, we form a set of pre-intermediate equations:

$$(5.1)\quad \| BX : \cdots : \cdots \| = \| AX : \cdots : \cdots \| - R \| A_3X - X : \cdots : \cdots \|.$$

TABLE I

Layout of Doolittle solution for the non-symmetric case. Coefficients not used further are indicated by ⋯ .

$$
\begin{array}{rrrcrrrcl}
a_{11}x_1 & + a_{12}x_2 & + a_{13}x_3 & = & c_{11}y_1 & + c_{12}y_2 & + c_{13}y_3 & = & d_1 \\
a_{21}x_1 & + a_{22}x_2 & + a_{23}x_3 & = & c_{21}y_1 & + c_{22}y_2 & + c_{23}y_3 & = & d_2 \\
a_{31}x_1 & + a_{32}x_2 & + a_{33}x_3 & = & c_{31}y_1 & + c_{32}y_2 & + c_{33}y_3 & = & d_3
\end{array}
$$

$$
\begin{array}{rrrcrrrcl}
s_1x_1 & + \cdots & \cdots & = & \cdots & \cdots & \cdots & = & d_{2:1} = d_1 \\
x_1 & + a_{3:12}x_2 & + a_{3:13}x_3 & = & c_{3:11}y_1 & + c_{3:12}y_2 & + c_{3:13}y_3 & = & d_{3:1} = d_1/s_1 \\
r_{21}x_1 & + s_2x_2 & + \cdots & = & \cdots & \cdots & \cdots & = & d_{2:2} = d_2 - r_{21}[d_{3:1} - x_1] \\
 & x_2 & + a_{3:23}x_3 & = & c_{3:21}y_1 & + c_{3:22}y_2 & + c_{3:23}y_3 & = & d_{3:2} = [d_2 - r_{21}d_{3:1}]/s_2 \\
r_{31}x_1 & + r_{32}x_2 & + s_3x_3 & = & \cdots & \cdots & \cdots & = & d_{2:3} = d_3 - r_{31}[d_{3:1} - x_1] - r_{32}[d_{3:2} - x_2] \\
 & & x_3 & = & c_{3:31}y_1 & + c_{3:32}y_2 & + c_{3:33}y_3 & = & d_{3:3} = [d_3 - r_{31}d_{3:1} - r_{32}d_{3:2}]/s_3
\end{array}
$$

$$
\begin{array}{rrrcrrrcl}
x_1 & & & = & c_{4:11}y_1 & + c_{4:12}y_2 & + c_{4:13}y_3 & = & d_{4:1} = d_{3:1} - a_{3:12}d_{3:2} - a_{3:13}d_{3:3} \\
 & x_2 & & = & c_{4:21}y_1 & + c_{4:22}y_2 & + c_{4:23}y_3 & = & d_{4:2} = d_{3:2} - a_{3:23}d_{3:3} \\
 & & x_3 & = & c_{4:31}y_1 & + c_{4:32}y_2 & + c_{4:33}y_3 & = & d_{4:3} = d_{3:3}
\end{array}
$$

TABLE II

Layout of Doolittle solution for the symmetric case. Coefficients not used further are indicated by ⋯ . Coefficients which can be filled by symmetry are indicated by —

$$
\begin{array}{rrrcrrrcl}
a_{11}x_1 & + a_{12}x_2 & + a_{13}x_3 & = & c_{11}y_1 & + c_{12}y_2 & + c_{13}y_3 & = & d_1 \\
— & + a_{22}x_2 & + a_{23}x_3 & = & c_{21}y_1 & + c_{22}y_2 & + c_{23}y_3 & = & d_2 \\
— & — & + a_{33}x_3 & = & c_{31}y_1 & + c_{32}y_2 & + c_{33}y_3 & = & d_3
\end{array}
$$

$$
\begin{array}{rrrcrrrcl}
s_1x_1 & + r_{21}x_2 & + r_{31}x_3 & = & \cdots & \cdots & \cdots & = & d_{2:1} = d_1 \\
x_1 & + a_{3:12}x_2 & + a_{3:13}x_3 & = & c_{3:11}y_1 & + c_{3:12}y_2 & + c_{3:13}y_3 & = & d_{3:1} = [d_1]/s_1 \\
 & s_2x_2 & + r_{32}x_3 & = & \cdots & \cdots & \cdots & = & d_{2:2} = d_2 - r_{21}d_{3:1} \\
 & x_2 & + a_{3:23}x_3 & = & \cdots & + c_{3:22}y_2 & + c_{3:23}y_3 & = & d_{3:2} = [d_2 - r_{21}d_{3:1}]/s_2 \\
 & & s_3x_3 & = & \cdots & \cdots & \cdots & = & d_{2:3} = d_3 - r_{31}d_{3:1} - r_{32}d_{3:2} \\
 & & x_3 & = & \cdots & \cdots & + c_{3:33}y_3 & = & d_{3:3} = [d_3 - r_{31}d_{3:1} - r_{32}d_{3:2}]/s_3
\end{array}
$$

$$
\begin{array}{rrrcrrrcl}
x_1 & & & = & c_{4:11}y_1 & + c_{4:12}y_2 & + c_{4:13}y_3 & = & d_{4:1} = d_{3:1} - a_{3:12}d_{4:2} - a_{3:12}d_{4:3} \\
 & x_2 & & = & — & + c_{4:22}y_2 & + c_{4:23}y_3 & = & d_{4:2} = d_{3:2} - a_{3:23}d_{4:3} \\
 & & x_3 & = & — & — & + c_{4:33}y_3 & = & d_{4:3} = d_{3:3}
\end{array}
$$

Then

$$\begin{aligned} B &= A - R[A_3 - I] \\ &= A - R[(I + S^{-1}T) - I) && \text{by (4.5),} \\ &= A - RS^{-1}T \\ &= R + S + T. && \text{by (3.5) and (3.6).} \end{aligned} \tag{5.2}$$

Therefore we see that the prediagonal coefficients of the x's in (5.1) are the elements of R and that the diagonal coefficients are the elements of S. The rest of the coefficients in this set of equations are not needed in the calculations and have been indicated by dots in Table I.

6. The symmetric case. If the A matrix is symmetric, advantage can be taken of the fact that the B matrix is also symmetric. Therefore

$$S + T = S + R' \tag{6.1}$$

and the elements of S and R' can be written down just before the division by s_{ii} in the calculation of the A_3 matrix:

$$\begin{aligned} A_3 &= (I + S^{-1}T) = S^{-1}(S + T) && \text{by (4.5)} \\ &= S^{-1}(S + R'). \end{aligned} \tag{6.2}$$

The layout of the work is given in Table II.

If the C_4 matrix is symmetric, the prediagonal elements of C_4 can be entered by symmetry. Therefore it is not only unnecessary to calculate the prediagonal elements of C_4, but we can also omit the prediagonal elements of C_3. Note that in this case C_4 must be calculated from the right to the left as well as from the bottom up.

The most important case where it is known in advance that C_4 is symmetric is the determination of the inverse of a symmetric matrix. Then $C = I$ and $C_4 = A^{-1}$. Also the postdiagonal elements of C_3 are all zero so that the only elements of C_3 which have to be calculated are the diagonal elements. These are the reciprocals of the s_{ii}'s.

A case where C_4 is symmetric though $C \neq I$ will appear in a subsequent paper.

7. Norms. In order to state the conditions for error control in a matrix calculation, a concept of the norm or the absolute value of a matrix is necessary. In this paper the norm will be defined as the square root of the sum of the squares of the elements of the matrix. That is

$$N(F) = \sqrt{\Sigma_i \Sigma_j (f_{ij})^2}. \tag{7.1}$$

The two basic inequalities satisfied by the norm are

$$N(F + G) \leq N(F) + N(G) \tag{7.2}$$

and

$$N(FG) \leq N(F)N(G). \tag{7.3}$$

All other properties of the norm are derived from these.

For future reference we list the following norm relations:

$$\begin{aligned} N[(F)(I + G)] &= N(F + FG) \\ &\leq N(F) + N(F)N(G) \\ &\leq N(F)[1 + N(G)]. \end{aligned} \tag{7.4}$$

If $N(F) < 1$ we have

$$\begin{aligned} N[(I + F)^{-1} - I] &= N[I - F + F^2 - \cdots - I] \\ &\leq N(F) + [N(F)]^2 + [N(F)]^3 + \cdots \\ &\leq \frac{N(F)}{1 - N(F)}. \end{aligned} \tag{7.5}$$

If $N(G - I) < 1$, (7.5) becomes

$$1 + N(G^{-1} - I) \leq \frac{1}{1 - N(G - I)}. \tag{7.6}$$

When $N(F - I) < 1$

$$\begin{aligned} N(F^{-1}G) &\leq N[I + (F^{-1} - I)][G] \\ &\leq [1 + N(F^{-1} - I)]\,N(G) \\ &\leq \frac{N(G)}{1 - N(F - I)} \qquad \text{by (7.6)} \end{aligned} \tag{7.7}$$

8. Error matrix and error norm. We shall also need a formal statement as to what we mean by error and we need a measure of the errors.

By an error matrix we mean the matrix whose elements consist of the differences between the value of the matrix elements as actually calculated and the true value of the matrix elements which would have been obtained if all calculations had been made exactly without any rounding or other approximations. The fundamental relation for the error matrix, $E[f(G)]$, of a function, $f(G)$, of G is

$$E[f(G)] = f[G + E(G)] - f(G). \tag{8.1}$$

If each element of a matrix is calculated to q decimal places and the matrix has p rows and p columns, the maximum rounding error introduced in any element is $.5 \times 10^{-q}$. The norm of the error introduced by rounding is less than

$$\begin{aligned} NE_1 &= \sqrt{p^2[.5 \times 10^{-q}]^2} \\ &= .5 \times 10^{-q} \times p. \end{aligned} \tag{8.2}$$

For triangular matrices

$$NE_2 = .5 \times 10^{-q}\sqrt{p(p+1)/2}. \tag{8.3}$$

For one column matrices

$$NE_3 = .5 \times 10^{-q}\sqrt{p}. \tag{8.4}$$

For future reference the following formulas for error norms are listed:

$$NE(F+G) \leq NE(F) + NE(G). \tag{8.5}$$

$$\begin{aligned} NE(FG) &\leq NE(F)N[G + E(G)] + NE(G)N(F) \\ &\leq NE(F)N(G) + NE(G)N(F) + NE(F)NE(G). \end{aligned} \tag{8.6}$$

$$\begin{aligned} NE[F(I+G)] &\leq NE(F)[1 + N[G + E(G)] + NE(G)N(F) \\ &\leq NE(F) + NE(F)N(G) + NE(G)N(F) \\ &\qquad + NE(F)NE(G). \end{aligned} \tag{8.7}$$

If $N(F-I) + NE(F) < 1$,

$$\begin{aligned} NE(F^{-1}) &= NE[I + (F-I)]^{-1} \\ &= N\left[\frac{1}{I + [F + E(F) - I]} - \frac{1}{I + (F-I)}\right] \\ &= N\{I + [F + E(F) - I]\}^{-1}\{I + (F-I)\}^{-1}\{E(F)\} \\ &\leq \frac{NE(F)}{\{1 - N[F + E(F) - I]\}\{1 - N(F-I)\}} \qquad \text{by (7.7)}, \\ &\leq \frac{NE(F)}{[1 - N(F-I) - NE(F)][1 - N(F-I)]}. \end{aligned} \tag{8.8}$$

If $N(F-I) + NE(F) < 1$,

$$\begin{aligned} NE(F^{-1}G) &= N\left[\frac{G + E(G)}{I + [F + E(F) - I]} - \frac{G}{I + (F-I)}\right] \\ &= N\left[\frac{[E(G)][I + (F-I)] - [E(F)][G]}{\{I + [F + E(F) - I]\}\{I + (F-I)\}}\right] \\ &\leq \frac{NE(G) + \{[NE(F)]/[1 - N(F-I)]\}\{N(G)\}}{1 - N[F + E(F) - I]} \qquad \text{by (7.7)}. \end{aligned} \tag{8.9}$$

9. Certain maxima. The R, S, and T matrices have no non-zero elements in common so that

$$\begin{aligned} N(B-I) &= N(R + S - I + T) \\ &= \sqrt{[N(R)]^2 + [N(S-I)]^2 + [N(T)]^2}. \end{aligned} \tag{9.1}$$

Similarly

$$NE(B) = \sqrt{[NE(R)]^2 + [NE(S)]^2 + [NE(T)]^2}. \tag{9.2}$$

The developments of the following formulas for certain maxima are not given here since they involve only well known calculus principles. It is understood that these inequalities hold whenever the quantities involved exist. Usually the maxima are given subject to the condition $N(T) = N(R)$ as well as for the unrestricted case where $N(T)$ may be zero.

$$\max \frac{N(R)N(T)}{1 - N(S - I)} = 1 - \sqrt{1 - [N(B - I)]^2}. \tag{9.3}$$

$$\max \frac{N(R)}{1 - N(S - I)} = \frac{N(B - I)}{\{1 - [N(B - I)]^2\}^{1/2}} \quad \text{if } N(T) = 0, \tag{9.4}$$

$$= \frac{N(B - I)}{\sqrt{2}\sqrt{1 - [N(B - I)]^2}} \quad \text{if } N(T) = N(R). \tag{9.5}$$

$$\max [N(R)NE(T) + N(T)NE(R)] = \sqrt{[N(B - I)]^2 - [N(S - I)]^2} \times \sqrt{[NE(B)]^2 - [NE(S)]^2}. \tag{9.6}$$

Any substitution satisfying the relation

$$\frac{NE(R)}{NE(T)} = \frac{N(T)}{N(R)} \tag{9.7}$$

will cause (9.6) to attain its maximum.

$$\max[NE(R) + kNE(S)] = \sqrt{\tfrac{1}{2} + k^2}\, NE(B) \quad \text{if } NE(R) = NE(T), \tag{9.8}$$

$$= \sqrt{1 + k^2}\, NE(B) \quad \text{if } NE(T) = 0. \tag{9.8.a}$$

$$\max \frac{k + N(R)}{1 - N(S - I)} = k + \frac{[N(B - I)]^2}{K} \tag{9.9}$$

where K is the root of

$$[1 + k^2]K^2 + 2[k\{N(B - \mathrm{I})\}^2]K + \{[N(B - I)]^4 - [N(B - I)]^2\} = 0. \tag{9.10}$$

10. Errors in the Doolittle method. In all that follows, we shall assume that $N(A - I)$ is small enough so that the "divisions" are permissible. First let us examine in the multipliers $B = R + S + T$. By (3.6)

$$B - I = (A - I) - RS^{-1}T. \tag{10.1}$$

Therefore

$$N(B - I) \leq N(A - I) + \frac{N(R)N(T)}{1 - N(S - 1)} \quad \text{by (7.3) and (7.7),}$$

$$\leq N(A - I) + 1 - \sqrt{1 - [N(B - I)]^2} \quad \text{by (9.3)} \tag{10.2}$$

Remembering that A has no error and letting NE_1 be the rounding error norm introduced in writing down the elements, we have

$$NE(B) \leq NE_1 + \frac{NE(R)N(T) - NE(T)N(R) + NE(R)NE(T)}{1 - N[S + E(S) - I]} + \frac{NE(S)N(T)N(R)}{\{1 - N[S + E(S) - I]\}\{1 - N(S - I)\}} \quad \text{by (8.9) and (8.6).} \tag{10.3}$$

We are interested only in the range of values where the errors are small. Therefore we shall ignore second order errors. Except for such errors,

$$NE(B) \leq NE_1 + \frac{N(T)}{1 - N(S - I)} NE(R) + \frac{N(R)}{1 - N(S - I)} NE(T) + \frac{N(T)N(R)}{[1 - N(S - I)]^2} NE(S). \tag{10.4}$$

The last term will be largest when $N(T) = N(R)$. Therefore the sum of the second and third will be largest when $NE(R) = NE(T)$ by (9.7).

$$NE(B) \leq NE_1 + 2\frac{N(R)}{1 - N(S - I)} NE(R) + \left[\frac{N(R)}{1 - N(S - I)}\right]^2 NE(S). \tag{10.5}$$

By (9.8) we now obtain

$$\begin{aligned} NE(B) &\leq NE_1 + K\sqrt{2 + K^2}\, NE(B) \\ &\leq \frac{NE_1}{1 - K\sqrt{2 + K^2}} \end{aligned} \tag{10.6}$$

where

$$K = \max \frac{N(\dot{R})}{1 - N(S - I)} = \frac{N(B - I)}{\sqrt{2}\sqrt{1 - [N(B - I)]^2}} \quad \text{by (9.5).}$$

Actually in practice we introduce the rounding error in $(S^{-1}T)$ instead of in T as assumed above. Our assumption is conservative since the division by S magnifies any error in T.

Next consider the errors in the C_3 (or D_3) matrix. From (4.2)

$$\begin{aligned} C_3 &= S^{-1}(C - RC_3) \\ &= S^{-1}C - S^{-1}RC_3 \\ &= (I + S^{-1}R)^{-1}S^{-1}C \qquad \text{by (2.5) and (2.6),} \\ &= [S(I + S^{-1}R)]^{-1}C \end{aligned} \tag{10.7}$$

$$= (S + R)^{-1}C. \tag{10.8}$$

Therefore

$$N(C_3) \leq \frac{N(C)}{1 - N(R + S - I)} \quad \text{by (7.7),} \tag{10.9}$$

$$\leq \frac{N(C)}{1 - N(R) - N(S - I)}. \tag{10.10}$$

From (10.7), remembering that C has no error, and letting the rounding error be NE_3,

$$\begin{aligned} NE(C_3) &\leq NE_3 + \frac{NE(R)N(C_3) + NE(C_3)N[R + E(R)]}{1 - N[S + E(S) - I]} \\ &\quad + \frac{NE(S)[N(C) + N(R)N(C_3)]}{\{1 - N[S + E(S) - I]\}\{1 - N(S - I)\}} \quad \text{by (7.7) and (7.3),} \\ &\leq \frac{\dfrac{NE(R) + NE(S)}{1 - N(R) - N(S - I)} N(C) + \{NE_3\}\{1 - N[S + E(S) - I)\}}{1 - N[R + E(R)] - N[S + E(S) - I]} \end{aligned} \tag{10.11}$$

since

$$\begin{aligned} &\frac{N(C) + N(R)N(C_3)}{1 - N(S - I)} \\ &\quad \leq \frac{N(C)}{1 - N(S - I)}\left[1 + \frac{N(R)}{1 - N(R) - N(S - I)}\right] \quad \text{by (10.10),} \\ &\quad \leq \frac{N(C)[1 - N(S - 1)]}{[1 - N(S - I)][1 - N(R) - N(S - I)]} \end{aligned} \tag{10.12}$$

and since transferring the terms in $NE(C_3)$ to the left requires that we divide through by

$$1 - \frac{N[R + E(R)]}{1 - N[S + E(S) + I]} = \frac{1 - N[R + E(R)] - N[S + E(S) - I]}{1 - N[S + E(S) - I]}. \tag{10.13}$$

Again we can ignore second order errors. Taking maxima by (9.8.a) gives

$$NE(C_3) \leq \frac{\dfrac{\sqrt{2}\, NE(B)N(C)}{1 - \sqrt{2}\, N(B - I)} + NE_3}{1 - \sqrt{2}\, N(B - I)} \tag{10.14}$$

and

$$\frac{NE(C_3)}{N(C)} \leq \frac{\sqrt{2}\, NE(B)}{[1 - \sqrt{2}\, N(B - I)]^2} + \frac{[NE_3]/[N(C)]}{1 - \sqrt{2}\, N(B - I)}. \tag{10.15}$$

Thus we see that the proportionate error in C_3 is made up of two parts: the first due to the rounding errors in the multipliers as given by the first term and the second due to the proportionate rounding error introduced in calculating C_3.

Finally we have the errors in the C_4 (or D_4) matrix. Since

$$C_4 = A^{-1}C \tag{10.16}$$

we have

$$N(C_4) \leq \frac{N(C)}{1 - N(A - I)}. \tag{10.17}$$

By (4.6)

$$C_4 = C_3 - (S^{-1}T)C_4. \tag{10.18}$$

If we let NE_4 be the rounding error introduced in this step

$$\begin{aligned} NE(C_4) &\leq NE(C_3) + NE_4 + \frac{NE(T)N(C_4) + NE(C_4)N[T + E(T)]}{1 - N[S + E(S) - I]} \\ &\quad + \frac{NE(S)N(T)N(C_4)}{\{1 - N[S + E(S) - I]\}\{1 - N(S - I)\}} \\ &\leq \frac{\{NE(C_3) + NE_4\}\{1 - N[S + E(S) - I]\} + \left[NE(T) + \frac{N(T)NE(S)}{1 - N(S - I)}\right]\left[\frac{N(C)}{1 - N(A - I)}\right]}{1 - N[T + E(T)] - N[S + E(S) - I]} \end{aligned} \tag{10.19}$$

by (10.17) and relations similar to (10.13). We now ignore second order errors and take maxima by (9.4) and (9.8.a):

$$\begin{aligned} \frac{NE(C_4)}{N(C)} &\leq \frac{\frac{NE(C_3)}{N(C)} + \frac{(NE_4)}{N(C)} + \frac{\sqrt{1 + K^2}\,NE(B)}{1 - N(A - I)}}{1 - \sqrt{2}\,N(B - I)} \\ &\quad \left[\frac{1}{[1-N(A-I)]\sqrt{1-[N(B-I)]^2}} + \frac{\sqrt{2}}{[1-\sqrt{2}\,N(B-I)]^2}\right] NE(B) \\ &\leq \frac{+ \frac{[NE_3]/[N(C)]}{1 - \sqrt{2}\,N(B-I)} + \frac{NE_4}{N(C)}}{1 - \sqrt{2}\,N(B - I)} \end{aligned} \tag{10.20}$$

where

$$K = \max \frac{N(T)}{1 - N(S - I)} = \frac{N(B - I)}{\sqrt{1 - [N(B - I)]^2}} \quad \text{by (9.4)}, \tag{10.21}$$

and

$$\sqrt{1 + K^2} = 1/\sqrt{1 - [N(B - I)]^2}. \tag{10.22}$$

If A is symmetric, $NE(B - I)$ remains unchanged but $NE(C_4)$ can be somewhat strengthened through the use of (9.5) and (9.8) instead of (9.4) and (9.8.a).

The result is

$$\frac{NE(C_4)}{N(C)} \leq \left[\frac{1}{[1 - N(A - I)]\sqrt{2}\sqrt{1 - [N(B - I)]^2}} + \frac{\sqrt{3/2}}{[1 - \sqrt{3/2}\, N(B - I)]^2}\right]\left[\frac{NE(B)}{1 - \sqrt{3/2}\, N(B - I)}\right] + \frac{1}{[1 - \sqrt{3/2}\, N(B - I)]^2}\frac{NE_3}{N(C)} + \frac{1}{1 - \sqrt{3/2}\, N(B - I)}\frac{NE_4}{N(C)}. \tag{10.23}$$

If C is approximately equal to I, a better formula is obtained by substituting $(I + C_1)$ for C. The final formulas then are identical to (10.20) and (10.23) with the substitution of $1 + N(C_1)$ for $N(C)$. Similarly if $C = I$ so that $C_4 = A^{-1}$, $N(C) = 1$ should be substituted in (10.20) and (10.23).

If A is symmetric and if $C = I$ so that $C_4 = A^{-1}$, the prediagonal elements of C_4 are filled in by symmetry as in Table II instead of being calculated directly. This complicates the analysis of error relations. The following inequality gives the error limit for the diagonal and postdiagonal elements of (A^{-1}). We have indicated these elements by F.

$$NE(F) \leq \frac{\sqrt{[3/2] - N(A - I) + [N(B - I)]^2/K}}{[1 - N(A - I)]\,[1 - \sqrt{2}\, N(B - I)]} NE(B) + \frac{NE_5}{1 - \sqrt{2}\, N(B - I)} \tag{10.24}$$

where K is the root of (9.10) when $k = 1 - N(A - I)$ and $NE_5 = 0.5 \times 10^{-q} \times \sqrt{p(p + 1)/2}$ by (8.3).

11. Results. Given the matrix A, we subtract one from each element on the principal diagonal to obtain the matrix $(A - I)$. By the norm of $(A - I)$ we mean the square root of the sum of the squares of the elements of $(A - I)$. We shall now show that a Doolittle process such as outlined in Table I is in a state of error control if the norm of $(A - I)$ is less than 0.35:

1. $N(A - I) \leq 0.35$
2. $N(B - I) \leq 0.4642$ by (10.2).
3. $NE(B) \leq 1.09\ p$ by (10.6) and (8.2)

if the maximum rounding error in any element is 0.5 and A has p rows. Thus no element of the multiplying matrices R, S, or T can have an error of greater than this amount.

4. $NE(C_4) \leq (44 \times p \times 10^{-q})N(C) + (6 \times p \times 10^{-r})$ by (10.20) and (8.2) where q decimal places are carried on the left and r on the right. Thus if the decimal point in C is shifted so that $N(C) \leq 1$, the error in any element of C_4 can

not amount to more than three significant figures if the same number of decimal places are carried on both the left and on the right (four significant figures for 21 to 200 rows).

5. $NE(A^{-1}) \leq$ three significant figures by substituting $N(C) = 1$ since $C = I$.

As we let $N(A - I)$ become larger than 0.35, the maximum errors indicated by our formulas rapidly become very large. In fact they become infinite if $N(A - I) = 0.414$.

Since for more than four equations the above formulas show errors in the second decimal place no matter how small $N(A - I)$ is, it is suggested that as a general practice:

1. The problem be arranged so that $N(A - I) \leq 0.35$.
2. The decimal points in C be shifted so $N(C) < 1$.
3. Three extra decimal places be carried in the calculations.

12. Preliminary adjustments. The requirement that $N(A - I)$ should be less than 0.35 is not normally met in practical problems. If, however, an approximation to A^{-1} is available, the problem can almost always be rearranged to satisfy this condition.

Thus if we are solving the equations such as given in Table I,

$$AX = CY = D, \tag{12.1}$$

we are perfectly free to multiply through by any non-singular matrix F without disturbing the solution:

$$(FA)X = (FC)Y = (FD). \tag{12.2}$$

Now if F is a sufficiently close approximation to A^{-1}, FA will be almost equal to I. Therefore $N(FA - I)$ will be less than 0.35 and a Doolittle solution of (12.2) will be in a state of error control.

Similarly for the inversion of A, we can apply the Doolittle process to the pair of matrices

$$FA : F$$

just as easily as to the pair

$$A : I$$

since

$$\begin{aligned}(FA)^{-1} \,||\, FA : F \,||\, &= A^{-1}F^{-1} \,||\, FA : F \,||\\ &= ||\, A^{-1}(F^{-1}F)A : A^{-1}(F^{-1}F) \,||\\ &= ||\, A^{-1}A : A^{-1} \,||\\ &= ||\, I : A^{-1} \,||.\end{aligned}$$

Thus by taking F as a sufficiently close approximation to A^{-1}, we can bring an inversion calculation into a state of error control.

The computor should be cautioned that the multiplication by F must be exact and that no rounding is allowable in this step. Our formulas assumed that we started with matrices free of error.

13. Further work. The principles used in this paper can be applied to the task of developing calculation routines which will be in a state of error control regardless of the size of $N(A - I)$. Enough work has been done to see that such routines do exist and do not involve prohibitive labor. The author expects that the most efficient routine will be to use these more elaborate methods to obtain an F such that $N(FA - I) \leq 0.35$ and then to use the normal Doolittle methods as outlined in section 12.

REFERENCES

The writer wishes to acknowledge particularly his debt to Prof. Hotelling's very complete paper on interative methods and to Prof. Dwyer's papers on the Doolittle methods. These papers have such complete bibliographies, we shall not give further references here.

[1] PAUL S. DWYER, "The Doolittle technique," *Ann. Math. Stat.*, Vol. 12 (1941), pp. 449–458.

[2] PAUL S. DWYER, "The solution of simultaneous equations," *Psychometrika*, Vol. 6 (1941), pp. 101–129.

[3] HAROLD HOTELLING, "Some new methods in matrix calculation," *Ann. Math. Stat.*, Vol. 14 (1943), pp. 1–34.