EXAMPLES OF SOLUTION ACCURACY IN CERTAIN LARGE SIMULTANEOUS EQUATION SYSTEMS

B. E. Gatewood*
The Ohio State University
Columbus, Ohio

Norik Ohanian**
North American Aviation, Inc.
Columbus Division, Ohio

It is demonstrated that for a fixed number of decimal places in the calculations the differential order of the system, the sequencing of the computations, and the relative magnitude of the coefficients are the primary sources of error in the solution of systems of simultaneous equations. Accuracy can be improved by reducing the order of the system, by keeping the multiplying factors small in the computation sequence, and by using realistic coefficients. The extended Choleski method using submatrices is derived and shown to have very little error accumulation in the calculation sequence. As in all methods it is affected by the differential order of the system, losing six places in a beam deflection problem for both 100 fourth order equations and 10,000 second order equations. The transfer matrix method is described and shown to be very accurate for a true first order system. It lost four places for 100,000 equations in the beam deflection problem. Methods using forward or back substitution in the original system are shown to have serious error accumulation in the calculation sequence. For certain physical problems these substitution methods must fail regardless of the finite number of places used in the calculations.

INTRODUCTION

Every engineer at one time or another can expect to run into accuracy problems in solving simultaneous equations, inverting matrices, and calculating high order derivatives from test data. There has been much discussion of these accuracy problems in the literature but there appears to be little information on the precise reasons for the difficulties and how to avoid them. If everyone had been familiar with a statement made by a famous mathematician in 1947 (Reference 1) that it is possible to lose eight decimal places in inverting a 15 by 15 matrix, probably even less work would have been done on the problem of solving large systems of equations.

In this report an effort is made to identify the sources of the errors that occur in the solution of large systems and to demonstrate methods of solution that keep these errors to a minimum. Three primary sources of error, (1) the differential order of the system, (2) the sequencing of the computations, and (3) the relative magnitude of the coefficients, are considered. Also, three methods of solution, (1) the extended Choleski method using submatrices, (2) the back substitution method, and (3) the transfer matrix method, are discussed. No effort has been made to consider all error sources, to determine exact magnitude of possible errors, or to discuss all methods of solution and matrix inversion. The discussion and examples are presented to provide some guidance to those working with large systems of simultaneous equations.

*Professor, Aeronautical and Astronautical Engineering.
**Specialist, Research Structural Development Group, Columbus.

ACCURACY REQUIREMENTS IN DIFFERENTIAL EQUATION SOLUTIONS

Consider the differential equations

$$\frac{d^4U}{dx^4} = p \tag{1}$$

$$\frac{d^2U}{dx^2} = q \tag{2}$$

$$\nabla^4 U = r \tag{3}$$

over intervals of unit length $x = 0.0$ to 1.0 and $y = 0.0$ to 1.0. Divide the intervals into N elements with $\Delta x = 1/N$, $\Delta y = 1/N$, and write the three equations in finite difference form

$$N^4 (U_{n-2} - 4U_{n-1} + 6U_n - 4U_{n+1} + U_{n+2}) = p_n \tag{4}$$

$$N^2 (U_{n-1} - 2U_n + U_{n+1}) = q_n \tag{5}$$

$$\begin{aligned} N^4 \Big[& U_{n,m-2} + (2U_{n-1,m-1} - 8U_{n,m-1} + 2U_{n+1,m-1}) \\ & + (U_{n-2,m} - 8U_{n-1,m} + 20U_{n,m} - 8U_{n+1,m} + U_{n+2,m}) \\ & + (2U_{n-1,m+1} - 8U_{n,m+1} + 2U_{n+1,m+1}) + U_{n,m+2} \Big] = r_{n,m} \end{aligned} \tag{6}$$

If the maximum error in U is $\pm U_e$, then the maximum possible errors in p, q, r in these three cases are

$$p_e = 16N^4 U_e \quad , \qquad N \text{ unknowns} \tag{7}$$

$$q_e = 4N^2 U_e \quad , \qquad N \text{ unknowns} \tag{8}$$

$$r_e = 64N^4 U_e \quad , \qquad N^2 \text{ unknowns} \tag{9}$$

If the permissible error in p, q, r is of the order of unity then Equations 7, 8 and 9 demonstrate the well known result that measured values cannot be differentiated several times to give reasonable derivatives. If $U_e = 0.01$ is the error in measuring U and p_e, q_e, $r_e = 1$, then Equations 7 and 9 give N between one and two while Equation 8 gives $N = 5$. For perfect computation of U in an eight-place machine with $U_e = 10^{-8}$, $N \approx 25$, 5000, 12, respectively, for Equations 7, 8, and 9. Thus it appears that the accuracy requirements in the solution

of differential equations by finite differences with a fixed finite number of places in the calculations depends upon the order of the equation, each order increase requiring approximately an order of magnitude increase in accuracy. Apparently, difficulties can be expected in fourth-order equations using eight-place machines. It will be demonstrated in a later section that one way to avoid part of this accuracy problem is to split the equation into lower-order equations.

There is a further accuracy problem involved in making the computations in solving the system of equations. This is the error accumulation arising in a long sequence of calculations.

ERROR ACCUMULATION IN SEQUENCE COMPUTATIONS

Consider the simple subtractions

$$x_3 = k_2 x_2 - k_1 x_1 , \qquad x_4 = k_2 x_3 - k_1 x_2 \tag{10}$$

where x_1, x_2, x_3 are of the same order of magnitude. Let

$$x_2 = x_1 + x_e , \qquad k_2 = k_1 + 1 ,$$

so that

$$x_3 = x_1 + k_2 x_e , \qquad x_4 = x_1 + (k_2^2 - k_2 + 1)x_e \tag{11}$$

If x_e is the error in x_2 and there is no error in x_1, then the error in x_3 is $k_2 x_e$ and in x_4 is $(k_2^2 - k_2 + 1)x_e$. For a sequence of N similar calculations such as in a recursion, the error is of the order of

$$\left.\begin{array}{lll} k_2 = 1 & : & x_{eN} = x_e \\ k_2 = 2 & : & x_{eN} = N x_e \\ k_2 = 3 & : & x_{eN} = (2^N - 1)x_e \\ k_2 > 3 & : & x_{eN} \approx (k_2)^{N-1} x_e \end{array}\right\} \tag{12}$$

It is evident the error accumulations can be quite large and rapid for large values of k_2. If $x_e = 10^{-8}$ and $k_2 = 10.0$, then $x_{eN} = 1.0$ after only nine calculations. It appears that when subtraction is involved in variables of the same relative size, the average multiplying factors must be less than two to keep the error accumulation small in a long sequence of calculations. As will be demonstrated in a later section the method of calculation has a big effect on this error accumulation.

It should be noted that difficulties may be expected if some of the multiplying factors are much larger than others and if the values change rapidly in magnitude in the sequence.

ERRORS IN SOLVING LARGE SYSTEMS OF SIMULTANEOUS EQUATIONS

Depending upon what the system of equations represents and upon how it is solved, it is possible for all the above factors that affect the accuracy of the calculations to be present. If a direct solution is made by inverting the matrix of coefficients, then depending upon the method of inversion errors in the inverse may be serious. However, obtaining the inverse in structural and finite difference equations does not appear to be as difficult as the conclusions of Von Neuman and Goldstine (Reference 1) that it is possible to lose eight decimal

places in inverting a 15 by 15 matrix, ten places in a 50 by 50 matrix, and twelve places in a 150 by 150 matrix.

Due to these error problems in direct solutions a large amount of work on indirect or iterative methods has been published in the literature; see for example Varga's book (Reference 2). However, these methods also have problems of convergence, divergence, and speed of convergence. In a large system the time for sufficient iterations to convergence may be much larger than for a direct solution. Speed up factors help but there is no way of knowing the optimum factors in a large system.

After consideration of all these factors it was decided to investigate in detail direct methods using partitioned matrices and taking advantage of the fact that the stiffness influence coefficient matrix in most structural problems as well as all matrices in finite differences can be arranged in a diagonal form with many zeros. The method described in the next section is a submatrix form of Choleski's method for elements of the matrix. See McMinn (Reference 3) for a description of Choleski's method. Fox (Reference 4) and Turing (Reference 5) favor Choleski's method as more accurate than other direct element methods.

The transfer matrix method as applied to nonhomogeneous equations is discussed in a later section. See Pestel (Reference 6) for use of transfer matrices in vibration problems.

DIRECT SOLUTION OF SIMULTANEOUS EQUATION SYSTEMS BY SUBMATRICES

Consider the matrix equation

$$\mathbf{S}\,\mathbf{U} = \mathbf{G} \tag{13}$$

where $\mathbf{S}$ is a (p x p) square matrix of known constants with $|\mathbf{S}| \neq 0$, $\mathbf{G}$ is a (p x q) matrix of known constants, and $\mathbf{U}$ is a (p x q) unknown matrix. If q = 1, Eq. (13) represents a system of simultaneous equations. In fact for each column in $\mathbf{G}$ the corresponding column in $\mathbf{U}$ gives a solution of a system of simultaneous equations. If the inverse $\mathbf{S}^{-1}$ is known then

$$\mathbf{U} = \mathbf{S}^{-1}\,\mathbf{G} \tag{14}$$

However, if $\mathbf{S}$ is large the calculation of $\mathbf{S}^{-1}$ may have many of the accuracy problems described above as well as storage and time problems on the computer. It appears that in many cases it may be preferable to partition $\mathbf{S}$ into submatrices and solve for $\mathbf{U}$ without using $\mathbf{S}^{-1}$.

Partition $\mathbf{S}$ into (M x M) submatrices $\mathbf{S}_{ij}$ with the properties

$$\begin{aligned} &\mathbf{S}_{ii} \text{ square}, \quad |\mathbf{S}_{ii}| \neq 0, \quad i = 1, 2, \ldots, M, \\ &\mathbf{S}_{i,\,i\pm j} = \mathbf{0}, \quad \text{for} \quad j > N, \; N < M, \; i = 1, 2, \ldots, M \end{aligned} \tag{15}$$

That is, all submatrices in $\mathbf{S}$ are zero outside of (2N+1) diagonals with N diagonals above the main diagonal and N diagonals below the main diagonal. If $\mathbf{S}$ has no zero submatrices then N = M-1. Partition $\mathbf{U}$ and $\mathbf{G}$ into a column of M submatrices $\mathbf{U}_i$ and $\mathbf{G}_i$, i = 1, 2, . . ., M, with the number of rows in $\mathbf{U}_i$ and $\mathbf{G}_i$ the same as in the corresponding $\mathbf{S}_{ii}$. Note that the $\mathbf{S}_{ii}$ do not have to be the same size, although in many practical problems they will be taken the same size. Now $\mathbf{S}$ can be expressed in the form

$$\mathbf{S} = \left[\mathbf{S}_{i,\,i-N} \cdots \mathbf{S}_{i,\,i-1} \; \mathbf{S}_{ii} \; \mathbf{S}_{i,\,i+1} \cdots \mathbf{S}_{i,\,i+N}\right]_{1}^{M} \tag{16}$$

The matrix $\mathbf{S}$ in Equation 16 can be expressed as the product of a lower triangular matrix $\mathbf{B}$ times an upper triangular matrix $\mathbf{C}$ with $\mathbf{B}$ and $\mathbf{C}$ partitioned in the same way as $\mathbf{S}$, or

$$S = BC, \quad B = \left[B_{i,i-N} \cdots B_{i,i-1} B_{ii} \right]_1^M$$

$$C = \left[I_{ii} \; C_{i,i+1} \cdots C_{i,i+N} \right]_1^M \tag{17}$$

Multiply B C and equate to the corresponding submatrices in S to get recursion formulas for the calulation of $B_{i,i-j}$ and $C_{i,i+j}$:

$$B_{i,i-j} = S_{i,i-j} - \sum_{k=1}^{N-j} B_{i,i-j-k} C_{i-j-k,i-j} ,$$

$$j = N, N-1, \ldots, 1, 0; \; j < i; \; k < (i-j)$$

$$C_{i,i+j} = B_{ii}^{-1} \left(S_{i,i+j} - \sum_{k=1}^{N-j} B_{i,i-k} C_{i-k,i+j} \right) , \tag{18}$$

$$j = 1, 2, \ldots, N; \; j \leq (M-i); \; k < i ,$$

$$i = 1, 2, 3, \ldots, M$$

If P is defined by

$$CU = P \tag{19}$$

where P is partitioned in a column of M submatrices P_i the same as U, then from Equations 13 and 17

$$BP = G; \quad P_i = B_{ii}^{-1} \left(G_i - \sum_{k=1}^{N} B_{i,i-k} P_{i-k} \right) ,$$

$$k < i; \; i = 1, 2, 3, \ldots, M \tag{20}$$

Note that this recursion for P_i can be carried out simultaneously with the recursion for $B_{i,i-j}$ and $C_{i,i+j}$ in Equation 18.

From Equations 17 and 19 the backsweep recursion for U_i is

$$U_i = P_i - \sum_{k=1}^{N} C_{i,i+k} U_{i+k} , \quad k \leq (M-i); \quad i = M, M-1, \ldots, 1. \tag{21}$$

Thus for any number of columns in G the solution for the matrix U can be obtained by a forward recursion calculating the submatrices $B_{i,i-j}$, $C_{i,i+j}$, P_i and a backward recursion calculating the submatrices U_i of U. The determinant of S can be calculated directly from the submatrices B_{ii}, as

$$|S| = |B||C| = |B_{11}| |B_{22}| \cdots |B_{MM}| |I| \tag{22}$$

The inverse of S can be obtained by the above procedure by taking $G = I$, whence $P = B^{-1}$ and $U = S^{-1}$. Since each column of S^{-1} is calculated independently of the others, the accuracy of S^{-1} is the same as that for one load column. Although this method is one of the shortest and most accurate for getting S^{-1} when S is large, it is evident that a system of simultaneous equations with one column for G can be solved much quicker than S^{-1} can be calculated.

The above procedure for the case of the submatrices as one by one, or elements, is credited to Choleski in References 3 and 14.

TRI-DIAGONAL MATRICES

A particular case of Equation 16 that arises in many structural problems and in finite difference equations is that of N = 1 giving S three diagonals in submatrices. For this case Equations 16 to 21 simplify to

$$\mathbf{SU} = \mathbf{G}, \quad \mathbf{S} = \left[\mathbf{S}_{i,i-1} \quad \mathbf{S}_{ii} \quad \mathbf{S}_{i,i+1} \right]_1^M \tag{23}$$

$$\mathbf{S} = \mathbf{BC}, \quad \mathbf{B} = \left[\mathbf{S}_{i,i-1} \; \mathbf{B}_{ii} \right]_1^M, \quad \mathbf{C} = \left[\mathbf{I}_{ii} \; \mathbf{C}_{i,i+1} \right]_1^M \tag{24}$$

$$\mathbf{B}_{11} = \mathbf{S}_{11}, \quad \mathbf{B}_{ii} = \mathbf{S}_{ii} - \mathbf{S}_{i,i-1}\mathbf{C}_{i-1,i}$$

$$\mathbf{C}_{i,i+1} = \mathbf{B}_{ii}^{-1}\mathbf{S}_{i,i+1}, \quad i = 1, 2, \ldots, M \tag{25}$$

$$\mathbf{P}_1 = \mathbf{B}_{11}^{-1}\mathbf{G}_1, \quad \mathbf{P}_i = \mathbf{B}_{ii}^{-1}(\mathbf{G}_i - \mathbf{S}_{i,i-1}\mathbf{P}_{i-1}), \quad i = 2, 3, \ldots, M \tag{26}$$

$$\mathbf{U}_M = \mathbf{P}_M, \quad \mathbf{U}_i = \mathbf{P}_i - \mathbf{C}_{i,i+1}\mathbf{U}_{i+1}, \quad i = M-1, M-2, \ldots, 1 \tag{27}$$

Once the $\mathbf{B}_{ii}^{-1}$ and $\mathbf{C}_{i,i+1}$ matrices are obtained the inverse of $\mathbf{S}$ can be calculated directly in submatrix form. For the symmetrical case with

$$\mathbf{S}^{-1} = \begin{bmatrix} \mathbf{Q}_{11} & \mathbf{Q}_{12} & \cdots & \mathbf{Q}_{1M} \\ \mathbf{Q}_{12}^T & \mathbf{Q}_{22} & \cdots & \mathbf{Q}_{2M} \\ \cdots & \cdots & \cdots & \cdots \\ \mathbf{Q}_{1M}^T & \mathbf{Q}_{2M}^T & \cdots & \mathbf{Q}_{MM} \end{bmatrix} \tag{28}$$

the main diagonal of the inverse is

$$\mathbf{Q}_{MM} = \mathbf{B}_{MM}^{-1}, \quad \mathbf{Q}_{ii} = \mathbf{B}_{ii}^{-1} + \mathbf{C}_{i,i+1}\mathbf{Q}_{i+1,i+1}\mathbf{C}_{i,i+1}^T, \quad i = M-1, M-2, \ldots, 1 \tag{29}$$

By columns above the main diagonal

$$\mathbf{Q}_{ik} = \mathbf{C}_{i,i+1}\mathbf{Q}_{i+1,k}, \quad k > i, \quad i = M-1, M-2, \ldots, 1 \tag{30}$$

The rest of the inverse is given by symmetry.

For some of the literature on tri-diagonal matrices see References 7 through 13.

FINITE DIFFERENCE SOLUTION FOR BEAM DEFLECTIONS (ONE-FOURTH ORDER EQUATION)

The beam deflection Equation 1 with definite difference form (Equation 4) was solved for the simple supported case by the procedure of Equations 15 to 21 using one by one submatrices. Equation 16 has five diagonals with $S_{i,i-2} = 1$, $S_{i,i-1} = -4$, $S_{ii} = 6$, $S_{i,i+1} = -4$, $S_{i,i+2} = 1$. The term p was selected to give an exact deflection of unity at the center of the beam. Table 1 shows the results for both single precision (8 places) and double precision (16 places) on the IBM 7090 computer for various values of M. Some computation times for double precision were 10 seconds for M = 200, 15 seconds for M = 500, 12 minutes and 12 seconds for M = 15,000. Table 1 also shows some single precision solutions for 100 elements using the tri-diagonal procedure with various sizes of the submatrices.

The results in Table 1 appear much better than might be expected on the basis of the maximum errors discussed in connection with Equations 7 and 12. The Choleski method of Equations 15 to 21 appears to have very little error accumulation due to sequence calculations in this case. That is, the inversions of the submatrices in Equation 18 at each step in the sequence appears to maintain the multiplying factors in Equations 20 and 21 near unity. Also the use of P_i and $C_{i,i+k}$ from the forward sweep calculations in the backsweep of Equations 21 reduces the accumulation effect.

TABLE 1

BEAM DEFLECTIONS BY FINITE DIFFERENCES IN FOURTH-ORDER DIFFERENTIAL EQUATION

Single Precision (8 places)		Double Precision (16 places)	
Number of elements M	Deflection at center	Number of elements M	Deflection at center
exact	1.0000000	exact	1.000000000000000
1 by 1 submat.		1 by 1 submat.	
50	0.99867544		
100	0.97705088	100	1.0000799
200	0.73315608	200	1.0000100
250	0.53907152		
		500	1.0000031
100		1000	1.0000008
2 by 2		1500	1.0000003
submat.	0.99986310	2000	1.0000002
4 by 4			
submat.	0.98772007	3000	1.0000002
5 by 5		4000	1.0000006
submat.	0.98101705	5000	1.0000015
10 by 10			
submat.	0.97795213	10000	1.0000210
		15000	1.0001084
75			
2 by 2			
submat.	1.0002720		

As might be expected the double precision calculations show an optimum number of elements for the best accuracy, the representation approximations showing up for a smaller number of elements and the round-off and method errors showing up for a larger number of elements. The decrease in accuracy for the larger number of elements probably arises from the large value of N in Equation 7 and the least value of U_e for the number of places in the calculations.

Table 1 shows an improvement in the results for the tri-diagonal calculations using submatrices over the five diagonal element calculations. However, this improvement may not always occur. Some eleven place calculations show 3 by 3 and 4 by 4 submatrices to be worse than the 1 by 1 elements but the 2 by 2 submatrices to be better than the 1 by 1 elements.

FINITE DIFFERENCE SOLUTION FOR BEAM DEFLECTION WITH EQUATION ORDER REDUCED

Since Equations 7, 8, and 9 show the order of the differential equation to have a large effect on the errors in the finite difference calculations, the fourth order beam equation was split into two second-order equations and the finite difference results for one equation used as input for the second equation. From Equation 5, the second-order equations are tri-diagonal in elements. The results are shown in Table 2. Comparison of the results in Tables 1 and 2 shows the 8-place double integration for 10,000 elements to be slightly better than the 8-place single integration for 100 elements (as Equations 7 and 8 predict).

TABLE 2

BEAM DEFLECTIONS BY FINITE DIFFERENCES WITH EQUATION ORDER REDUCED

Single Precision (8 places)		Double Precision (16 places)	
Number of elements M	Deflection at center	Number of elements M	Deflection at center
Two Second-Order Equations			
100	1.0000507		
200	1.0000213		
500	0.99999984	500	1.0000032
1000	1.0000211		
2000	1.0000467		
3000	0.99995088		
4000	0.99974568		
5000	0.99929824		
7000	0.99752952		
10000	0.98744416	20000	0.99999944
Four First-Order Equations			
1000	1.0000005		
20000	0.99998515		
100000	0.99997171		

The fourth-order equation was also split into four first-order equations. Direct transfer solutions were made in this case (see discussion of transfer matrices in a later section), the matrix procedure of Equation 13 not being necessary. Since all terms in the calculations were positive, there was no subtraction error accumulations so that only round-offs in addition affected the accuracy. There is little order effect in the first order equations. The result in Table 2 for 100,000 elements is much better than the possible round-off error accumulation of $(10^5)\ (10^{-8}) = 10^{-3}$ for just 10^5 simple multiplications.

From Tables 1 and 2 it is evident that reducing the order of the differential equations has a large effect on the accuracy of the calculations in finite differences. Also, the time of computation is reduced for the lower order equations. The time for the first order case of 100,000 elements on the IBM 7090 computer was less than two minutes.

BACKSWEEP USING ORIGINAL EQUATIONS

In some cases, after the $\mathbf{U}_i$ are obtained for i = M, M-1, . . . , M-N+1, the backsweep recursion in Equations 21 and 27 can be replaced by a backsweep using the original Equation 13. If the system is arranged so that $|\mathbf{S}_{i,i-N}| \neq 0$, then

$$\mathbf{U}_i = \mathbf{S}^{-1}_{i+N,i}\left(\mathbf{G}_{i+N} - \sum_{k=1}^{N} \mathbf{S}_{i+N,i+k}\,\mathbf{U}_{i+k}\right), \quad i = M-N, \ldots, 1. \tag{31}$$

This procedure avoids the storage of the $\mathbf{C}_{i,i+j}$ and $\mathbf{P}_i$ submatrices in the computer and can save considerable time. However, it may not be possible to make the backsweep in Equation 31 due to error accumulation in the subtractions. From Equations 4 and 6 the k_2 factor in Equation 12 may be quite large so that the error builds up rapidly. This effect does not appear to be serious in the simple beam problems discussed above, but in a plate problem using Equation 6 with 11 by 11 submatrices the stresses became meaningless after about ten steps in the sweepback using Equation 31. No difficulty arose in using the Equation 21 sweepback for forty steps in the same problem.

There is a technical reason for not using Equation 31 in certain physical problems. If the boundary conditions produce local effects and show only average or no effects elsewhere in the structure, then the sequence in Equation 31 coming from an average region cannot pick up the local effects. In a long plate with concentrated loads on the ends there is a local shear lag problem near the ends with a uniform stress away from the ends. In a backsweep from one end with Equation 31 there is no way to detect the shear lag at the other end.

Thus, it appears that regardless of the finite number of places used in the calculations direct methods of solution based on forward or back substitution in the original system of equations must fail for certain types of physical problems under certain boundary and loading conditions. In such a problem it would appear that an inversion method based on this solution method must always fail, since the inverse must give a solution for any load case.

This difficulty with local effects produced by certain loads in certain physical problems does not occur in the modified Choleski method. The backsweep in Equations 21 and 27 uses a load term $\mathbf{P}_i$ calculated on the forward sweep so that the effect of all loads is included at the point. In Equation 31 the $\mathbf{P}_i$ effect is included only indirectly through its effect, if any, on $\mathbf{P}_M$ at the far end.

SINGLE SUBMATRIX INVERSION IN TRI-DIAGONAL SYSTEM

If $|\mathbf{S}_{i,i-1}| \neq 0$ in Equation 23, all submatrices are the same size, and $\mathbf{S}_{i,i-1}$ is easy to obtain, then Schechter (Reference 8) gives a procedure which avoids all the $\mathbf{B}_{ii}$ inversions in Equation 25 except the last one. Multiply Equation 23 by the diagonal matrix $\mathbf{S}^{-1}_{i,i-1}$ to give

$$\bar{S}U = \bar{G}, \quad \bar{S} = \left[I \;\; \bar{S}_{i,i} \;\; \bar{S}_{i,i+1} \right]_1^M, \quad \bar{S}_{ii} = S^{-1}_{i,i-1} S_{ii},$$

$$\bar{S}_{i,i+1} = S^{-1}_{i,i-1} S_{i,i+1}, \quad \bar{G}_i = S^{-1}_{i,i-1} G_i \tag{32}$$

Equation 25 becomes

$$B_{11} = \bar{S}_{11}, \quad B_{ii} = \bar{S}_{ii} - C_{i-1,i},$$

$$C_{i,i+1} = B^{-1}_{ii} \bar{S}_{i,i+1}, \quad i = 1, 2, \ldots, M \tag{33}$$

Define

$$B_{ii} \equiv H^{-1}_{i-1} H_i, \quad Q_i \equiv H_i P_i \tag{34}$$

whence Equations 33 and 26 become, respectively,

$$H_0 = I, \quad H_1 = S_{11}, \quad H_i = H_{i-1} \bar{S}_{ii} - H_{i-2} \bar{S}_{i-1,i} \tag{35}$$

$$Q_1 = \bar{G}_1, \quad Q_i = H_{i-1} \bar{G}_i - Q_{i-1} \tag{36}$$

The recursion formula for Q_i in Equation 36 can be written as a sum so that

$$Q_M = \sum_{i=1}^{M} (-1)^{M-1} H_{i-1} \bar{G}_i \tag{37}$$

whence Equations 27 and 34 give

$$U_M = P_M = H_M^{-1} Q_M = H_M^{-1} \sum_{i=1}^{M} (-1)^{M-i} H_{i-1} \bar{G}_i \tag{38}$$

and only H_M has to be inverted. The backsweep by Equation 31 gives

$$U_{M-1} = \bar{G}_M - \bar{S}_{MM} U_M, \quad U_{i-1} = \bar{G}_i - \bar{S}_{ii} U_i - \bar{S}_{i,i+1} U_{i+1},$$

$$i = M-1, M-2, \ldots, 2 \tag{39}$$

By taking $G_i = G_k = I$ for $i = k$ and $G_i = 0$ for $i \neq k$ and solving for U_i as a square matrix for each i, the inverse of $\bar{S}$ in Equation 32 can be obtained by rows from Equations 38 and 39 as follows (see Equation 28 for notation for submatrices in the inverse)

$$Q_{Mk} = H_M^{-1} (-1)^{M-k} H_{k-1} S^{-1}_{k,k-1}, \quad S^{-1}_{10} = I,$$

$$Q_{M-1,M} = S_{M,M-1} - \bar{S}_{MM} Q_{MM}, \quad Q_{M-1,k} = -\bar{S}_{MM} Q_{Mk},$$

$$Q_{i-1,k} = -\bar{S}_{ii} Q_{ik} - \bar{S}_{i,i+1} Q_{i+1,k}, \quad i = M-1, M-2, \ldots, 2, i \neq k, \tag{40}$$

$$Q_{k-1,k} = S^{-1}_{k,k-1} - \bar{S}_{kk} Q_{kk} - \bar{S}_{k,k+1} Q_{k+1,k}.$$

This method gives very simple solutions for certain classes of problems. For example, in Reference 13 the three-moment equation of a uniform beam is solved by this procedure for any number of supports. Also, in Reference 13 any element of the inverse of the matrix of coefficients can be written down from a simple formula. However, for more involved problems

the method has serious error accumulations. The numerical values of the H_i matrices increase in magnitude so that it may not be possible to calculate the inverse of H_M. With eleven place calculations on the beam problem of Table 1 above, this procedure on a single matrix inversion (2 by 2 submatrices in Equation 35) failed after about M = 200 elements when the terms in H_M reached the order of 10^{11} and 10^{12}. The regular tri-diagonal procedure gave satisfactory results to M = 1200 elements with eleven place calculations. On a plate problem with 18 by 18 submatrices H_8 could not be inverted after eight steps. The regular tri-diagonal was satisfactory with forty steps. The H_i matrices apparently will grow in size if the $\bar{S}_{ii}$ matrices have elements much larger than those in $\bar{S}_{i-1,i}$.

RELATIVE MAGNITUDE OF MATRIX ELEMENTS

If some terms in a matrix are very large or very small compared to the other terms it may not be possible to invert the matrix. In a structure such a situation may occur if one member is very stiff or very flexible compared to the others. To get an idea of the affect a 6 by 6 stiffness matrix for a simple structure was inverted for various increased stiffnesses of one member. For eleven places in the calculations the inverse was satisfactory up to a relative stiffness ratio of 10^6, off some at 10^7, and no good at 10^8. There was little effect of the stiffness ratio on the other terms in the inverse in the range above 10^3 for the ratio, so that an effective stiffness can usually be used to keep within the accuracy range. If a member is quite flexible, omit it.

TRANSFER MATRICES

Basically, transfer matrices can be used for the solution of a system of first order equations by finite differences. From the finite difference form

$$\frac{dw}{dx} = g \ , \qquad w_i = w_{i-1} + g(\Delta x) \tag{41}$$

the matrix form of a two equation system can be written as

$$\begin{bmatrix} w_i \\ v_i \\ 1 \end{bmatrix} = \begin{bmatrix} a_{11}(i) & a_{12}(i) & a_{13}(i) \\ a_{21}(i) & a_{22}(i) & a_{23}(i) \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} w_{i-1} \\ v_{i-1} \\ 1 \end{bmatrix} \tag{42}$$

or

$$U_i = D_i U_{i-1} \ ,$$

where the third column represents the non-homogeneous terms in the equation. It can be deleted in homogeneous systems. For M elements with

$$K_i = D_i D_{i-1} \cdots D_2 \ , \quad i = 2, 3, \ldots, M \tag{43}$$

it follows that

$$U_i = K_i U_1 \ , \quad U_M = K_M U_1 \ , \quad i = 2, 3, \ldots, M \tag{44}$$

where D_i is the transfer matrix from U_{i-1} to U_i and K_i transfers from U_1 to U_i. For given boundary conditions U_1 and U_M can be calculated from $U_M = K_M U_1$ and all the values of U_i determined from Equation 42 by starting at i = 2. Except for the solution for U_1 and U_M only simple matrix multiplications are involved in the transfer matrix procedure. It also allows various boundary conditions to be used without recalculating K_M. Any number of load columns can be used and these load columns do not affect the calculations for the other

columns. Since first order equations are involved, there is little order effect so that the calculations will be quite accurate (see Table 2).

In many cases, in order to maintain accuracy and to simplify calculations, it would appear desirable to convert the problem to first order equations and use transfer matrices. Ordinarily, this conversion should be made on the physical problem before the system of equations are set up. However, it may be possible to split the matrix equation system (Equation 13) into a transfer matrix form. Consider the tri-diagonal Equation 32 with all submatrices the same size:

$$U_{i-1} + \bar{S}_{ii}\bar{U}_i + \bar{S}_{i,i+1}\bar{U}_{i+1} = \bar{G}_i \tag{45}$$

Define

$$\bullet_i = U_{i-1} + A_i U_i \tag{46}$$

where A_i is to be selected to give as simple and as accurate transfer matrix as possible. From Equations 45 and 46

$$\begin{bmatrix} U_i \\ \bullet_i \\ 1 \end{bmatrix} = \begin{bmatrix} -A_{i+1} & I & 0 \\ Q_i & R_i & \bar{G}_i \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} U_{i+1} \\ \bullet_{i+1} \\ 1 \end{bmatrix} \tag{47}$$

$$Q_i = (\bar{S}_{ii} - A_i)A_{i+1} - \bar{S}_{i,i+1}, \quad R_i = -(\bar{S}_{ii} - A_i) \tag{48}$$

It is evident that if A_i is selected as $A_i = \bar{S}_{ii}$ the transfer matrix is considerably simplified. However, this makes Equation 47 equivalent to a backsweep in the original equations 31 and makes the recursion in Equation 43 equivalent to the recursion on the H matrices in Equation 35, except the recursion is made from the right end instead of the left end. Thus, accuracy problems as discussed above for these procedures can arise with this value of A_i in the transfer matrix.

If Equation 45 represents a true second-order system then $A_i = -I$ will ordinarily convert the system to first order so that Equation 47 will have the accuracy of a first order system. However, if Equation 45 represents a repartitioned higher order system, it is unlikely an A_i can be found to maintain the tri-diagonal accuracy in the transfer form of Equation 47.

From Equations 26 and 27 it is apparent that the tri-diagonal solution can be expressed in transfer matrix form on both the forward sweep and the backsweep. The forms are

$$\begin{bmatrix} P_i \\ 1 \end{bmatrix} = \begin{bmatrix} -B_{ii}^{-1}S_{i,i-1} & B_{ii}^{-1}G_i \\ 0 & 1 \end{bmatrix} \begin{bmatrix} P_{i-1} \\ 1 \end{bmatrix} = \begin{bmatrix} D_i & F_i \\ 0 & 1 \end{bmatrix} \begin{bmatrix} P_1 \\ 1 \end{bmatrix} \tag{49}$$

$$\begin{bmatrix} P_M \\ 1 \end{bmatrix} = \begin{bmatrix} U_M \\ 1 \end{bmatrix} = \begin{bmatrix} -B_{MM}^{-1}S_{M,M-1} & B_{MM}^{-1}G_M \\ 0 & 1 \end{bmatrix} \begin{bmatrix} D_{M-1} & F_{M-1} \\ 0 & 1 \end{bmatrix} \begin{bmatrix} P \\ 1 \end{bmatrix} \tag{50}$$

$$\begin{bmatrix} U_i \\ 1 \end{bmatrix} = \begin{bmatrix} -C_{i,i+1} & P_i \\ 0 & 1 \end{bmatrix} \begin{bmatrix} U_{i+1} \\ 1 \end{bmatrix} \tag{51}$$

$$\begin{bmatrix} U_i \\ P_1 \\ 1 \end{bmatrix} = \begin{bmatrix} -C_{i,i+1} & D_i & F_i \\ 0 & I & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} U_{i+1} \\ P_1 \\ 1 \end{bmatrix} = \begin{bmatrix} A_{iM} & A_{i1} & G_{i1} \\ 0 & I & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} U_M \\ P_1 \\ 1 \end{bmatrix} \tag{52}$$

$$\begin{bmatrix} U_1 \\ P_1 \\ 1 \end{bmatrix} = \begin{bmatrix} A_{1M} & A_{11} & G_{11} \\ 0 & I & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} U_M \\ P_1 \\ 1 \end{bmatrix} \tag{53}$$

Once D_{M-1}, F_{M-1}, A_{1M}, A_{11}, and G_{11} have been calculated, Equations 50 and 53 permit U_1 and U_M to be calculated for any selected values of G_1 and G_M. Also, G_1 and G_M can be determined for specified values of U_1 and U_M.

CONCLUSIONS

It has been demonstrated that, for a fixed number of decimal places in the calculations, the differential order of the system, the sequencing of the computations, and the relative magnitude of the coefficients are the primary sources of error in the solution of systems of simultaneous equations. All these error sources can be avoided or reduced by reducing the order of the system, by keeping multiplying factors small in the sequence of computations, and by using realistic coefficients in the equations. The transfer matrix procedure for a true first order system is very accurate and has practically no limit on the number of equations (it lost four places for 100,000 equations in a beam deflection problem). The modified Choleski method using submatrices has little error accumulation in the calculation sequence but depends on the differential order of the system (it lost six places for 100 fourth-order equations and lost six places for 10,000 second-order equations repeated to give the beam deflections). Matrix inversion using this method is as accurate as the solution method for each column independently.

Also, it has been shown that any direct method using forward or back substitution in the original system of equations has not only the differential order problem but also an error accumulation in the sequence of calculations so that it may fail much quicker on both the solution and the inverse than the Choleski procedure. It was also found that efforts to reduce the number of submatrix inversions can make the elements of the matrices large and lead to difficulties in the sequence of calculations. Finally, it was shown that the tri-diagonal procedure can be written in a transfer matrix form which maintains accuracy but that splitting the tri-diagonal form to force an apparent first order transfer matrix system may fail unless the system actually becomes a true first order system.

REFERENCES

1. Von Neuman, John, and Goldstine, H. H., "Numerical Inverting of Matrices of High Order," Bulletin American Mathematical Society, Vol. 53, p. 1021, 1947.

2. Varga, Richard S., Matrix Iterative Analysis, Prentice Hall, Inc., Englewood Cliffs, N. J., 1962.

3. McMinn, S. J., Matrices for Structural Analysis, John Wiley and Sons, Inc., New York, 1962.

4. Fox, L., Huskey, H. D., and Wilkinson, J. H., "Notes on the Solution of Algebraic Linear Simultaneous Equations," Quart. Journal of Mechanics and Applied Mathematics, Vol. 1, p. 149-173, 1948.

5. Turing, A. M., "Rounding-Off Errors in Matrix Processes," Quart. Journal of Mechanics and Applied Mathematics, Vol 1, pp. 287-308, 1948.

6. Pestel, Eduard, and Leckie, F. A., Matrix Methods in Elastomechanics, McGraw-Hill Book Co., New York, 1963.

7. Kosko, Eric, "Matrix Inversion by Partitioning." Aeronautical Quart., Vol. VIII, p. 157-184, May 1957.

8. Schechter, Samuel, "Quasi-Tri-Diagonal Matrices and Type-Insensitive Difference Equations," Quart. of Appl. Maths., Vol. 18, p. 285-295, October 1960.

9. Oliphant, Thomas A., "An Extrapolation Procedure for Solving Linear Systems," Quart. of Appl. Maths,. Vol. 20, p. 257-265, October 1962.

10. Cornock, A. F., "The Numerical Solution of Poisson's and the Bi-Harmonic Equations by Matrices," Proc. of Cambridge Philosophical Society, Vol. 50, p. 524-535, 1954.

11. Karlqvist, Olle, "Numerical Solution of Elliptic Differential Equations," Tellus 4, p. 374-384, 1952.

12. Asplund, S. O., Inversion of Band Matrices, ASCE 2nd Conference on Electronic Computation, Pittsburgh, Pa., No. 1960-42, September 1960.

13. Gatewood, B. E., and Norik Ohanian, "Note on Solution of a System of Three Moment Equations," AIAA Journal, Vol. 1, p. 1965, January 1963; "Tri-Diagonal Matrix Method for Complex Structures," ASCE Structural Division Journal, Vol. 91, No. ST2, p. 27-41, April 1965.

14. Bodewig, E., "Matrix Calculus," North Holland Publishing Co., Amsterdam 1959.