NUMERICAL CONDITIONING OF STIFFNESS MATRIX FORMULATIONS
FOR FRAME STRUCTURES

R. A. Rosanoff,* J. F. Gloudeman,** and S. Levy***

North American Rockwell Corp., Space Division,
Downey, California

This is a numerical experiment employing one hundred structural models. Over five hundred inversions are performed and more than three hundred eigensolutions accomplished. The Householder-Givens routine for the eigenvalue problem is found to have a previously unreported weakness. It is shown that, among good algorithms, there is a secondary effect of algorithm upon numerical precision, but that the major effect is caused by truncation of lower mode information in the original numerical representation of the matrix. Matrix equilibration, or scaling, has a major effect on the measurement of matrix conditioning. The most important effect on the conditioning of stiffness matrices is observed to be the ratio of element stiffness to the stiffness of the supporting structure. Preferred coordinates have a beneficial effect on numerical error in stiffness matrix problems.

* Member, Technical Staff, Structures,
** Assistant Director, Computing,
*** Member, Technical Staff
SECTION I

CONCEPTUAL FRAMEWORK AND METHODOLOGY

The set of numbers available on any computer is finite and non-dense. It follows from this that it lacks the properties of closure and associativity exhibited by the real number set. Since these properties are included among the properties of a field, and a field is implied in the definition of an algebra, it is clear that the computer number set and the real number set are fundamentally different.

They are different in kind, not just in degree. This observation is valid, whether one is thinking about the algebra involved in the formulation of stiffness matrices, or whether one is thinking about the description of algorithms for the manipulation of matrices once formulated. Numerical error arises solely because of these differences between the real number set and the computer number set. Of course, the success of most computation rests on the fact that the computer number set approximates the real number set adequately for the solution of most problems. However, it should not be overlooked that numerical error measures the extent of this approximation. This means that numerical error measures the failure of the most elementary and fundamental theorems to predict the results of computation. For example, there is the theorem which states that a matrix is singular if, and only if, its determinant is zero. However, the fact of the matter is that a computed zero determinant is more likely to demonstrate the failure of the property of closure for the computer number set than the singularity of a matrix. Not only that, but it is easy to produce hopelessly ill-conditioned matrices with determinants equal to one. And other theorems could be cited. The significant point is that numerical error measures specifically the failure of mathematical theorems to establish truth for computation. Research without a valid means for establishing truth will cause trouble. Hence the first question for error analysis is the choice of methodology. Our choice is to employ the methods of experimental science.

It is important to recognize that, although our work draws heavily on the accomplishments of the many fine numerical analysts who have built the foundations of matrix error analysis, we are operating in a different conceptual framework. We use different definitions and we attach different significance to some of the same terms and entities. Specifically, most of the work done in the past is primarily concerned with the comparison and improvement of algorithms for matrix manipulation. For this work it is common practice to "take the matrix representation in the computer as given." We, on the other hand, address ourselves
to the questions of the eigensolution or the inversion of matrices defined on the real number field. We define numerical error to be the error in numerically obtained inverses, or eigensolutions as compared to the real number solutions of the corresponding problems.

Let $A$ be a matrix defined over the real numbers. Let $B$ be its inverse. Let $C$ be the machine representation of $A$, and let $X$ be a candidate numerical inverse. The traditional approach would be to study the effect of various algorithms in the inversion of $C$. Sometimes bounds are set up about $C$ which include a matrix for which $X$ is the exact inverse. Both Reference 1, p. 319 and Reference 2, p. 84 go to some pains to point out that it is the matrix $C$ whose inverse they are studying, not the matrix $A$!

We, on the other hand, define the error matrix to be the matrix $B - X$. In so doing, we avoid prejudging the effects on the final answer of errors in representing the matrix in a finite number set, the effects of inaccuracies in the physical parameters of the problem, or for that matter, the effect of manipulation error itself. Philosophically, our position is that structural analysis implicitly assumes the properties of the real number set. Numerical error, whatever its immediate source in an analysis, erodes the value of the computed solution.

The first problem encountered in designing an experimental technique for error analysis is the determination of error. It is most important to keep in mind at this stage that the equivalence of different error measures in the "real number matrix algebra" are completely unreliable for the comparison of error measures in computation. In particular, in the use of mixed precision it is easy to demonstrate cases (Reference 3) in which the norms of $CX - I$ are decreased by eight orders of magnitude, while the norms of $B - X$ are actually increasing.

For the Hilbert tenth-order matrix we have performed an interesting little experiment to illustrate the care which is necessary. We took the matrix $C$ to be a single precision representation of the 10th order Hilbert $A$. The numerical inverse $X$ was obtained using a double precision inversion routine. We then inverted $X$, using the same routine. We found that $X^{-1}$ was about a seven-digit approximation to $A$, whereas $X$ was three orders magnitude away from the true inverse $B$.

For some problems, (usually not interesting in practical applications) the inverse $B$ of the real number matrix $A$ is known in closed form. With many such problems it is easy to demonstrate that the real number matrix $A$, or its real number inverse $B$ are represented to the full precision of the machine. In such a case one may directly observe the matrix of
errors as we have defined them, that is, $B - X$. The problem remains to obtain such a measure for the significant cases where the Matrix $B$, the real number inverse, is not known in closed form. Our solution (and this is critical to the validity of our experiments) is to use full, double precision, as a standard for error measurement on single precision.

Since the validity of our experiment depends on this solution, the solution itself requires careful consideration. By full double precision we mean that our matrix shall be represented in double precision, and the manipulation, be it the solution of equations, matrix inversion or the eigenvalue problem, shall also be accomplished in double precision. The rationale of requiring that the representation as well as the manipulation of the matrix be in double precision will be clearer after a discussion of the error of representation in matrices, which is the subject of the next section. Immediately, however, the idea is reasonable.

Because of the importance of the issue, some preliminary experiments were conducted to validate the standard. Forty matrices, including some Hilberts, some stiffness matrices, and some transfer matrices, all of whose real number solutions were known in closed form, were used in the following manner, Double precision representations of the matrices were generated and the problems, inversion or the eigenvalue problem, were solved in double precision. The matrices were truncated into single precision, and the resulting problems were solved in both single and double precision. The real number solutions were generated in double precision and were used for a standard of comparison for the other solutions. Matrices with integer coefficients were avoided because their single precision representations for double precision manipulation are indistinguishable from their double precision representations. The results showed that, in all cases, the full double precision solutions were eight orders of magnitude better than either the single or mixed precision solutions. Based on this, and such evidence as is developed later in support of this experimental model, we shall consider the double precision matrix as being, for our purposes, equivalent to the real number matrix. The matrix $B$ shall be considered to be equivalent to the numerically obtained double precision inverse of $A$.

The matrix $B - X$ is the primary data of our inverse experiments, but it is frequently not the matrix of major concern to the engineer. The matrix of relative errors, defined by $(b_{i,j} - x_{i,j})/b_{i,j}$ ($b_{i,j}$ is an element of $B$) has the advantage that it directly mirrors the number of good digits in the solutions. Of course, this matrix has one disadvantage — the relative error may become large for the smaller elements of $B$.
During our experiment, more than five hundred inversions and in excess of three hundred eigensolutions were performed. For the inversion solutions the elements of $\mathbf{A}$, $\mathbf{B}$, $\mathbf{X}$, the matrix of absolute errors, and the matrix of relative errors were determined and printed out. Further, the absolute errors were summarized as mean errors, standard deviation of the errors, and maximum errors. For the relative errors the same information was output in addition to frequency distributions of the relative errors by magnitude. It is estimated that the total output count of the matrix terms and the error terms was approximately one-half million. Over three thousand matrix norms and conditioning measures were determined and printed. At one time or another during the investigation, fourteen inversion and five eigenvalue routines were used. The bulk of the investigations were carried out on the IBM 360, 1410, 7094, 7094 Emulator on the 360, and the Univac 1108.

Despite this rather massive assault, significant elements of bias in the experiment are recognized. For example, the largest matrix investigated was of order 75. There was a preponderance of beam problems with their topologically simple networks. Almost all of the matrices were symmetric, positive definite. Although iterative improvement schemes were among the methods used, no purely iterative or relaxation solutions of equations were performed. The QR and LU algorithms were not employed. The philosophy behind these restrictions was mainly economic. It was felt that an unbiased experiment could not be devised. The best allocation of resources seemed to dictate a detailed look at the experimental sample which was most readily available.

In the next section, the problems of error of representation of matrices is examined along with some measures of matrix conditioning. The third section describes the experiment and the final section provides results and conclusions.
SECTION II

ERROR OF REPRESENTATION

The columns of the $n^{th}$ order identity matrix serve as a particularly convenient, or
"natural" basis, or coordinate system for the representation of vectors in an $n$-space. That
is, any vector in the space may be represented as

$$\mathbf{x} = \sum_{i=1}^{n} a_i \mathbf{e}_i$$

If one takes only $m$ of these vectors $m \leq n$, they serve to describe an $m^{th}$ order subspace.
These particular subspaces, formed as linear combinations of $m$ of the columns of the
identity, $e_i$, shall be called standard subspaces. Thus the vectors

$$3 \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \end{bmatrix} + 7 \begin{bmatrix} 0 \\ 1 \\ 0 \\ 0 \end{bmatrix} = \begin{bmatrix} 3 \\ 7 \\ 0 \\ 0 \end{bmatrix} \text{ and } 5 \begin{bmatrix} 0 \\ 0 \\ 1 \\ 0 \end{bmatrix} + 2 \begin{bmatrix} 0 \\ 0 \\ 0 \\ 1 \end{bmatrix} = \begin{bmatrix} 0 \\ 5 \\ 2 \end{bmatrix}$$

belong to disjoint standard subspaces of order 2. (Disjoint because there is no vector belonging
to the smallest standard subspace containing the first vector, which also belongs to the
smallest standard subspace containing the second vector.) We shall need this bit of jargon,
the idea of standard subspaces, shortly. The dynamicist may be more at home with the idea
if we point out that when the eigenvectors or mode shapes of a structure fall into disjoint
standard subspaces, the modes are said to be uncoupled.

It can be shown, without too much trouble, that a symmetric matrix can be represented as

$$\mathbf{A} = \sum_{i=1}^{n} \lambda_i \mathbf{U}_i \mathbf{U}_i^T$$

(1)

where the $\lambda_i$ are the eigenvalues and the $\mathbf{U}_i$ are the corresponding eigenvectors. The rank
one nxn matrices $\mathbf{U}_i \mathbf{U}_i^T$ are idempotent (i.e., $(\mathbf{U}_i \mathbf{U}_i^T)^p = (\mathbf{U}_i \mathbf{U}_i^T)$) and are referred
to as the principal idempotents of the matrix. Since the eigenvalues of the inverse, if it
exists, are the reciprocals of the eigenvalues of the matrix itself, then

$$\mathbf{A}^{-1} = \sum_{i=1}^{n} \frac{1}{\lambda_i} \mathbf{U}_i \mathbf{U}_i^T$$

(2)
For example, consider the matrix

\[
A(z) = \begin{bmatrix}
  z & 0 & 0 \\
  0 & \sinh(z) & \cosh(z) \\
  0 & \cosh(z) & \sinh(z)
\end{bmatrix}
\]

whose inverse is easily checked to be

\[
A^{-1}(z) = B(z) = \begin{bmatrix}
  \frac{1}{z} & 0 & 0 \\
  0 & -\sinh z & \cosh z \\
  0 & \cosh z & -\sinh z
\end{bmatrix}
\]

The eigenvalues and eigenvectors are

\[
\begin{align*}
\lambda_1 &= e^z \\
\lambda_2 &= z \\
\lambda_3 &= -e^{-z}
\end{align*}
\]

\[
\begin{align*}
U_1 &= \begin{bmatrix} 0, & \frac{\sqrt{2}}{2}, & \frac{\sqrt{2}}{2} \end{bmatrix} \\
U_2 &= \begin{bmatrix} 1, & 0, & 0 \end{bmatrix} \\
U_3 &= \begin{bmatrix} 0, & \frac{\sqrt{2}}{2}, & -\frac{\sqrt{2}}{2} \end{bmatrix}
\end{align*}
\]

Forming the linear combination of idempotents, one has

\[
A(z) = e^z \begin{bmatrix}
  0 & 0 & 0 \\
  0 & 1/2 & 1/2 \\
  0 & 1/2 & 1/2
\end{bmatrix} + z \begin{bmatrix}
  1 & 0 & 0 \\
  0 & 0 & 0 \\
  0 & 0 & 0
\end{bmatrix} - e^{-z} \begin{bmatrix}
  0 & 0 & 0 \\
  0 & 1/2 & -1/2 \\
  0 & -1/2 & 1/2
\end{bmatrix}
\]

For \(z = 8\) this gives in eight-digit arithmetic the machine representation \(C(8)\).

\[
C(8) = \begin{bmatrix}
  \text{\ldots} & \text{\ldots} & \text{\ldots} \\
  .00000000 & .00000000 & .00000000 \\
  8.00000000 & .00000000 & .00000000 \\
  .00000000 & .00000000 & .00000000 \\
  8.00000000 & \text{\ldots} & \text{\ldots} \\
  .00000000 & .00000000 & .00000000 \\
  .00000000 & .00000000 & .00000000 \\
  .00000000 & 1490.4790 & 1490.4790 \\
  .00000000 & .00000000 & +.00000000 \\
  .00000000 & -.0016773132 & .00016773132 \\
  .00000000 & 1490.47883228888 & 1490.47883228888 \\
  .00000000 & .0016773132 & -.00016773132 \\
  .00000000 & 1490.47916773132 & 1490.47916773132 \\
  .00000000 & .00000000 & .00000000 \\
  .00000000 & .00016773132 & -.00016773132 \\
  .00000000 & 1490.47883228888 & 1490.47883228888
\end{bmatrix}
\]
Inverting this numerically, one obtains

\[
X(8) = \begin{bmatrix}
0.1250000 & 0 & 0 \\
0 & -1667.781748 & 1667.816964 \\
0 & 1667.781715 & -1667.816629
\end{bmatrix}
\]

The true inverse, as one sees by inspecting the closed form inverse shown, is obtained by substituting the reciprocal of \(a_{11}\) for \(b_{11}\) and changing the sign of \(a_{22}\) and \(a_{33}\). Clearly, the lower right two-by-two sub-matrix of the numerically obtained inverse \(X(8)\) shown above has only one good digit.

The matrix \(A(8)\) illustrates what we consider to be the most important concept in matrix error analysis. In the lower right-hand two-by-two submatrix it is seen that the contribution of the idempotent \(U_3 U_3^\dagger\) is represented to fewer digits than the idempotent \(U_1 U_1^\dagger\). Further, the difference in the number of decimal digits in the representation is approximately the number of powers of ten in the ratio of the extreme eigenvalues associated with that standard subspace. Since the idempotent which must dominate the inverse is \(U_3 U_3^\dagger\), it is clear that even an eight-digit representation \(C(8)\) approximating \(A(8)\) is quite deficient in information about \(A^{-1}(8) = B(8)\). It is also clear that the measure of this deficiency is the size of the shift from the first nonzero decimal digit in the dominating idempotent to the first nonzero decimal digit in the subordinate idempotent. This is measured by the logarithm of the ratio of the eigenvalues.

Another example will serve to enlarge our understanding of this phenomenon.

Consider the two-spring system shown above. The stiffness matrix is

\[
K = \begin{bmatrix}
K_a & -K_a \\
-K_a & K_a + K_b
\end{bmatrix}
\]

and

\[
K^{-1} = \begin{bmatrix}
\frac{1}{K_a + K_b} & \frac{1}{K_b} \\
\frac{1}{K_b} & \frac{1}{K_b}
\end{bmatrix}
\]

1036
Solving the eigenvalue problem for the matrix $K$ one has:

$$\lambda_1 = K_a + \frac{K_b}{2} + \sqrt{K_a^2 + \frac{K_b^2}{4}}$$

and

$$\lambda_2 = K_a + \frac{K_b}{2} - \sqrt{K_a^2 + \frac{K_b^2}{4}}$$

The ratio of the eigenvalues may be shown to be

$$\frac{\lambda_1}{\lambda_2} = 1 + 2 \frac{K_a}{K_b} + \frac{K_b}{2K_a} + \frac{1}{2} \sqrt{\left(\frac{4K_a}{K_b} + \frac{K_b}{K_a}\right)^2 + 4 \left(\frac{K_a}{K_b} + \frac{K_b}{K_a}\right)}$$

which clearly becomes large in the case of a rigidity mismatch whether $K_a$ is larger or smaller than $K_b$.

The eigenvectors may be written as $U_1 = (u_{1,1}, u_{2,1})$

where

$$u_{1,1} = -\frac{K_b}{2} - \sqrt{K_a^2 + \frac{K_b^2}{4}}$$

and $U_2 = (u_{1,2}, u_{2,2})$

where

$$u_{2,2} = -\frac{K_b}{2} + \sqrt{K_a^2 + \frac{K_b^2}{4}}$$

Normalizing the vector $U_1$ one has

$$(u_{1,1})^2 = \frac{K_a^2}{2K_a^2 + \frac{K_b^2}{2} + K_b \sqrt{K_a^2 + \frac{K_b^2}{4}}}$$
Using also the fact that the eigenvectors of a symmetric matrix are orthogonal, it can be seen that

1. If \( K_a \), the spring away from the support, is very large compared to \( K_b \), the eigenvectors approach

\[
\begin{align*}
\mathbf{u}_1 &= \left[ \frac{\sqrt{2}}{2}, \frac{\sqrt{2}}{2} \right] \\
\mathbf{u}_2 &= \left[ \frac{\sqrt{2}}{2}, -\frac{\sqrt{2}}{2} \right]
\end{align*}
\]

2. If \( K_b \), the spring next to the support is very large compared to \( K_a \), \( u_{1,1} \rightarrow 0 \) and the eigenvectors approach

\[
\begin{align*}
\mathbf{u}_1 &= \begin{bmatrix} 0, & 1 \end{bmatrix} \\
\mathbf{u}_2 &= \begin{bmatrix} 1, & 0 \end{bmatrix}
\end{align*}
\]

That is to say, as the spring constant, \( K_b \), next to the support becomes great relative to \( K_a \), the eigenvectors tend to disjoint standard subspaces.

An experiment was conducted with this family of matrices along the following lines:

\( K_a \) was maintained constant at \( \sqrt{2} \)

\( K_b \) was varied from \( \sqrt{3}, 10^{-9} \) to \( \sqrt{3}, 10^{-8} \) by reducing \( K_b \) by a factor of ten after each experiment.

The matrix was determined to full double precision on the Emulate 7094 system operating on the 360.

The “Real Number” inverse \( A \) was determined in full double precision from the closed form inverse.

The double precision stiffness matrix was inverted using a double precision Gauss-Jordan inverse routine.

The eigenvalue problem was solved using a double precision Jacobi routine.
A single precision stiffness matrix was determined by truncation from the double precision stiffness matrix. This matrix was inverted using the same double precision Gauss-Jordan routine and a single precision Gauss-Jordan routine derived from the double precision routine by change of specification cards. The eigensolution for the single precision stiffness matrix was performed using the same double precision Jacobi routine used on the double precision stiffness matrix. The results were as follows:

1. The direct inverse and the full double precision inverse were indistinguishable for determination of relative errors in the single precision and mixed precision inverses.

2. For $K_b > K_a$, the relative errors in the single and mixed precision matrices were of order $10^{-8}$. During these runs the ratio of eigenvalues decreased from .12247499 10 to approximately 1.

3. For $K_a > K_b$ the ratio of eigenvalues increased by approximately a factor of ten and the relative errors in the single precision, double precision and mixed precision runs increased by a factor of ten for each succeeding experiment.

4. The relative errors in the mixed precision eigenvalue solutions, as compared to the double precision solutions, were approximately $10^{-8}$ for the dominant eigenvalues and approximately equal to the single precision inversion errors for the subdominant eigenvalue.

5. There was neither a significant nor a consistent difference between the magnitudes of the errors in the mixed and the single precision inversions.

Returning to the question of double precision representation, as a standard for comparison of single precision matrix inversions and eigensolutions, one may now face the frequently raised question: How might one justify the requirement for high-digit representation and manipulation of a matrix representing a system which is itself not defined to a high number of digits. Specifically, for this two-spring system, it is ridiculous to consider an eight-digit spring constant, let alone a sixteen-digit spring constant for any physical spring. For the case where the numerical error became serious as the mismatch became pronounced, i.e., $K_a$ greater than $K_b$, an inspection of the stiffness matrix and its inverse clearly indicated the nature of the problem. $K_a$ dominates the stiffness matrix. If it is much greater than $K_b$, it will have to be entered consistently in the matrix to more digits than have physical meaning, or $K_b$ will be lost in the system noise.
On the other hand, it is $K_a$ which, through its reciprocal, must dominate the inverse. One could hardly say that there is little physical significance to $K^{-1}$ simply because $K_a$ is not known to some great precision. What we see is that a matrix is essentially a defective device for storing information. Information which is vital to the flexibility matrix is subject to being lost in the stiffness matrix and vice versa. More generally, the stiffness properties of a member enter the stiffness matrix at the diagonals representing the joints which are joined by the member and at the corresponding coupling terms. Errors in the physical parameters will have corresponding effects on the upper modes of the stiffness matrix. However, if these stiffness terms are not entered consistently at those places where they occur in the stiffness matrix, this inconsistency will have the physical effect of a spring to ground, i.e., a weak support condition. Now, support conditions have a major effect on the lower modes of the stiffness matrix.

On the other hand, the flexibility matrix is dominated by terms which represent integrations over the entire structure. Integration is a smoothing operation, and hence, local minor physical errors would tend to be smoothed out in the upper modes of the flexibility matrix. The lower modes of a flexibility matrix would tend to reflect local patterns and, incidentally, represent that part of the structural model which normally contains the poorest physical information in any event. Therefore, one should expect that the flexibility matrix would be a preferred tool for structural computation since it emphasizes the physically meaningful modes by holding them as upper modes. Unfortunately, one cannot directly obtain the flexibility matrix for structures of interest and, as is well known, the choice of redundant cuts for the flexibility matrix method is fraught with great numerical difficulties.

Before leaving this subject, the question could be asked: Is the analysis valid in the real number system? If the answer is yes, then one must see that an artificial number of digits may be necessary to protect the information from truncation by less meaningful upper modes.

Finally, consider a hypothetical system of two springs in which $K_a$ is of order $10^3 \cdot K_b$. Let us also assume that $K_a$ is only known to two digits whereas $K_b$ is a superb spring constant known to four digits. Now the stiffness matrix does not represent the physical system to more than two digits. However, if the system is represented and solved to sufficient precision, it is clear that the flexibility matrix can be determined to four good digits because it is dominated by $K_b$. Thus we start out with two digits, lose three more en route and wind up with four digits. Obviously the limitation on the meaningful precision in the inverse is dependent upon the precision which dominates the lower modes, not on the precision which dominates the upper modes. The precision requirement for the matrix and the manipulation is a requirement of consistency.
One of Bauer's theorems (Reference 4) has received wide circulation in Forsythe and Moler, Reference 2 (p. 38), Computer Solution of Linear Algebraic Systems, which is most instructive at this point. The gist of the theorem is that if one preconditions or equilibrates the matrix $A$ by pre- and post-multiplication by diagonal matrices

$$D_1 A D_2 = \tilde{A}$$

and if $D_1$ and $D_2$ are integer powers of the computer number base (i.e., powers of two for a binary machine) and if one solves the set of equations

$$A x = b \quad \text{and} \quad \tilde{A} y = b$$

using Gaussian elimination with the same successive choice of pivot elements, then the floating point mantissas of the vector $x$ will be identical to those in the vector $y$.

In general, $D_1$ is not $D_2^{-1}$, and therefore, this is not a similarity transformation. Hence, in general, the eigenvalues and the conditioning number of the matrix $A$ will be changed by the transformation. We interpret this to mean that such a transformation permits an improvement in the measure which the conditioning number provides for truncation of information.

In particular, consider the transformation

$$
\begin{bmatrix}
\frac{1}{\sqrt{K_a}} & 0 \\
0 & \frac{1}{\sqrt{K_a + K_b}}
\end{bmatrix}
\begin{bmatrix}
K_a & -K_a \\
-K_a & K_a + K_b
\end{bmatrix}
\begin{bmatrix}
\frac{1}{\sqrt{K_a}} & 0 \\
0 & \frac{1}{\sqrt{K_a + K_b}}
\end{bmatrix}
$$

with eigenvalues

$$\lambda_1 = 1 + \sqrt{\frac{K_a}{K_a + K_b}}$$
$$\lambda_2 = 1 - \sqrt{\frac{K_a}{K_a + K_b}}$$
and with eigenvectors

\[ \mathbf{u}_1 = \left[ \frac{\sqrt{2}}{2}, \frac{\sqrt{2}}{2} \right] \]

\[ \mathbf{u}_2 = \left[ -\frac{\sqrt{2}}{2}, \frac{\sqrt{2}}{2} \right] \]

Now notice if \( K_a \gg K_b \lambda_1 \rightarrow 2 \lambda_2 \rightarrow 0 \) correctly signalling severe ill-conditioning. On the other hand, if \( K_b \gg K_a \lambda_1 \rightarrow \lambda_2 \rightarrow 1 \) correctly signalling no ill-conditioning.

This phenomenon, the partial separation of the eigenvectors into disjoint standard subspaces, is really the same thing Lanczos (Reference 5) referred to as "artificial ill-conditioning" as long ago as 1956.

Reference 3 (p. 13) notes the effect of dimensional units on the stiffness matrix conditioning numbers. For their example, the preconditioning technique just delineated, again reveals the problem to be perfectly conditioned. They noted at that time that such methods as conjugate gradient might be hampered by what we shall now call artificial ill-conditioning. Their speculation has been confirmed in Reference 6 (p. 36) who indeed found that the conjugate gradient method was much more sensitive in this respect than was Gauss-Jordan. They also found that preconditioning improved the estimate of digit loss provided by the log of the conditioning number.

Another instructive spring problem may be introduced at this point. Consider the spring system illustrated below.
The stiffness matrix for this system may be written as

\[
K(\theta) = \begin{bmatrix}
\cos \theta & -\sin \theta \\
\sin \theta & \cos \theta
\end{bmatrix}
\begin{bmatrix}
K_1 & 0 \\
0 & K_2
\end{bmatrix}
\begin{bmatrix}
\cos \theta & \sin \theta \\
-\sin \theta & \cos \theta
\end{bmatrix}
\]

Thus:

\[
K(0) = \begin{bmatrix}
K_1 & 0 \\
0 & K_2
\end{bmatrix}
\]

whereas

\[
K(\frac{\pi}{4}) = \begin{bmatrix}
\frac{K_1 + K_2}{2} & \frac{K_1 - K_2}{2} \\
\frac{K_1 - K_2}{2} & \frac{K_1 + K_2}{2}
\end{bmatrix}
\]

Now suppose \( K_1 \) is much less than \( K_2 \). The favorable choice of coordinates, with \( \theta = 0 \) involves no numerical error. The problem is represented in principal coordinates. Notice, however, that the unfavorable choice of coordinates, \( \theta = \frac{\pi}{4} \), involves the truncation of part of the information about the smaller spring constant \( K_1 \). Of course, changing the orientation, i.e., the value of \( \theta \), amounts to an orthogonal transformation of the stiffness matrix. What we see here is a demonstration of the fact that a matrix may contain full information about widely separated eigenvalues, provided they are associated with eigenvectors of disjoint standard subspaces. However, when this disjoint character of the standard subspaces is disturbed, information about the smaller eigenvalue is lost. Notice that this information can not be recovered by performing the inverse orthogonal transformation. This is a fundamentally different view of error analysis than is commonly held. Recall that an orthogonal transformation is a similarity transformation and hence, by theorem, holds the eigenvalues invariant. It is customary to point out that orthogonal transformations are very desirable because of the above fact plus the fact that the Euclidean lengths of the column vectors are also unchanged by orthogonal transformations. Now, we hold, that the matrix \( K(0) \) and the matrix \( K(\frac{\pi}{4}) \) can be written to equal precision and that \( K(\frac{\pi}{4}) \) can be derived from \( K(0) \) by the above orthogonality transformation without increasing the relative errors in the terms of the matrices. However, at the same time the information about the inverse of \( K \) or the lower eigenvalues is lost. Presently, we shall show how such a mechanism does indeed occur, in some problems, with the Householder-Givens method.

However, we should observe at this point that a favorable choice of coordinates may be most important in the formulation of stiffness matrices. In particular, the principal rigidities
of members may be greatly different. For some structures, it is possible to choose a coordinate system which is parallel or nearly parallel to the member principal rigidities. This was pointed out by Melosh and Palacol (Reference 7).

Our experiment involved some investigation of this phenomenon of artificial ill-conditioning and the separation of the eigenvectors into disjoint standard subspaces. In the first place, the stiffness matrix program which was used to generate the matrices is a plane frame program. Each joint is assigned three degrees of freedom; an $X_1$ coordinate, perpendicular to that an $X_2$ coordinate, and a rotation in the plane of the frame. In the full stiffness matrix the degrees of freedom are then ordered by joint number so that there is a cyclical pattern of degrees of freedom in the $X_1$ direction, the $X_2$ direction, and a rotational degree. Completely disjoint standard subspaces were arranged for beam problems by assigning the $X_1$ and $X_2$ directions parallel to the principal axes of the beam. For comparison, the same beam problems were solved with the coordinate axes rotated by an angle of $\pi/4$. This destroys the disjoint basis. The eigenvalues and eigenvectors were determined and printed out. In addition, a graphical display of the magnitudes of the eigenvectors components was determined and printed out. This display is illustrated below for the case of a one-inch bar treated as a simple beam with equally spaced joints at six stations including the supports.

```
10110110110111
01001001001000
10110110110111
10110110110111
01001001001000
10110110110111
10110110110111
01001001001000
10110110110111
10110110110111
01001001001000
10110110110111
10110110110111
01001001001000
10110110110111
10110110110111
10110110110111
```

In this illustration each row of the diagram represents an eigenvector. The absolute values of the eigenvector elements was tested against $10^{-4}$. Those which were greater than the test figure were represented on the diagram with a numeral one. The others were represented as zeros. For this case, it is clearly seen that the vectors did indeed separate into disjoint standard subspaces.
Inspection of the vector components indicated that the criterion of \(10^{-4}\) was completely arbitrary. For the completely disjoint subspaces solved with a Jacobi routine the vector components were true zeros. However, for most of the matrices investigated in the study, there was a pattern of reduced components in one extreme eigenvector corresponding with the larger components at the opposite extreme eigenvector. Because of this, the logarithm of the ratio of extreme eigenvalues overestimates the number of digits which are lost in the representation of the matrix. The effect of equilibration is to compact the eigenvectors of the transformed matrix and to simultaneously reduce the ratio of extreme eigenvalues.

We sought a quantitative measure for this partial separation into disjoint standard subspaces which led to experimentation with a new measure of matrix conditioning. For each pair of eigenvectors in the matrix, the dot product of the absolute values of the eigenvectors was found. The logarithm of this was added to the absolute value of the logarithm of the ratio of the corresponding eigenvalues. These new conditioning measures were formed for all pairs of vectors and the maximum was selected as a measure of the number of lost digits. We did this for several matrices before and after equilibrating the matrices. The resulting conditioning measures were in all cases close to the logarithm of the extreme eigenvalues of the equilibrated matrix. For those new conditioning numbers formed with the equilibrated matrices, the new conditioning numbers were always less than those formed from the extreme eigenvalues of the equilibrated matrix, but they were trivially less. We do not foresee the use of this conditioning measure for other than research purposes, but we do feel the need for further experimentation with it to convince ourselves that equilibration is providing a satisfactory optimum conditioning measure.

A useful measure of an approach to a scientific problem is its ability to provide new knowledge and insights. One of the major, practical new findings of this study is a weakness in the Householder-Givens method for the solution to the eigenvalue problem. This finding arises directly from consideration of the conditioning of disjoint standard subspaces.

Consider a symmetric matrix \(A\) whose columns, and hence rows, belong to two disjoint standard subspaces \(M_1\) and \(M_2\). Consider also a vector \(V_1\) belonging to \(M_1\). The property "belonging to \(M_1\)" is invariant over the transformation \(AV_1\). Thus the eigenvectors of \(A\) may be found in the same disjoint standard subspaces as the columns and rows of the matrix \(A\). In the typical orthogonal transformation of the Jacobi method, the fact that the \(i\)th diagonal element and the \(ij\) element are nonzero implies that the \(i\) and \(j\) rows or columns of the matrix belong to the same standard subspace. From this, it follows that the transformed matrix must again have rows and columns which belong to the same.
standard subspaces as before the transformation. Since the Jacobi method consists of a set of such transformations, it is clear that at no time in the process is the disjoint nature of the matrix columns or rows disturbed.

On the other hand, for the Householder-Givens method, there are conditions under which this disjoint pattern is disturbed. The typical transformation matrix of the Householder-Givens method is $\begin{pmatrix} I - 2 \mathbf{u}_k \mathbf{u}_k^t \end{pmatrix}$ where $\mathbf{u}_k^t = (0, 0, \ldots, a_{k+1}, k + s, a_{k+2}, k \ldots \ldots)$ and $s^2 = \| a_k \|^2 - a_k^2, k$. In the event that the standard subspace to which the $k^{th}$ column belongs is marked by the presence of a zero in the $k+1^{st}$ element, and that some element of the $k^{th}$ row below the $k+1^{st}$ element is nonzero, $s$ will be nonzero and $\mathbf{u}_k$ will not belong to the same standard subspace as the $k^{th}$ column of $A$. This condition was realized in our beam problems with the coordinate systems oriented along the principal axes of the beams. As a result, the eigenvector components which were known to be zeros were assigned nonzero values. Perhaps more serious, however, was the fact that when the problem was artificially ill-conditioned (i.e., the dominant eigenvalue of one subspace was much greater than the sub-dominant eigenvalue of the other subspace), the lower eigenvalue of the matrix could not be obtained with single precision Householder-Givens, even when it was readily obtained by single or double precision Jacobi analysis or double precision Householder-Givens. In some cases this effect was so serious that for the lower eigenvalue, the single precision Householder-Givens method failed by two orders of magnitude while the single precision Jacobi method obtained several good digits, as compared to the double precision solutions. In these cases, the log of the conditioning number obtained from the Householder-Givens method was computed to be approximately six, indicating that there are approximately six digits in single-precision 360 floating point arithmetic. This last comment is significant because it points out that clues of the failure of the routine are output with the proposed solution. Because the Householder-Givens routine is so much faster than the Jacobi method, we anticipate continuing to use it in double precision work.

Two considerations attest that, although the example described is contrived, the weakness of the method is nevertheless significant. The effect was noted in matrices with partially disjoint subspaces and the effect was again noted in production computation with dynamics matrices.

This is a very suggestive example. One sees that it is not sufficient to tell a customer, as does Ortega (Reference 8) in his excellent description of the Householder-Givens method, that the error in eigenvalues is less than or equal to $\| E \|$ where $\| E \|$ is the norm of the perturbation matrix of $A$. The lower eigenvalue may itself be much smaller than $\| E \|$ and
still be contained in the finite representation of the matrix. Further, the genius of an Argyris (Reference 9) in his fuselage book brings to light methods of representing some matrices in at least partially disjoint standard subspaces. One's attention is also focused on the effect of matrix routines on disjoint subspaces and one wonders if this may not be a larger factor in the weakness of numerical techniques than is the number of arithmetic operations.

SECTION III
DESCRIPTION OF THE EXPERIMENT

An existing double precision, plane frame stiffness matrix program was used as a basic tool for this experiment. Small programs were written to generate card input data for this basic program. The stiffness matrix program was then modified to write on tape the double precision stiffness matrices and their numerically determined double precision inverses. This was arranged so that a large number of cases could be handled in one computer run. A set of third-level programs was then written to read the tapes generated in the stiffness matrix program to perform the numerical experiments desired. In these third-level programs, error in the matrices was studied as follows:

1. The stiffness matrices were read from tape and truncated to single precision.

2. Using any of several test inversion routines, the stiffness matrix was inverted.

3. The results of the double precision inversion, accomplished in the stiffness matrix program, were read from tape.
4. Systematically, through the row and column indices, the double precision inverse element, the single precision inverse element, the difference between the two, and the relative error of the single precision element, were printed out.

5. The mean, standard deviation and maximum value of the absolute value of the difference between the double and single precision inverse elements were determined.

6. The same determinations were made for the relative errors.

7. A frequency distribution of the relative errors was determined.

8. The results of 5, 6, and 7 were printed out.

The eigensolutions were handled in a somewhat similar manner. For double precision eigensolutions the stiffness matrix was not truncated. The output included the graph of the eigenvectors, previously described, the eigenvalues and eigenvectors and a frequency distribution of the eigenvalues. Automated comparisons between the single and double precision eigensolutions were not obtained.

SUBROUTINES FOR MATRIX INVASION

The subroutines used for matrix inversion in the investigation were as follows:

1. Subroutine DINV, a double precision, Gauss-Jordan inversion with full pivoting which automatically determines and prints out:

   a. The Holder 1 and \( \infty \) and Schur norms for the matrix and computed inverse and the associated conditioning numbers.

   b. The Argyris conditioning measure of the matrix and the computed inverse.

   c. The successive pivot elements found during the inversion.

   d. The Turing conditioning number

   e. The trace of the matrix and of the computed inverse
f. The logarithm of the determinant

g. The mean and geometric mean eigenvalues of the matrix and of the computed inverse.

2. Subroutine INV which is a single precision version of DINV.

3. Subroutine CINV which is a single precision Cholesky square root routine. This routine did not make use of double precision accumulation of dot products, nor did it use a double precision square root routine.

4. Subroutines DECOMP, SOLVE, IMPRUV, and SING are a family of routines published in Reference 2 (p. 68). Together they perform Gaussian elimination with scaling and partial pivoting and include an iterative scheme for improving the solution of simultaneous equations by minimizing the residual vector. The results of the Gaussian elimination from DECOMP and the iterated improvement from IMPRUV were analyzed separately.

5. The same subroutines listed in item 4 were modified as follows:

   Statement 4, page 70, ibid, was altered to read:

   \[ 4 \quad \text{SUM} = \text{SUM} + \text{DBLE(A(I,J))}*X(J) \]

   The alteration was the introduction of the DBLE function which was found necessary to accomplish double precision accumulation of SUM on the IBM 360. In the second modification double precision accumulation for the word SUM was added to subroutine SOLVE.

6. Subroutine SIR, a second Gauss–Jordan subroutine will full pivoting furnished by another organization.

7. Subroutines GAUB and MULIBD which perform Gaussian decomposition without pivoting for band matrices.

EIGENVALUE ROUTINES EMPLOYED

1. A crude power method used in the stiffness matrix program for determination of conditioning information.
2. Subroutine GIVHO, a Householder-Givens routine presented by Ortega (Reference 8). This routine was altered to normalize eigenvectors to a Euclidean length of one. The routine includes an inverse power method for the determination of the eigenvectors of the tridiagonal matrix developed during the computation. This inverse power method is organized to perform one and one-half iterations. This was changed to perform two and one-half iterations as a part of the experiment.

3. DGIVHO, a double precision version of GIVHO, was also used.

4. JACOB, a conventional single precision threshold Jacobi method eigenvalue routine.

5. DAKE, a modified threshold Jacobi eigenvalue routine. The modification consisted of testing off-diagonal elements against their corresponding diagonal elements rather than an absolute test for selecting elements to rotate to zero.

A simple-minded equilibration scheme was used in the following manner: Diagonal matrices were formed by determining the reciprocals of the double precision square roots of the diagonal elements of the single precision stiffness matrices. The stiffness matrices were then pre- and post-multiplied by these diagonal matrices. To determine the effect of this operation on the inversion process, the resulting stiffness matrix was inverted using subroutine INV and the result again multiplied by the same diagonal matrices. To determine the effect upon conditioning, the equilibrated stiffness matrices were rotated to principal axes using subroutine JACOB.

The experimental sample consisted of many beam-columns and a few frames characterized as follows:

Beam boundary conditions included simple, cantilever, and fixed end. The beams were selected as 36 in. WF 245 lbs, 24 in. I 120 lbs, 10 in. WF at 54 lbs, 10 in. I at 35 lbs and 1 in. sq rod. Various types of rigidity mismatches were included. The frames were all rectangular frames consisting of the same horizontal and vertical structural sections. The number of degrees of freedom varied from fifteen to seventy-five. Shear energy was included in some but not all of the models. In excess of 110 models were used.
SECTION IV

EXPERIMENTAL RESULTS AND CONCLUSIONS

The basic data of the inversion portion of the experiment is the matrix of errors determined by subtracting a candidate, single or mixed precision, numerically obtained inverse from a double precision numerically obtained inverse used as a standard. This matrix of errors, along with the double precision standard matrix, the single precision candidate inverse, and a matrix of relative errors, were printed out term by term. Where the double precision flexibility matrix contained exact zero elements, division by zero was avoided and the relative error for that term was set to zero. Inspection of the results showed that the exact zeros were not destroyed by any of the inversion routines investigated.

The authors feel strongly about the value of printing out this primary data of the experiment. It was immediately clear from inspection of the data that the maximum relative errors were consistently associated with relatively small terms in the flexibility matrices. No case was seen in which the largest relative error was associated with the largest absolute error. For some cases, particularly the frames, the flexibility matrices contained some very small terms. In such cases, the maximum relative errors were several orders of magnitude larger than the average relative errors. From the data, one could also generalize that the larger relative errors for a particular structure tended to occur at the same place in the inverse matrix, regardless of the algorithm employed. There was a remarkable tendency for the relative errors to be of one algebraic sign, usually positive, and this was usually, but not always, independent of algorithm. The significance of this finding is not yet known, but the finding has been made and the data are available for later reinspection when the significance might contribute to our understanding of the numerical error process. Perhaps, most important, is the fact that our preconceived notions, of which we have many, are not permitted to prevent our seeing the errors as they did occur.

Of course, the examination of a half-million error terms is a tedious task, no matter how motivated the investigator. Further, a human being is a remarkably unsatisfactory device for reliably selecting the largest number from a really long list. Accordingly, summaries of the error data were accomplished by the computer and also printed out. The summaries used included the largest, the average, and the standard deviation of the absolute values of the errors and the absolute values of the relative errors were determined and printed out. The absolute values of the relative errors were classified into twenty classes, the smallest being less than $10^{-9} \times \sqrt{10}$, and the largest, greater than one. For each of the
twenty classes, the upper class limit was determined as the $\sqrt{10}$ times the lower class limit.

For structures with completely disjoint, standard subspaces, containing the columns of the matrices, the frequency distributions were trimodal. The class of smallest relative errors included the zeros of the "coupling" matrices. There was then a point of accumulation for the errors associated with the longitudinal degrees of freedom, and above this, a point of accumulation for the errors associated with the transverse and rotational degrees of freedom. Depending upon the relative stiffness of the beams, longitudinally and transversely, these distributions about the upper two accumulation points were sometimes completely separated and sometimes merged. However, for the cases observed the trimodal patterns were still clear. When the disjoint subspaces were disturbed by the less favorable choices of coordinates, the relative error frequency distributions became roughly, symmetric, “bell-shaped” functions. The trimodal character disappeared, in such cases all summary measures of the error usually increased.

Had we foreseen the extent to which maximum relative error would be affected by the variation in size of the elements in the inverse matrix for the structures being studied, we would most surely have included norms of the matrices of errors in our error summaries. We shall certainly do this in the future. Then, it is very likely that somewhat crisper statements can be made about error magnitudes in terms of “normed relative errors” than we are able to make from this experiment. However, in a sense, this would be passing some of the problem to the user of our results. It is unlikely that a clean-cut bound for the normed relative error is really more useful in judging the quality of a matrix inverse than is a somewhat fuzzier description of the behavior of the relative errors themselves, along with qualitative comments about the effect of matrix element size on mean errors versus maximum errors. Certainly, the norms of the error matrices will not replace the display of the errors themselves.

The comparison of inversion algorithms produced the following results:

1. Stiffness matrices for forty-three of the experimental models were subjected to inversion by several subroutines. For seven of these models, the spread in relative error indicated substantially more than one-digit variation in the significance of the inverses. (That is, in six out of seven cases, the spread between the best and the poorest algorithm performance was about one digit or less.)
2. Subroutine DINV, operating on a single precision stiffness matrix representation, and subroutine DECOMP with IMPRUV, using the DBLE function, were usually of about equal quality and usually better than the other subroutines. That is to say, mixed precision tended to be better than single precision. But, there were a few exceptions. In particular, there was a model for which subroutine IMPRUV "blew up."

3. As noted by Forsythe and Moler, it is essential that subroutine IMPRUV use double precision accumulation for the calculation of residuals. When the DBLE function was not used, IMPRUV increased the error as frequently as it decreased it.

4. Subroutine DECOMP, using scaling and partial pivoting in Gaussian decomposition, was usually very close to subroutine GAUB, the band matrix routine using Gaussian decomposition without pivoting or scaling. There were exceptions to this, but the exceptions did not include the case where DECOMP produced such a poor solution that IMPRUV blew up.

5. Equilibration was used with subroutine INV and usually reduced the error, compared to subroutine INV without equilibration. However, there were exceptions to this. Equilibration was not used with other inverse routines.

6. Double precision accumulation of dot products in subroutine SOLVE, used with DECOMP, had a consistently trivial effect.

7. With one exception, the seven cases in which the digit spread between the various algorithms was substantially more than one digit, the matrices being inverted exhibited considerable artificial ill-conditioning. On the other hand, not all of the artificially ill-conditioned matrices showed this variation in precision with different algorithms.

8. In those seven cases in which the digit spread was essentially greater than one, DECOMP, GAUB, IMPRUV, without the DBLE function, and INV were almost consistently found among the poorer routines. On the other hand, DECOMP was the best routine in one of these cases.

9. For those cases in which conditioning measures were obtained for equilibrated matrices, the effect of equilibration on the conditioning measure was to reduce it by an amount varying from approximately zero to about four. The new conditioning measure was usually also lower than the original conditioning number, but not as low as the equilibrated measure. The exponent of the mean, relative error for a good subroutine, say IMPRUV with
double precision accumulation, may be estimated by subtracting the logarithm of the equilib-
ibrated conditioning number for seven. Since the floating point word length of the IBM 360 is
about 6.1 digits, this is seen to be about a one-digit conservative estimate, based on the
concept of truncation of principal idempotents. In twenty-six cases this was most frequently
"on the button." It was off by more than one digit only once.

Several conclusions and tentative conclusions are drawn from these results. From
the performance of the mixed precision routines, DINV and IMPREV, it is clear that the
single precision routines used are not preserving quite all of the information stored in the
stiffness matrices. Because, in six out of seven cases, this amounted to a digit or less, it is
concluded that this is a second-order effect. The evidence suggests that this should be con-
sidered from the point of view of the possible coupling of disjoint standard subspaces.

The effect of partial, double precision should definitely be investigated in connection
with the Cholesky method. For IBM 360 users, the idea of mixed precision for structural
problems is not, in itself, particularly interesting because the single precision word of
the IBM 360 is not adequate for structural work. The effect of equilibration on conditioning
number is about as conclusive as could be hoped for, given the inherent bias of the experiment.
However, the effects of equilibration on the inversion process have not been adequately
investigated. In particular, a less simple-minded equilibration scheme, i.e., "beta equivalent
transformations, according to Reference 2 (p. 38), should be tried. This scheme should be
applied with other algorithms. Although various machines and systems were used at different
points in the research, the work reported so far in this section was all accomplished on the
IBM 360. Thus, this work is subject to the particular bias of the hexadecimal normalization
of the IBM 360 floating point word. This bias should be corrected with further work.

The comparison of the eigenvalue routines has been partially covered in Section II of
this paper. There it was shown that the Householder-Givens routine has such weakness
that it can destroy favorable basis choices. As noted, this can become serious when the
matrix has substantial artificial ill-conditioning. The experimental evidence supporting our
analysis of the Householder-Givens behavior consisted of two parts. On the one hand, one
could observe the destruction of the exact zeros in the eigenvectors which were known to
belong to truly disjoint standard subspaces. In some cases these zeros became numbers as
large as \(10^{-2}\), which in a vector scaled to a Euclidean length of one is substantial. On the
other hand, there were comparisons with eigensolutions from subroutine JACOB, subroutine
DAKE, and the double precision Householder-Givens routine. That the single-precision
Householder-Givens routine was indeed the routine in error was further attested by the
behavior of the log of the conditioning number of the matrix. For problems with artificial ill-conditioning, it was not difficult for subroutine JACOB to produce values of the log of the conditioning number as high as 8.5 or 9. For the single precision Householder-Givens routine, the observed log of the conditioning numbers varied from about 5.8 to 6.5. This is consistent with a computer word length of from 19 to 22 bits. The floating-point word length of the 360 is 24 bits. However, due to the hexadecimal normalization of floating point words, the minimum effective word length is 21 bits. For the double-precision Householder-Givens routine the destruction of the zeros was reduced by about eight order of magnitude. The agreement of the eigenvalues between the Householder-Givens double precision routine and the single precision Jacobi routine, suggested that the double precision Householder-Givens was the more accurate of the two.

The major numerical weakness of the threshold Jacobi method was not demonstrated during this study, but we have observed it in practice. For very severe, artificial ill-conditioning it is possible to obtain such a large value for the original threshold that it is not possible to assign a proper number of threshold reductions a priori. For this reason, it is felt that the Jacobi method should always include the determination of Gershgorin radii for each of the eigenvalues as a demonstration of convergence. Close roots disturb the Jacobi method because the determination of an angle of rotation depends upon the difference between the affected diagonal elements. It has been found that the computation of Rayleigh quotients, using the eigenvectors calculated by the Jacobi method, produces improved eigenvalues, except for lower eigenvalues which have been truncated.

A by-product of our investigations of the new conditioning measure was the observation that the single precision Jacobi routine used, produces a set of eigenvectors which are not normalized to the Euclidean length of one. They tend to be slightly smaller than one, the amount apparently depending upon the number of rotations required to reach principal axes. Greenstadt, (Reference 10) and J. H. Wilkinson (Reference 11) discussing the Jacobi method, point out the necessity of maintaining strict orthogonality in the eigenvalue transformation problem. It is believed that the use of partial double precision would be justified for the Jacobi routine in question. This has not been experimentally verified.

For each eigenvalue solution by a transformation method, in this experiment, a frequency distribution function for the eigenvalues was determined and printed out. It was found from this that stiffness matrices of structures, not including severe rigidity mismatches, tended to have clusters of eigenvalues toward the top of the spectrum. It seems reasonable that this would be true, in general, of sparse matrices with uniform diagonal elements. The result
of this is to make the dominant eigenvalue of such stiffness matrices surprisingly difficult to obtain by the power method. On the other hand, the stiffness matrices investigated tended to have isolated lower roots. Because of this, the lower eigenvalues were very easy to obtain by the power method, using the inverse, or, where available, an LU decomposition.

In the literature it is claimed that the Householder-Givens method requires about one-eighteenth the time of the Jacobi method. One-fifth seems to be better estimate for stiffness matrices.

We see two overriding considerations in the conditioning of stiffness matrices. Stiff elements with flexible supports lead to numerical difficulties. The separation, or partial separation, of the eigenvectors into disjoint standard subspaces, each with lower conditioning numbers than the entire matrix, provides improvement in the numerical behavior of the problem. Specifically, stiff members at supports do not cause numerical difficulties. It may be said with less certainty that flexible members supported by stiff structures do not lead to serious numerical troubles.

Considering first the case of stiff structures with flexible supports, these may be caused by physical rigidity mismatches, or they may be caused by an overly fine representation of the structures. They may also be caused by severe mismatches of the principal rigidities of the structural elements. The more complex the structure, the less the likelihood that the disjoint character can be preserved. Although we did not conduct adequate experiments along these lines, we theorize from what we did see that the selection of elements with uniform rigidities throughout the model will be inherently good modeling from the numerical point of view. Again, without specific experimental backup, we suggest that nondimensional formulations will be superior.

In the matter of disjoint standard subspaces, we should most certainly recommend that, where possible, coordinates be chosen parallel to the principal rigidities of members. However, we do see an inherent conflict between the analyst's desire to model all of the physics he can by including the types of distortions which were normally excluded in classical analysis and the numerical error problem. For example, we suspect that modeling of a shell to include radial stresses will increase the digit demand of the calculation.

It was observed that upper modes of the stiffness matrix are quite independent of the boundary conditions of the structure. It can also be readily seen that the introduction of a support to a stably supported structure will inherently improve the conditioning. This
follows from the fact that the effect of a support is to delete the corresponding row and column in the stiffness matrix. On the other hand, the bordering theorem assures use that this will decrease the upper eigenvalue and increase the lower eigenvalue.

The authors feel that they have found precious little that is new about the factors which affect the conditioning of stiffness matrices. Since all of the conclusions reached had been anticipated, it is felt, in particular, that the experimental sample was somewhat contrived. Examination of the study from this point of view suggests several remedies for future work. The most important observation is that, through this experiment, we have obtained sufficient improvement in the correlation between conditioning measures and error to warrant the widespread measurement of matrix conditioning as a quality control measure. If we are able to achieve improved communications with personnel engaged in production computations, there will be a greatly improved sample for experimental investigation available.

In the second place, surprises were found in the effects of algorithms, which attracted considerable attention in this experiment. In our preliminary work with small matrices, we had repeatedly found mixed precision solutions to be on a par with single precision solutions. Although in this experiment the effects of algorithms was of a secondary nature, it was found that mixed precision solutions were almost uniformly slightly better than single precision solutions. A small additional study is planned for the purpose of optimizing certain subroutines so that this will be a matter of less concern in future, large-scale studies.

Because of the reconnoitering nature of this experiment, automatic comparisons were not made between different structures and different subroutines. Sufficient experience has now been gained to permit this further level of automatic comparison which will greatly simplify the work.

We turn now to the consideration of some relatively small findings which are nevertheless practical and useful. In the first place, there is the lingering folklore that the size of a determinant is a useful measure of conditioning of a matrix. As is well known, the determinant of a matrix is equal to the product of the eigenvalues. On the other hand, multiplying a matrix by a scalar multiplies each of the eigenvalues by the same scalar. Hence, it multiplies the determinant by that scalar to the $n^{th}$ power. If the scalar is close enough to one so as not to cause overflow or underflow in the matrix inversion process, it will have no effect on the inversion. Thus, for most matrices of interest, it is possible to scale the matrix in such a manner as to make the determinant huge or miniscule without affecting the inversion. Further, there are families of matrices whose determinants are invariant while the matrices
range from beautifully conditioned to hopelessly ill-conditioned. We have computed with moderately ill-conditioned matrices in this study whose determinants were of order $10^{1000}$. Should this point not yet be clear, consider the corresponding flexibility matrix whose conditioning is identical to the stiffness matrix—it would have a determinant of order $10^{-1000}$.

For structures with uniform rigidities, there was a cluster of eigenvalues at the top of the spectrum. As a result, the power method turned out to be a poor way to get the dominant eigenvalue of the matrix. The Holder one norm, however, which is given by

$$\| A \|_1 = \max_j \sum_i |a_{i,j}|$$

is very easy to calculate and is a close bound to the upper eigenvalue. The dominant eigenvalue is bounded from below by the largest of $|a_{i,i}|$ and from above by $\| A \|_1$. The Schur norm is given by

$$\| A \|_E = \sqrt{\sum_i \sum_j a_{i,j}^2}$$

which can be shown to be the square root of the sum of the squares of the eigenvalues of $A$. Thus, where there is a cluster of eigenvalues toward the top of the spectrum, as in the stiffness matrix, the Schur norms will be larger than the Holder one norm. When it is smaller, this is proof that there is an isolated dominant eigenvalue indicating artificial ill-conditioning. On the other hand, the lower eigenvalues of all of the stiffness matrices observed were isolated. As a result, the Schur norm of the inverse matrix was a consistently good approximation to the reciprocal of the smallest eigenvalue of the stiffness matrix. Thus, if the inverse is calculated as part of the computation, the Schur norm is a cheap norm for the inverse. On the other hand, the fact that the lower mode is isolated, means that the power method is quite effective in the determination of the dominant eigenvalue of the flexibility matrix. This means that the measuring of the conditioning number of a stiffness matrix is not an expensive operation. Further, if the matrix is equilibrated prior to inversion or decomposition, the resulting conditioning number was, in our experiment, consistently a good measure of the loss of significance.

When a matrix is so ill-conditioned that the process has broken down, the conditioning measures obtained from the numerical representation of the matrix and the numerical inversion thereof, may indeed be small as compared to the true conditioning number. Their logarithm, however, will indicate that there is no significance in the result. That is to say, the logarithm of the conditioning measure becomes equal to, or greater than the number of digits in the arithmetic.
Of course, one of the cheapest measures of matrix conditioning is the extreme ratio of pivots in the elimination process. This number tends to be slightly smaller than the ratio of extreme eigenvalues.

In final conclusion, we have found that the major factor in matrix round-off error is the original truncation of information when the matrix is written in finite arithmetic. Among good subroutines, the effect of algorithm is secondary, but it does enter the picture. The Householder-Givens routine for the eigenvalue problem has a previously unreported weakness which does not, however, require its abandonment. Equilibration is of major importance in measurement of matrix conditioning. The dominating factor in the conditioning of stiffness matrices seems to be the ratio of element stiffness to the stiffness of the supporting structure. The accurate measurement of matrix conditioning for stiffness matrices is now an inexpensive, reliable quality control technique which should be generally used. The experimental method applied to the study of numerical error is productive and leads to definitive results of use to engineering computation. More work is needed and can be undertaken with high hope of success in the general area of conditioning of stiffness matrices.

Acknowledgment

Appreciation is expressed to the Space Division of North American Rockwell Corporation, and specifically, Dr. Leonard Harris and Dr. Robert Brown for the support of this work. In addition, grateful acknowledgment to Milton Jones for unusual help and cooperation.
SECTION V

REFERENCES


