ILL-CONDITIONED EQUATIONS

by

DONALD LEE MYERS

B. A., Washburn University of Topeka, 1962

A MASTER'S REPORT

submitted in partial fulfillment of the

requirements for the degree

MASTER OF ARTS

Department of Mathematics

KANSAS STATE UNIVERSITY
Manhattan, Kansas
1964

Approved By:

S. Thomas Parker
Major Professor

TABLE OF CONTENTS

INTRODUCTION

The purposes of this report are the examination of existing techniques of classifying a system of linear equations, and, if the system is classified as ill-conditioned, the examination of iterative methods available by use of which it is possible to obtain an accurate representation of the solution of the system.

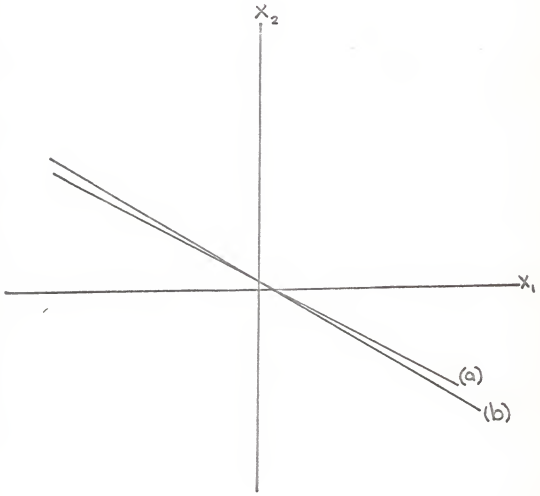
A system of equations is said to be "ill-conditioned" if it has a solution that is extremely sensitive to slight changes in the coefficients of the variables. Such sensitivity of the coefficients impairs the accuracy and dependability of the solution obtained in the usual iterative procedures, especially those involving matrix techniques, because of round-off error which is inherent in these techniques.

Many systems of equations occur as a result of experimental work. The coefficients are subject to error resulting from measurement techniques, etc. Hence it is necessary that the stability of the system be known so that appropriate precautions may be taken to insure a reasonably accurate solution of the system.

The solution of a two-dimensional system is, in a geometric sense, represented by the intersection of the graphs of the functions. As illustration, consider the set of linear equations:

$$4.001x_1+4.012x_2=0.001 \quad (a)$$

$$4.012x_1+4.014x_2=0.002. \quad (b)$$



It is obvious from the graph that a slight change in the coefficients in this system may result in a relatively much greater change in the coordinates of the point of intersection of the lines.

CLASSIFICATION TECHNIQUES

Consider the equations

$$a_{11}x + a_{12}y = d_1$$

$$a_{21}x + a_{22}y = d_2$$

whereas the lines corresponding to these equations are nearly parallel. The angle(s) between the lines must be nearly zero. To utilize this fact, one classification technique is the consideration of the angle between the lines. One can calculate $\cos \theta$ or $\cos^2\theta$ and arbitrarily decide how near unity must $\cos \theta$ or $\cos^2\theta$ be to term the system of equations ill-conditioned. Stanton (11) infers that as a reference point a criterion number could be set in the range $0.90 \pm .05$, and that ill-conditionedness may result if $\cos^2\theta$ is greater than this number. This would make θ slightly less than 18 degrees. To calculate $\cos \theta$ recall,

$$\cos \theta = \frac{\left| a_{11}a_{21} + a_{12}a_{22} \right|}{\sqrt{a_{11}^2 + a_{12}^2}\ \sqrt{a_{21}^2 + a_{22}^2}}$$

where θ is the angle between the lines with direction numbers a_{11}, a_{12} and a_{21}, a_{22}. (Recall that these are also the coefficients in the two equations.) Generalized to n dimensions this becomes:

$$\cos \theta_{ij} = \frac{\left| \sum a_{ik} \quad a_{jk} \right|}{\sqrt{\sum a_{ik}^2} \sqrt{\sum a_{jk}^2}}$$

This criterion is not wholly satisfactory because of the arbitrariness involved in choosing a critical value for $\cos^2\theta$.

One of the most widely used methods of classifying a system of equations is in terms of the value of the determinant of coefficients. If the determinant is near zero, or relatively small as compared with the determinant formed by replacing the coefficients of one of the variables by the constants, the system probably is ill-conditioned. Another way of expressing this fact involves the determinant of the normalized coefficient matrix. If the absolute value of this determinant is very small as compared with unity (i.e. $|A_N| \ll 1$) then the system is probably ill-conditioned.

Neither of the two preceding methods is entirely satisfactory by itself. As illustration, consider the set of linear equations already discussed:

$$4.001x_1 + 4.012x_2 = 0.001$$

$$4.012x_1 + 4.014x_2 = 0.002.$$

The true solution of this system is $x_1 = -1$ and $x_2 = +1$. The corresponding solution obtained by method of elimination with round-off in

the twelfth position is $x_1 = -1.00000001819$ and $x_2 = +1.00000001819$. This solution agrees quite well with the true solution considering the fact that the normalized determinant of the coefficient matrix is computed to be 0.000705 which is quite small relative to unity.

A. M. Turning (12), while considering the solution of the system AX=B, suggests that instead of using the coefficient matrix A, the matrix A-S be used where S is a small predetermined matrix which he calls an average or typical matrix. The solution of this new system will be $X_0+A^{-1}SX_0$ where X_0 is the true solution. By averaging the effect of the "average" transformation he arrives at two classification schemes, one termed the "N-condition number of A" and the other is the "M-condition number of A." The "N-condition number is equal to $(1/n)N(A)N(A^{-1})$ where n is the dimension of A and N(A) is the norm of A. The second condition number is used in conjunction with the first. The "M-condition number" is equal to $n(M(A)M(A^{-1}))$ where n is again the dimension of A and M(A) is $\max_{ij} |a_{ij}|$. The coefficients of a well-conditioned matrix give the "N-condition number" of order $n^{1/2}$ and the "M-condition number" approximately ln(n) times larger. The larger the condition numbers, the more ill-conditioned the system. However, as in the other classification procedures, this scheme has its disadvantages. For one, it is often quite tedious to compute A^{-1}. For another, it is somewhat arbitrary as to how large

the condition numbers need be, before the system is considered ill-conditioned.

Among the most popular measures of the ill-conditionedness of a system are those which employ knowledge of eigenvalues and eigenvectors of the coefficient matrix, determination of which in itself is often a challenging problem. The usual criterion is the quotient $\frac{|\lambda|_{\max}}{|\lambda|_{\min}}$. The larger the resulting quotient, the more ill-conditioned the system is. The disadvantage of this method, as previously mentioned, lies in the computation of the eigenvalues and eigenvectors.

The last classification scheme being considered in this report is attributed to Riley (9). When attempting to solve a system AX=B where A has real eigenvalues, it consists, instead, of considering the solution of the system CY=B where C=A+kI. The matrix C is "better conditioned" than is the matrix A. This will be proved by showing that if $\lambda_1, \lambda_2, \cdots, \lambda_n$ are the eigenvalues of A then $\lambda_1+k, \lambda_2+k, \cdots, \lambda_n+k$ are the eigenvalues of C.

To show that this is so, write the characteristic equation for C:

$$|rI-C| = 0.$$

It follows, successively, that

$$|rI-(A+kI)| = 0,$$

$$|rI-kI-A| = 0,$$

$$|(r-k)I-A| = 0.$$

If the eigenvalues of C are λ_i, then the eigenvalues of A are λ_i-k.

Now, without loss of generality, it is possible to assume that the eigenvalue of least numerical value is positive.

Note, $\lambda_j+k > \lambda_j$ where λ_j+k and λ_j are the minimum eigenvalues of C and A respectively. Thus $\dfrac{\lambda_j+k}{\lambda_i-\lambda_j} > \dfrac{\lambda_j}{\lambda_i-\lambda_j}$ where $\lambda_i-\lambda_j$ is positive. Hence, it follows, successively, that

$$\frac{\lambda_i-\lambda_j}{\lambda_j+k} < \frac{\lambda_i-\lambda_j}{\lambda_j},$$

$$\frac{\lambda_i-\lambda_j}{\lambda_j+k} - \frac{\lambda_i-\lambda_j}{\lambda_j} < 0,$$

$$\frac{\lambda_i}{\lambda_j+k} - \frac{\lambda_j}{\lambda_j+k} + 1 - \frac{\lambda_i}{\lambda_j} < 0,$$

$$\frac{\lambda_i}{\lambda_j+k} - \frac{\lambda_j}{\lambda_j+k} + \frac{\lambda_j+k}{\lambda_j+k} - \frac{\lambda_i}{\lambda_j} < 0,$$

$$\frac{\lambda_i}{\lambda_j+k} + \frac{k}{\lambda_j+k} - \frac{\lambda_i}{\lambda_j} < 0,$$

$$\frac{\lambda_i+k}{\lambda_j+k} < \frac{\lambda_i}{\lambda_j}.$$

Therefore

$$\frac{|\lambda_i+k|}{|\lambda_j+k|} < \frac{|\lambda_i|}{|\lambda_j|}.$$

Thus the matrix C is better conditioned than the matrix A.

Riley notes that his experience suggests that a reasonable choice is for k, a small positive constant, to be somewhere between 10^{2-s} and 10^{3-s} where s is the number of decimals being carried. X is computed from a series expansion in a matrix Y as follows:

$C=A+kI$ implies $A=C-kI$. Thus one has

$$AA^{-1}=(C-kI)A^{-1}.$$

Hence it follows successively, that

$$I=CA^{-1}-kA^{-1},$$

$$C^{-1}=C^{-1}CA^{-1}-C^{-1}kA^{-1},$$

$$A^{-1}=C^{-1}+kC^{-1}A^{-1},$$

$$A^{-1}=C^{-1}+kC^{-1}(C^{-1}+kC^{-1}A^{-1})=C^{-1}+kC^{-2}+k^2C^{-2}A^{-1},$$

Hence, $A^{-1}=C^{-1}+kC^{-2}+\cdots+k^nC^{-(n+1)}+\cdots$. Thus

$$X=A^{-1}B=C^{-1}B+kC^{-2}B+\cdots+k^nC^{-(n+1)}B+\cdots.$$

Now, letting $Y=C^{-1}B$, it follows that

$$X=Y+kC^{-1}Y+\cdots+(kC^{-1})^nY+\cdots.$$

To obtain an indication of the condition of the system, use a value of k of one in the next to last decimal place ($k=0.000\cdots010$). Y is then computed by an inversion technique such as the Cholesky or Square

Root method. Then using Y as the right member, one computes the second term in the series expansion of X; this is then used to obtain the next member of the series. If it appears that additional computation would contribute nothing, it is likely that the system is well-conditioned, providing k was appropriately chosen. On the other hand, if it appears that additional computation would contribute to the sum of the series then the system is ill-conditioned and thus the original system must have been quite ill-conditioned. The disadvantages of the method lie in the choosing of suitable k and in the series computation of X. It should be noted that in order for the condition of C to be better than that of A, k must be approximately as large as $|\lambda_i|_{min}$. However it must not be too large, lest convergence of the series be slowed down. The rate of convergence depends on k times the eigenvalues of C^{-1}.

This technique presents a method to improve an approximate solution of the system. There exist matrices A such that A^{-1} cannot be expressed exactly carrying any specified number of decimals. However, for such matrices the inverse of the corresponding C can be expressed exactly with appropriate choice of k and can be used to improve the approximate solution. For example:

Let X_0 be an approximate solution of AX=B.

Consider "$CZ_0=B-AX_0$".

Now $$C(X-(X_0+Z_0)) = CX-CX_0-CZ_0$$

$$= AX+kX-AX_0-kX_0-CZ_0$$

$$= B-AX_0-CZ_0+k(X-X_0).$$

Thus $X-(X_0+Z_0)=kC^{-1}(X-X_0)$. Now letting $X_1=X_0+Z_0$ and $Z_1=C^{-1}(B-AX_1)$, repeat the process. Continuing, one notes $Z_n=C^{-1}(B-AX_n)$. Thus if $Z_n \to 0$, one has $X \to X_n$. Convergence of Z_n is highly dependent upon the appropriate choice of k, as is seen from the following development.

If one has

$$CZ_0=B-AX_0,$$

$$CZ_1=B-AX_1,$$

$$CZ_2=B-AX_2,$$

$$CZ_3=B-AX_3,$$

$$\vdots$$

$$CZ_n=B-AX_n,$$

then

$$Z_0=C^{-1}(B-AX_0),$$

$$Z_1=C^{-1}(B-AX_1)=C^{-1}(B-A(X_0+Z_0)),$$

$$Z_2=C^{-1}(B-AX_2)=C^{-1}(B-A(X_1+Z_1)),$$

$$Z_3=C^{-1}(B-AX_3)=C^{-1}(B-A(X_2+Z_2)),$$

$$\vdots$$

$$Z_n=C^{-1}(B-AX_n)=C^{-1}(B-A(X_{n-1}+Z_{n-1})).$$

Consider:

$$
\begin{aligned}
Z_3 &= C^{-1}(B-A(X_2+Z_2)) \\
&= Z_2 - C^{-1}AZ_2 \\
&= Z_2 - C^{-1}A(Z_1 - C^{-1}AZ_1) \\
&= Z_2 - C^{-1}AZ_1 + (C^{-1}A)^2 Z_1 \\
&= Z_2 - C^{-1}AZ_1 + (C^{-1}A)^2 (Z_0 - C^{-1}AZ_0) \\
&= Z_2 - C^{-1}AZ_1 + (C^{-1}A)^2 Z_0 - (C^{-1}A)^3 Z_0 \\
&= Z_2 - C^{-1}AZ_1 + (C^{-1}A)^2 Z_0 - (C^{-1}A)^3 (C^{-1}(B-AX_0)).
\end{aligned}
$$

In general, for Z_n, one has

$$Z_n = Z_{n-1} - C^{-1}AZ_{n-2} + \cdots + (-1C^{-1}A)^{n-1} Z_0 + (-1C^{-1}A)^n (C^{-1}(B-AX_0)).$$

Now the eigenvalues of C^{-1} $(k/\lambda_i + k)$ are numerically less than one. Thus since raising C^{-1} to the nth power also raises the eigenvalues of C^{-1} to the nth power, the eigenvalues will approach zero. Recall, if the eigenvalues of a matrix are zero, then the matrix is the zero matrix. Thus an appropriate choice of k which will allow the eigenvalues of C^{-1} to approach zero rapidly, will insure convergence of the series.

MODIFICATION OF AN ILL-CONDITIONED MATRIX

The following technique is one which presents a method of improving the condition of the coefficient matrix of a system of equations. The procedure is concerned with the eigenvalues of the coefficient matrix,

principally those which are very small. Thus, since most methods for extraction of eigenvalues obtain the dominant eigenvalue first, a technique for obtaining the smallest eigenvalue is presented in Appendix A.

As illustration, it is convenient to apply the method to a four by four system where the coefficient matrix is real and symmetric. Further, for the present it is assumed that the eigenvalues are distinct and are such that $|\lambda_1|>|\lambda_2|>|\lambda_3|>|\lambda_4|$. A method attributed to J. W. Head and G. M. Oulton (2) is as follows.

Consider the ill-conditioned system AX=D where

$$A=\begin{bmatrix} a_{11} & a_{12} & a_{13} & a_{14} \\ a_{21} & a_{22} & a_{23} & a_{24} \\ a_{31} & a_{32} & a_{33} & a_{34} \\ a_{41} & a_{42} & a_{43} & a_{44} \end{bmatrix} \quad \text{with } a_{ij}=a_{ji}, \quad X=\begin{bmatrix} x \\ y \\ z \\ t \end{bmatrix} \quad \text{and} \quad D=\begin{bmatrix} d_1 \\ d_2 \\ d_3 \\ d_4 \end{bmatrix}.$$

The equation

$$\begin{bmatrix} a_{11}-\lambda & a_{12} & a_{13} & a_{14} \\ a_{21} & a_{22}-\lambda & a_{23} & a_{24} \\ a_{31} & a_{32} & a_{33}-\lambda & a_{34} \\ a_{41} & a_{42} & a_{43} & a_{44}-\lambda \end{bmatrix} \begin{bmatrix} x \\ y \\ z \\ t \end{bmatrix} = 0$$

is satisfied by the values λ_1, λ_2, λ_3, λ_4 which allow the column vector X to have a non-trival solution. If to the non-trival solution corresponding to λ_1 the additional condition that $x_1>0$; $x_1^2+y_1^2+z_1^2+t_1^2=1$

be imposed, the resulting $X_1 = \begin{bmatrix} x_1 \\ y_1 \\ z_1 \\ t_1 \end{bmatrix}$ will be a normalized eigenvector

corresponding to λ_1. This vector furnishes the "mode", $U_1 = xx_1+yy_1+zz_1+tt_1$, associated with λ_1. In a like manner the modes corresponding to the other eigenvalues can be represented.

Since these eigenvectors X_i correspond to distinct eigenvalues of a real symmetric matrix they are orthogonal. Hence the matrix $P=[X_1\ X_2\ X_3\ X_4]$ is normal, orthogonal, and unitary. Thus these modes satisfy the orthornormal properties:

1.) $x_r^2+y_r^2+z_r^2+t_r^2=1$

2.) $x_r x_s+y_r y_s+z_r z_s+t_r t_s=0$ for all $r \neq s$

3.)* $\sum x_r^2=1, \quad \sum y_r^2=1, \quad \sum z_r^2=1, \quad \sum t_r^2=1$

4.) $\sum x_r y_r = \sum x_r z_r = \sum x_r t_r = \sum y_r z_r = \sum y_r t_r = \sum z_r t_r = 0$

Then the solution of the original system is:

$$x=d_1\sum\frac{x_r^2}{\lambda_r} + d_2\sum\frac{x_r y_r}{\lambda_r} + d_3\sum\frac{x_r z_r}{\lambda_r} + d_4\sum\frac{x_r t_r}{\lambda_r}$$

$$y=d_1\sum\frac{x_r y_r}{\lambda_r} + d_2\sum\frac{y_r^2}{\lambda_r} + d_3\sum\frac{y_r z_r}{\lambda_r} + d_4\sum\frac{y_r t_r}{\lambda_r}$$

*The limit of summation on all sums ranges from one to four.

$$x=d_1\sum\frac{x_r z_r}{\lambda_r}+d_2\sum\frac{y_r z_r}{\lambda_r}+d_3\sum\frac{z_r^2}{\lambda_r}+d_4\sum\frac{z_r t_r}{\lambda_r}$$

$$t=d_1\sum\frac{x_r t_r}{\lambda_r}+d_2\sum\frac{y_r t_r}{\lambda_r}+d_3\sum\frac{z_r t_r}{\lambda_r}+d_4\sum\frac{t_r^2}{\lambda_r}\quad .$$

To show that this is so, consider the original system AX=D.

Let $X=\sum a_r X_r$ where X_r is the eigenvector corresponding to λ_r.

Also let $D=\sum b_r X_r$ where $b_j=X_j'D$. Thus AX=D becomes $A(\sum a_r X_r)=\sum b_r X_r$. This is a finite sum, hence one has $\sum a_r AX_r=\sum b_r X_r$. Thus $\sum a_r \lambda_r X_r=\sum b_r X_r$ which implies $a_i=b_i/\lambda_i$. Hence

$$X=\sum\frac{b_r X_r}{\lambda_r}=\sum\frac{X_r b_r}{\lambda_r}=\begin{bmatrix}\frac{X_1}{\lambda_1} & \frac{X_2}{\lambda_2} & \frac{X_3}{\lambda_3} & \frac{X_4}{\lambda_4}\end{bmatrix}\begin{bmatrix}b_1\\ b_2\\ b_3\\ b_4\end{bmatrix}.$$

Now let $\hat{\lambda}=\begin{bmatrix}\lambda_1 & 0 & 0 & 0\\ 0 & \lambda_2 & 0 & 0\\ 0 & 0 & \lambda_3 & 0\\ 0 & 0 & 0 & \lambda_4\end{bmatrix}$.

Then $X=\begin{bmatrix}X_1 & X_2 & X_3 & X_4\end{bmatrix}\hat{\lambda}^{-1}\begin{bmatrix}X_1 & X_2 & X_3 & X_4\end{bmatrix}'D$.

Thus $\begin{bmatrix}x\\ y\\ z\\ t\end{bmatrix}=\begin{bmatrix}x_1 & x_2 & x_3 & x_4\\ y_1 & y_2 & y_3 & y_4\\ z_1 & z_2 & z_3 & z_4\\ t_1 & t_2 & t_3 & t_4\end{bmatrix}\hat{\lambda}^{-1}\begin{bmatrix}x_1 & y_1 & z_1 & t_1\\ x_2 & y_2 & z_2 & t_2\\ x_3 & y_3 & z_3 & t_3\\ x_4 & y_4 & z_4 & t_4\end{bmatrix}\begin{bmatrix}d_1\\ d_2\\ d_3\\ d_4\end{bmatrix}$.

From this expansion it is clear that large coefficients of d_i are contributed mainly by terms having λ_4 in the denominator. The terms involving this eigenvalue can be isolated by multiplying the expression for x by x_4, the expression for y by y_4, the expression for z by z_4, the expression for t by t_4 and adding. The result is $U_4 = xx_4+yy_4+zz_4+tt_4$, $U_4=(d_1x_4+d_2y_4+d_3z_4+d_4t_4)/\lambda_4$. Head and Oulton infer that all uncertainty in the solution of the system can be attributed to the fact that U_4 is uncertain. This is a consequence of the fact that slight errors in the d_i are greatly magnified when the expression is divided by λ_4.

The procedure in the next step is to replace one of the equations of the original system by the expression for U_4. As for which equation to replace, it is suggested that if the left side of any equation is approximately proportional to the expression for U_4, then that equation is the one to be replaced. If no equation appears proportional to U_4 then the equation to replace to improve the condition of the system is arbitrary. It may be necessary to try omitting several of the original equations before the condition of the system is improved. The condition will be improved when the determinant of coefficients of the new system is greater than that of the original system. If, in the example the fourth equation is replaced by $xx_4+yy_4+zz_4+tt_4=U_4$, the

determinant of the coefficient matrix will be $\lambda_1\lambda_2\lambda_3 \begin{vmatrix} x_1 & x_2 & x_3 \\ y_1 & y_2 & y_3 \\ z_1 & z_2 & z_3 \end{vmatrix}$.

To show that this expression is the determinant of the new system,

let $B=\begin{bmatrix} a_{11} & a_{12} & a_{13} & a_{14} \\ a_{21} & a_{22} & a_{23} & a_{24} \\ a_{31} & a_{32} & a_{33} & a_{34} \\ x_4 & y_4 & z_4 & t_4 \end{bmatrix}$. Consider

the matrix product $B\begin{bmatrix} X_1 & X_2 & X_3 & X_4 \end{bmatrix}$. Note, since $\begin{bmatrix} X_1 & X_2 & X_3 & X_4 \end{bmatrix}$ is unitary, its determinant is one.

Now $$BX_1=\begin{bmatrix} a_{11} & a_{12} & a_{13} & a_{14} \\ a_{21} & a_{22} & a_{23} & a_{24} \\ a_{31} & a_{32} & a_{33} & a_{34} \\ x_4 & y_4 & z_4 & t_4 \end{bmatrix}\begin{bmatrix} x_1 \\ y_1 \\ z_1 \\ t_1 \end{bmatrix}$$

$$=\begin{bmatrix} a_{11}x_1+a_{12}y_1+a_{13}z_1+a_{14}t_1 \\ a_{21}x_1+a_{22}y_1+a_{23}z_1+a_{24}t_1 \\ a_{31}x_1+a_{32}y_1+a_{33}z_1+a_{34}t_1 \\ 0 \end{bmatrix}.$$

Thus $BX_1 = \lambda_1 \begin{bmatrix} x_1 \\ y_1 \\ z_1 \\ 0 \end{bmatrix}$. A similar procedure

confirms that $BX_2 = \lambda_2 \begin{bmatrix} x_2 \\ y_2 \\ z_2 \\ 0 \end{bmatrix}$ and $BX_3 = \lambda_3 \begin{bmatrix} x_3 \\ y_3 \\ z_3 \\ 0 \end{bmatrix}$.

However, $BX_4 =$ $\begin{bmatrix} a_{11}x_4+a_{12}y_4+a_{13}z_4+a_{14}t_4 \\ a_{21}x_4+a_{22}y_4+a_{23}z_4+a_{24}t_4 \\ a_{31}x_4+a_{32}y_4+a_{33}z_4+a_{34}t_4 \\ x_4x_4 +y_4y_4 +z_4z_4 +t_4t_4=1 \end{bmatrix}$

Hence if the determinant of $B\begin{bmatrix} X_1 & X_2 & X_3 & X_4 \end{bmatrix}$ is expanded in terms of the elements of the fourth row the indicated expression is obtained. Similar expressions are obtained when one of the other equations is replaced by the expression for U_4. These determinants have the property that

$$\begin{vmatrix} x_1 & y_1 & z_1 \\ x_2 & y_2 & z_2 \\ x_3 & y_3 & z_3 \end{vmatrix}^2 + \begin{vmatrix} x_1 & y_1 & t_1 \\ x_2 & y_2 & t_2 \\ x_3 & y_3 & t_3 \end{vmatrix}^2 + \begin{vmatrix} x_1 & z_1 & t_1 \\ x_2 & z_2 & t_2 \\ x_3 & z_3 & t_3 \end{vmatrix}^2 + \begin{vmatrix} y_1 & z_1 & t_1 \\ y_2 & z_2 & t_2 \\ y_3 & z_3 & t_3 \end{vmatrix}^2 = 1$$

since P is orthogonal and unitary. Thus not all of the determinants can be small. Consequently, the condition of the system will be improved when the equation which corresponds to one of these minors is replaced by U_4.

It should be noted that if additional data cannot be obtained to fix the value of U_4 very precisely, then useful information can be gained

with $U_4=0$.

If there exist two small eigen values then the homogeneous system of equations from which the eigenvectors are derived reduces to two equations with four unknowns. Thus a non-trival solution must be obtained in terms of two of the unknowns. The procedure is to find the non-trival solution with the condition $x^2+y^2+z^2+t^2=1$ and the additional condition that one of the unknowns be zero. Let this solution be denoted as x_3, y_3, z_3, and t_3. Then another non-trival solution is obtained with the condition $x^2+y^2+z^2+t^2=1$ and $xx_3+yy_3+zz_3+tt_3=0$. These two non-trival solutions yield the corresponding "modes":

$$U_3=xx_3+yy_3+zz_3+tt_3=(x_3d_1+y_3d_2+z_3d_3+t_3d_4)/\lambda_3$$

$$U_4=xx_4+yy_4+zz_4+tt_4=(x_4d_1+y_4d_2+z_4d_3+t_4d_4)/\lambda_4 \quad .$$

The procedure now is the same as that for the distinct eigenvalue case with the exception that two equations of the original system are replaced by the expressions for U_3 and U_4.

After employing the preceding technique for improving the condition of the coefficient matrix one may proceed to solve the resulting system of linear equations by the same method used to solve a system of well-conditioned equations.

AN ITERATIVE TECHNIQUE

This technique employs a modification of the coefficient matrix similar to that used by Riley in his classification technique. The procedure is, when contemplating the solution of the ill-conditioned system AX=D, to consider instead the system (A+K)X=D+KX where K=kI and k is a small positive constant. These two systems are mathematically equivalent, but the matrix A+K is better conditioned, as was previously shown.

The method is to delete the term KX and solve the resulting system (A+K)Y=D. The solution is denoted as $Y^{(1)}$. Thus if X is the true solution then $X-Y^{(1)}=E^{(1)}$ is the error after the first iteration. Now if one subtracts (A+K)Y=D from AX=D, the result is AX-(A+K)Y=0. Thus A(X-Y)=KY. Therefore, $AE^{(1)}=KY^{(1)}$. Hence one has obtained a new system with the same coefficient matrix A. Modifying the coefficient matrix in the manner above one then solves the system $(A+K)Y^{(2)}=KY^{(1)}$. Then the error after the second iteration is given by $E^{(2)}=E^{(1)}-Y^{(2)}=X-Y^{(1)}-Y^{(2)}$. After m iterations, $E^{(m)}=E^{(m-1)}-Y^{(m)}=X-\sum_{i=1}^{m} Y^{(i)}$.

Thus $X=\sum_{i=1}^{m} Y^{(i)}+E^{(m)}$. The formula for the determination of the $Y^{(j)}$ is

$(A+K)Y^{(j)}=D$ if $j=1$

$=KY^{(j-1)}$ if $j>1$.

If the method converges, the solution of the original system is $X=\sum_{j=0}^{\infty} Y^{(j)}$.

To obtain conditions for convergence of the series, consider:

$$Y^{(1)}=(A+K)^{-1}D$$
$$Y^{(2)}=(A+K)^{-1}KY^{(1)}$$
$$\vdots$$
$$Y^{(m)}=(A+K)^{-1}KY^{(m-1)}.$$

Therefore, let $(A+K)^{-1}D=C$ and $(A+K)^{-1}K=B$. Thus

$$Y^{(m)}=BY^{(m-1)}$$
$$Y^{(m-1)}=BY^{(m-2)}$$
$$\vdots$$
$$Y^{(3)}=BY^{(2)}$$
$$Y^{(2)}=BY^{(1)}.$$

Hence $Y^{(m)}=B^{m-1}C$. Also

$$\sum Y^{(j)}=(I+B+B^2+\cdots+B^{m-1})C.$$

If $\sum_{j=0}^{\infty} B^j$ converges then $(A+K)X=D+KX$ implies

$$X=(A+K)^{-1}D+(A+K)^{-1}KX$$
$$=C+BX.$$

Thus $X-BX=C$ and $X=(I-B)^{-1}C=\sum_{j=1}^{\infty} B^jC$.

Hence, $X-\sum_{j=1}^{m} Y^{(j)}=(\sum_{j=0}^{\infty} B^j-\sum_{j=0}^{m-1} B^j)C$

$$=\sum_{m}^{j=0} B^jC=(I-B)^{-1}B^mC$$

$$=B^m(I-B)^{-1}C=E^{(m)}.$$

Thus if $\sum_{j=0}^{\infty} B^j$ converges then $\lim_{m\to\infty} B^m=0$. Therefore $E^{(m)}=0$ and $X=\sum_{j=0}^{\infty} Y^{(j)}$.

Since the choice of k, which will insure the convergence of $\sum_{j=0}^{\infty} B^j$, is a difficult problem and then, the computation of $E^{(m)}$ can become tedious, consider an expression for the upper bound of error. This expression is developed as follows.

Define $(A+K)^{-1}=D=[d_{ij}]$. Recall $Y^{(m)}=DKY^{(m-1)}$.

$$\text{Thus } Y^{(m)}=\begin{bmatrix} y_1^{(m)} \\ y_2^{(m)} \\ \vdots \\ y_n^{(m)} \end{bmatrix}=\begin{bmatrix} d_{11} & d_{12} \cdots & d_{1n} \\ d_{21} & d_{22} \cdots & d_{2n} \\ \vdots & & \vdots \\ d_{n1} & d_{n2} \cdots & d_{nn} \end{bmatrix}\begin{bmatrix} ky_1^{(m-1)} \\ ky_2^{(m-1)} \\ \\ ky_n^{(m-1)} \end{bmatrix}.$$

Hence, $y_r^{(m)}=d_{r1}ky_1^{(m-1)}+d_{r2}ky_2^{(m-1)}+\cdots+d_{rn}ky_n^{(m-1)}: r=1,..n$.

Thus one has

$$\left|y_r^{(m)}\right| \leq \left|d_{r1} k y_1^{(m-1)}\right| + \left|d_{r2} k y_2^{(m-1)}\right| + \cdots + \left|d_{rn} k y_n^{(m-1)}\right|$$

$$\leq |k| \left|\sum_{i=1}^{n} d_{ri}\right| \left|y_m^{(m-1)}\right| \quad \text{where } y_m^{(m-1)} = \max y_i^{(m-1)}.$$

Since this holds for all r, it will hold if $\left|y_r^{(m)}\right|$ is replaced by $\left|y_M^{(m)}\right| = \left\{\left|y_2^{(m)}\right|, \ldots, \left|y_n^{(m)}\right|\right\}$.

Therefore

$$\left|y_M^{(m)}\right| \leq |k| \left\{\sum |d_{ri}|\right\} \left|y_m^{(m-1)}\right| : r = i, 2, \ldots n. \tag{c}$$

Let $d_M = \max \left\{\sum |d_{1i}|, \sum |d_{2i}|, \cdots, \sum |d_{ni}|\right\}$.

Then define a convergence constant H and restrict H so that $|k| d_M = H < 1$. This restriction can always be satisfied by choosing k sufficiently small.

From (c), for $m > 1$, one has

$$\left|\frac{y_M^{(m)}}{y_m^{(m-1)}}\right| \leq H \quad \text{provided } y_m^{(m-1)} \neq 0. \tag{d}$$

Hence the series $\sum_{m=1}^{\infty} y_M^{(m)}$ is absolutely convergent since $H < 1$.

Considering the expressions for the individual components of the matrix X one observes

$$x_i = \sum_{0}^{\infty} y_i^{(j)} = \sum_{j=1}^{m} y_i^{(j)} + \sum_{j=m+1}^{\infty} y_i^{(j)} : i = 1, 2, \ldots, n.$$

Thus the error after m iterations for each x_i is

$$e_i^{(m)} = x_i - \sum_{1}^{m} y_i^{(j)} = \sum_{m+1}^{\infty} y_i^{(j)} : i=1, 2, \ldots, n.$$

Also $\left|e_i^{(m)}\right| \leq \sum_{m+1} \left|y_i^{(j)}\right| : k=1, 2, \ldots, n.$

Now since $\left|y_i^{(m)}\right| \leq \left|y_M^{(m)}\right|$, from (d) one obtains

$$\left|e_i^{(m)}\right| < H\left|y_M^{(m)}\right| (1+H+H^2+\cdots) = \left|y_M^{(m)}\right| \frac{H}{1-H} .$$

This technique is highly adaptable to machine computation as the coefficient matrix used when obtaining a solution of the modified system is the same for each of the $Y^{(i)}$. It is presented here merely as a supplement to the usual methods for solution of systems of equations.

APPENDIX A

In the determination of small eigenvalues only those which are less than 0.2 in absolute value are being considered. This is so since if the eigenvalue exceeds 0.2, then the maximum contribution possible to any coefficient is numerically less than 5.0.

In this procedure it is assumed that the evaluation of a determinant or the solution of a polynomial in with numerical coefficients presents no great problem.

Since attention is focused on those values of less than 0.2, the technique is to evaluate the determinant of the characteristic matrix with = -0.2+0.4r/n, r=0, 1, ...n, where is possibly negative. The evaluation of these determinants givesa rough idea of the number and position of the roots.

To illustrate the procedure, consider the following characteristic matrix:

$$\begin{bmatrix} 2.557-\lambda & 2.624 & 2.468 & 2.361 \\ 2.624 & 3.493-\lambda & 2.351 & 1.532 \\ 2.468 & 2.351 & 2.557-\lambda & 2.624 \\ 2.351 & 1.532 & 2.624 & 3.493-\lambda \end{bmatrix} .$$

Since all values of λ are positive for this example, the determinant, $\Delta(\lambda)$, is evaluated with λ =0, λ =0.04, λ =0.08, λ =0.12, λ =0.16.

The corresponding values for $\Delta(\lambda)$ are

$$\Delta(0) = 0.05$$

$$\Delta(0.04) = 0.0019526$$

$$\Delta(0.08) = 0.01761476$$

$$\Delta(0.12) = 0.09101456$$

$$\Delta(0.16) = 0.2190776.$$

Now $\Delta(\lambda)$ for a system of four equations is a polynomial of degree four in λ. Thus $f(\lambda)$, defined by

$$f(\lambda) = \Delta(0)\ \frac{(\lambda-0.04)(\lambda-0.08)(\lambda-0.12)(\lambda-0.16)}{0.04\quad 0.08\quad 0.12\quad 0.16}$$

$$- \Delta(0.04)\ \frac{\lambda(\lambda-0.08)(\lambda-0.12)(\lambda-0.16)}{0.04\quad 0.04\quad 0.08\quad 0.12}$$

$$+ \Delta(0.08)\ \frac{\lambda(\lambda-0.04)(\lambda-0.12)(\lambda-0.16)}{0.08\quad 0.04\quad 0.04\quad 0.08}$$

$$- \Delta(0.12)\ \frac{\lambda(\lambda-0.04)(\lambda-0.08)(\lambda-0.16)}{0.12\quad 0.08\quad 0.04\quad 0.04}$$

$$+ \Delta(0.16)\ \frac{\lambda(\lambda-0.04)(\lambda-0.08)(\lambda-0.12)}{0.16\quad 0.12\quad 0.08\quad 0.04}$$

is identical to $\Delta(\lambda)$ when λ has these five values. Hence $f(\lambda)$ will represent $\Delta(\lambda)$ very accurately for small values of λ.

Thus

$$f(\lambda) = 0.05\,\frac{(\lambda^4 - 0.4\,\lambda^3 + 0.056\,\lambda^2 - 0.0032\,\lambda + 0.00006144}{0.00006144}$$

$$- 0.0019526\,\frac{(\lambda^4 - 0.36\,\lambda^3 + 0.0416\,\lambda^2 - 0.001536\,\lambda)}{0.00001536}$$

$$+ 0.0171476\,\frac{(\lambda^4 - 0.32\,\lambda^3 + 0.0304\,\lambda^2 - 0.000768\,\lambda)}{0.00001024}$$

$$- 0.09101456\,\frac{(\lambda^4 - 0.28\,\lambda^3 + 0.0224\,\lambda^2 - 0.000512\,\lambda)}{0.00001536}$$

$$+ 0.2190776\,\frac{(\lambda^4 - 0.24\,\lambda^3 + 0.0176\,\lambda^2 - 0.000384\,\lambda)}{0.00006144}\,.$$

This simplifies to

$$f(\lambda) = 9449.49115\,\lambda^4 - 835.28266\,\lambda^3 + 74.19092\,\lambda^2$$

$$- 3.8285\,\lambda + 0.05.$$

The technique at this point is to equate $f(\lambda)$ to zero and solve by an iterative measure for λ.

In this particular example the value of λ obtained is 0.05.

BIBLIOGRAPHY

1. Frazer, R. A.
Elementary Matrices, Cambridge University Press, 1938.

2. Head, J. W. and Oulton, G. M.
"The Solution of 'Ill-Conditioned' Linear Simultaneous Equations", Aircraft Engineering, Volume 30 (October, 1958), London: Dunhill Publications.

3. Hildebrand, F. B.
Introduction to Numerical Analysis, New York: McGraw-Hill Book Company, Inc., 1956.

4. Householder, A. S.
Principles of Numerical Analysis, New York: McGraw-Hill Book Company, Inc., 1953.

5. Kunz, K. S.
Numerical Analysis, New York: McGraw-Hill Book Company, Inc., 1957.

6. Lanczos, Cornelius
Applied Analysis, Englewood Cliffs, New Jersey: Prentice-Hall, Inc., 1956.

7. Nielson, K. L.
Methods in Numerical Analysis, New York: The Macmillan Company, 1956.

8. Perlis, Sam
Theory of Matrices, Reading, Massachusetts: Addison-Wesley Publishing Company, Inc., 1958.

9. Riley, J. D.
"Solving Systems of Linear Equations with a Positive Definite, Symmetric, but Possibly Ill-Conditioned Matrix", Math Tables and Other Aids to Computation, Washington, D. C.: The National Research Council, July, 1955.

10. Scott, E. J., Li, H. L., and Chao, B. T.
On the Solution of Ill-Conditioned Simultaneous, Linear, Algebraic Equations by Machine Computation, Urbana, Illinois: University of Illinois Bulletin, 1961.

11. Stanton, Ralph G.
Numerical Methods for Science and Engineering, Englewood Cliffs, New Jersey: Prentice-Hall, Inc., 1961.

12. Turning, A. M.
"Rounding Off Errors in Matrix Process," Quarterly Journal of Mechanics and Applied Mathematics, Volume 1, 1948.

ACKNOWLEDGMENT

The author wishes to express his sincere thanks and appreciation to Professor S. Thomas Parker for his helpful suggestions and assistance with the preparation of this report.

ILL-CONDITIONED EQUATIONS

by

DONALD LEE MYERS

B. A., Washburn University of Topeka, 1962

AN ABSTRACT OF A MASTER'S REPORT

submitted in partial fulfillment of the

requirements for the degree

MASTER OF ARTS

Department of Mathematics

KANSAS STATE UNIVERSITY
Manhattan, Kansas
1964

The purposes of this report are to define "Ill-Conditioned Equations," and examine existing methods of classifying a system of equations, develop a procedure to improve the condition of the system of equations, and present an iterative technique for use in the solution of a system of "Ill-Conditioned Equations."

A set of "Ill-Conditioned Equations" is a system of equations which has a solution that is extremely sensitive to slight changes in the coefficients of the variables. As a result, the accuracy and dependability of the solution obtained in the usual iterative procedures is impaired.

The classification techniques considered are:

1. The (generalized) angle(s) between the "lines" which represent the system of equations in a geometric sense.
2. The ratio of the determinant of the coefficient matrix to the determinant formed from it by replacing the coefficient of one of the variables by the constant.
3. The consideration of the determinant of the normalized coefficient matrix as compared with unity.
4. The use of "M-condition numbers" and "N-condition numbers."
5. The largest ratio of the absolute values of two eigenvalues.

6. A consideration of the convergence of a series expansion of the system of equations.

A procedure developed to improve the condition of the original system is to replace one of the original equations with a "mode." These modes are derived from eigenvalues and their corresponding eigenvectors. A technique for determination of eigenvalues less in absolute value than 0.2 is presented in Appendix A.

The remainder of the report deals with an iterative technique which is highly adaptable to electronic computation. An expression for the determination of the error after m iterations is also presented.