ERRORS IN SIMULTANEOUS LINEAR EQUATIONS[1]

By RAYMOND REDHEFFER (*Research Lab. of Electronics, M. I. T.*)

1. The problem. After computing the unknowns in a linear system $\sum_1^n a_{ij}\, x_i = b_j$ with non-vanishing determinant, one may check the work by substitution into the original equations. If the j-th equation is satisfied within an error e_j, it is natural to inquire into what can be said about the corresponding error in the unknowns. This question is the subject of the present note. Analytically we have

$$\sum a_{ij}(x_i + E_i) = b_j + e_j, \tag{1}$$

where x_i are the exact values of the unknowns, and where the errors E_i in the unknowns are to be estimated in terms of the substitution errors e_j. This problem was suggested by Prof. P. Franklin.

2. A solution. Regarding $\{E_i\}$ and $\{e_i\}$ as vectors, let us estimate the maximum length of one in terms of the length of the other; in other words let us find max $(\sum E_i^2)^{1/2}$ subject to $(\sum e_i^2)^{1/2} = e$. This is equivalent to finding the minimum of $\sum e_i^2$ subject to a constant value of $\sum E_i^2$, and such a problem lends itself to standard methods. Using the fact that the x_i satisfy the equations exactly, we have

$$\sum a_{ij}E_i = e_j\,. \tag{2}$$

We are thus required to minimize

$$\sum_j \left(\sum_i a_{ij}E_i\right)^2 \tag{3}$$

subject to the condition that $\sum E_i^2 =$ constant. By collecting terms in (3) we find that the matrix of the quadratic form is AA', where $A = (a_{ij})$ is the matrix of the original system and A' is its transpose. Hence, if e is the length of the known error vector, $e = (\sum e_i^2)^{1/2}$, and E is the length of the unknown error vector, $E = (\sum E_i^2)^{1/2}$, then we have[1]

$$E \leq e/(\lambda_m)^{1/2}, \tag{4}$$

where λ_m is the minimum characteristic value of the matrix AA'. When the matrix A is symmetrical, in particular, we have

$$E \leq e/|\,\lambda_m'\,|, \tag{5}$$

where λ_m' is (in absolute value) the minimum characteristic value of A. These results are the best possible, in that equality is always attained, for given e, with some set of values e_i.

3. Approximation. If λ_i are the characteristic values of AA' we know that

$$\sum \lambda_i = \sum\sum a_{ij}^2 = T, \tag{6}$$

the trace of AA'. Also

$$\Pi\lambda_i = |\,A\,|^2 \tag{7}$$

[1] Received Nov. 21, 1947. This work has been supported in part by the Signal Corps, the Air Materiel Command, and the Office of Naval Research.

[2] R. Courant and D. Hilbert, *Methoden der Mathematischen Physik*, vol. 1, J. Springer, Berlin, 1931, p. 21.

if $| A |$ stands for the determinant of A. By (3) it is clear that AA' is positive definite, so that $\lambda_i > 0$, and hence we may use (6) to find

$$\sum_{i \neq m} \lambda_i < T, \tag{8}$$

while in any case we have, by (7),

$$\lambda_m \prod_{i \neq m} \lambda_i = | A |^2. \tag{9}$$

If the sum of n positive quantities is constant the product is maximum when they are all equal. Hence we find, by (8) and (9),

$$\lambda_m > \frac{| A |^2}{T^{n-1}} (n - 1)^{n-1} \tag{10}$$

with the final result

$$E < \frac{e}{|| A ||} \left(\frac{T}{n - 1} \right)^{(n-1)/2} \tag{11}$$

which is not difficult to compute numerically.

From the derivation it is clear that this upper bound is not far from optimum, whenever λ_m is small compared to the sum of the λ_i, and in addition the other λ_i do not vary over too wide a range of values. In the case of any normal orthogonal transformation, for example, we have

$$E < \frac{e}{1} \left(\frac{n}{n - 1} \right)^{(n-1)/2} \rightarrow e(2.718 \cdots)^{1/2} \cong 1.65e$$

by (11), whereas the optimum inequality is $E \leq e$.

4. Numerical examples. Suppose that approximate values of the unknowns are substituted in a system with the matrix

$$A = \begin{pmatrix} 3 & -2 & -2 \\ -2 & 5 & 1 \\ -2 & 1 & 4 \end{pmatrix} \tag{12}$$

and are found to satisfy the equations with errors of 0.1, 0.1 and 0.2 respectively. To find the maximum possible error in any one unknown we proceed as follows. The determinant of the system A is 29, and for T we take the sum of the squares of the entries in (12), which gives $T = 68$. Substituting in (11) with $n = 3$, we find

$$E < 1.17e \tag{13}$$

In the present case $e = (.01 + .01 + .04)^{1/2} = .22$ so that E, and hence the maximum error in any one unknown, cannot exceed 0.26.

The characteristic values were found by J. G. Linvil to be 1.13, 7.39, and 3.49. Using $\lambda_m = 1.13$ in (5) we obtain the optimum inequality

$$E < 0.89e \tag{14}$$

which is only a slight improvement on (13).