SYSTEMS OF LINEAR EQUATIONS WITH COEFFICIENTS SUBJECT TO ERROR

By A. T. Lonseth

Iowa State College

1. Introduction. Various scientific problems lead to non-homogeneous systems of n linear equations in n unknowns, in which the $n^2 + n$ coefficients (including "absolute" terms) are subject to error. Such errors may be errors of observation, or errors introduced by rounding off decimal expansions. If the system has a non-vanishing determinant, the ordinary rules yield the solution. But the question arises: how may the possible errors in the coefficients affect the solutions? In particular, one would like to know how to exclude the fatal event that some malicious combination of errors might make the determinant zero. One would further like to have limitations on the solution-errors in terms of maximum coefficient-errors. Considering the coefficient-errors as random variables, one may also inquire as to the probability distributions of the solution-errors.

The principal result obtained in this paper is the Taylor's expansion of the error in any unknown, considered as a function of the $n(n + 1)$ errors in the coefficients. An upper bound is obtained for each term of this series, and the sum of these upper bounds (when convergent) is expressed in closed form. Thus are obtained not only approximations to the maximum error, but an actual upper limit. Convergence of the power series is established for sufficiently small coefficient-errors; "sufficient smallness" is specified in terms of a simple criterion, which simultaneously provides a sufficient condition for the non-vanishing of a determinant with elements subject to error.

These results were obtained before I learned that work had already been done on the problem. The earliest seems to be that of F. R. Moulton [2] in 1913; he found the first order approximation (6) for $n = 3$, and discussed the geometrical reasons for sensitivity. Much later I. M. H. Etherington [1], evidently unaware of Moulton's paper, found the expression for the total error of a determinant whose elements may be in error, and applied this to the present problem. He thus found limits for the first and second order errors, in a rather different form from mine. The probabilistic considerations of section 5 were suggested by Etherington's article. L. B. Tuckerman [3] recently discussed the question of estimating computational errors incurred in the course of solution. He considered only errors of first order.

My original procedure was to compute the terms of the Taylor's series as successive differentials of the unknown, from Cramer's formula. This soon becomes laborious, and I found only the first two terms. The linear matrix equation (4) was then kindly suggested to me by R. Oldenburger. Here (4) is solved by iteration, resulting in a simple recursion formula for successive terms of the Taylor's series.

2. Formal matrix solution. Let the system of equations be

$$\sum_{j=1}^{n} a_{ij}x_j = c_i \qquad i = 1, 2, \cdots, n. \tag{1}$$

In terms of the matrices

$$\mathbf{A} = \begin{pmatrix} a_{11} \cdots a_{1n} \\ \vdots \quad \vdots \\ a_{n1} \cdots a_{nn} \end{pmatrix}, \qquad \mathbf{X} = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}, \qquad \mathbf{C} = \begin{pmatrix} c_1 \\ \vdots \\ c_n \end{pmatrix},$$

system (1) can be written

$$\mathbf{AX} = \mathbf{C}. \tag{2}$$

Supposing that not all c's vanish, and that A, the determinant of $\mathbf{A}$, does not vanish, there is a unique solution $\mathbf{X}$. But the a's and c's, and consequently the x's, are subject to error: let the true value of a_{ij} be $a_{ij} + \alpha_{ij}$; of c_i, $c_i + \gamma_i$; and of the resulting x_j, $x_j + \xi_j$. We must actually deal with the system

$$(\mathbf{A} + \mathbf{a})(\mathbf{X} + \mathbf{x}) = \mathbf{C} + \mathbf{c}, \tag{3}$$

where we have written

$$\mathbf{a} = \begin{pmatrix} \alpha_{11} \cdots \alpha_{1n} \\ \vdots \quad \vdots \\ \alpha_{n1} \cdots \alpha_{nn} \end{pmatrix}, \qquad \mathbf{x} = \begin{pmatrix} \xi_1 \\ \vdots \\ \xi_n \end{pmatrix}, \qquad \mathbf{c} = \begin{pmatrix} \gamma_1 \\ \vdots \\ \gamma_n \end{pmatrix}$$

Expanding (3) and using (2), we find for the error-matrix $\mathbf{x}$

$$\mathbf{x} = \mathbf{m} + \mathbf{nX} + \mathbf{nx}, \tag{4}$$

with $\mathbf{m} = \mathbf{A}^{-1}\mathbf{c}$, $\mathbf{n} = -\mathbf{A}^{-1}\mathbf{a}$; $\mathbf{A}^{-1}$ is the inverse of $\mathbf{A}$. We solve (4) formally for $\mathbf{x}$ by iteration. Thus

$$\mathbf{x} = \mathbf{m} + \mathbf{nX} + \mathbf{n}(\mathbf{m} + \mathbf{nX}) + \mathbf{n}^2\mathbf{x}, \text{ etc.:}$$

and there results the infinite expansion

$$\mathbf{x} = \sum_{k=1}^{\infty} \mathbf{x}^{(k)}\colon \mathbf{x}^{(1)} = \mathbf{m} + \mathbf{nX}; \qquad \mathbf{x}^{(k)} = \mathbf{nx}^{(k-1)}, \qquad k > 1. \tag{5}$$

In section 4 convergence of (5) will be established for sufficiently small $|\alpha_{ij}|$.

3. The elements of $\mathbf{x}^{(k)}$. It is necessary to consider closely the individual elements of $\mathbf{x}^{(k)}$. Writing

$$\mathbf{x}^{(k)} = \begin{pmatrix} \xi_1^{(k)} \\ \vdots \\ \xi_n^{(k)} \end{pmatrix},$$

we note from (5) that

$$\xi_j = \sum_{k=1}^{\infty} \xi_j^{(k)} ;$$

this is precisely the Taylor's series for the error in x_j : each $\xi_j^{(k)}$ is a homogeneous polynomial of degree k in the α's and γ's. Writing A_{ij} for the cofactor of a_{ij} in A,

$$\mathbf{x}^{(1)} = \mathbf{m} + \mathbf{nX} = \mathbf{A}^{-1}(\mathbf{c} - \mathbf{aX})$$

$$= \begin{pmatrix} \frac{A_{11}}{A} & \cdots & \frac{A_{n1}}{A} \\ \vdots & & \vdots \\ \frac{A_{1n}}{A} & \cdots & \frac{A_{nn}}{A} \end{pmatrix} \left\{ \begin{pmatrix} \gamma_1 \\ \vdots \\ \gamma_n \end{pmatrix} - \begin{pmatrix} \alpha_{11} & \cdots & \alpha_{1n} \\ \vdots & & \vdots \\ \alpha_{n1} & \cdots & \alpha_{nn} \end{pmatrix} \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} \right\}$$

$$= \begin{pmatrix} \frac{A_{11}}{A} & \cdots & \frac{A_{n1}}{A} \\ \vdots & & \vdots \\ \frac{A_{1n}}{A} & \cdots & \frac{A_{nn}}{A} \end{pmatrix} \begin{pmatrix} \gamma_1 - \alpha_{11}x_1 - \cdots - \alpha_{1n}x_n \\ \vdots \\ \gamma_n - \alpha_{n1}x_1 - \cdots - \alpha_{nn}x_n \end{pmatrix},$$

whence (summing hereafter from 1 to n on Greek-letter subscripts)

$$\xi_j^{(1)} = \frac{1}{A} \left\{ \sum_{\mu} \gamma_\mu A_{\mu j} - x_1 \sum_{\mu} \alpha_{\mu 1} A_{\mu j} - \cdots - x_n \sum_{\mu} \alpha_{\mu n} A_{\mu j} \right\}. \tag{6}$$

From (5), if $k > 1$,

$$\mathbf{x}^{(k)} = \mathbf{n}\mathbf{x}^{(k-1)} = -\mathbf{A}^{-1}\mathbf{a}\mathbf{x}^{(k-1)}$$

$$= \begin{pmatrix} -\frac{1}{A}\Sigma\alpha_{\mu 1}A_{\mu 1} & \cdots & -\frac{1}{A}\Sigma\alpha_{\mu n}A_{\mu 1} \\ \vdots & & \vdots \\ -\frac{1}{A}\Sigma\alpha_{\mu 1}A_{\mu n} & \cdots & -\frac{1}{A}\Sigma\alpha_{\mu n}A_{\mu n} \end{pmatrix} \begin{pmatrix} \xi_1^{(k-1)} \\ \vdots \\ \xi_n^{(k-1)} \end{pmatrix},$$

so that

$$\xi_j^{(k)} = -\frac{1}{A} \sum_{\nu} \xi_\nu^{(k-1)} \sum_{\mu} \alpha_{\mu\nu} A_{\mu j}, \qquad k > 1. \tag{7}$$

The sums $\Sigma\gamma_\mu A_{\mu j}$, $\Sigma\alpha_{\mu l}A_{\mu j}$ have obvious interpretations as determinants.

4 Bounds and convergence of the series. Assuming $|\alpha_{ij}|$, $|\gamma_i| \leqq \delta$ and taking absolute values in (6),

$$|\xi_j^{(1)}| \leqq \frac{\delta}{|A|}(1 + \sum_{\mu} |x_\mu|)(\sum_{\mu} |A_{\mu j}|). \tag{8}$$

It will be observed that equality can be attained for a particular choice of α's and γ's as $\pm\delta$: the bound for first-order errors is best possible. But it is not in general possible by a single choice of α's and γ's to obtain equality for all j.

Similarly from (7)

$$|\xi_j^{(k)}| \leqq \frac{\delta}{|A|}\left(\sum_\mu |\xi_\mu^{(k-1)}|\right)\left(\sum_\mu |A_{\mu j}|\right), \qquad k > 1;$$

whence by induction

$$|\xi_j^{(k)}| \leqq \left(\frac{\delta}{|A|}\right)^k \left(1 + \sum_\mu |x_\mu|\right)\left(\sum_\nu \sum_\mu |A_{\mu\nu}|\right)^{k-1}\left(\sum_\mu |A_{\mu j}|\right). \tag{9}$$

Summing on k,

$$\sum_{k=1}^{m} |\xi_j^{(k)}| \leqq \frac{\delta}{|A|}\left(1 + \sum_\mu |x_\mu|\right)\left(\sum_\mu |A_{\mu j}|\right)\left(\sum_{k=1}^{m} \rho^{k-1}\right),$$

with

$$\rho = \frac{\delta}{|A|}\sum_\nu \sum_\mu |A_{\mu\nu}|.$$

If $\rho < 1$, we can let $m \to \infty$:

$$|\xi_j| \leqq \frac{\delta}{|A|}\left(1 + \sum_\mu |x_\mu|\right)\left(\sum_\mu |A_{\mu j}|\right)/(1 - \rho). \tag{10}$$

Observing that the γ's occur linearly in (6) and (7), we conclude that (5) converges if

$$|\alpha_{ij}| \leqq \delta < |A|/\left(\sum_\nu \sum_\mu |A_{\mu\nu}|\right). \tag{11}$$

It follows that the determinant of the system (3) cannot vanish if (11) holds. This is rather remarkable, in that $\delta\Sigma\Sigma\,|A_{\mu\nu}|$ is merely the maximum first-order term in the error of that determinant ([1], p. 108); the effect of higher order terms (i.e., of any but first-order minors) in producing a zero determinant can be wholly ignored.

From the remark after (8), it appears that equality in (9) and (10) cannot generally be attained.

If (10) is written $|\xi_j| \leqq B/(1 - \rho)$, it is easily seen that the remainder after the hth approximation does not exceed $\rho^h B/(1 - \rho)$.

5. Probability distributions. We now consider some consequences of the following assumptions: the α's and γ's are identical, independent random variables, bounded by a δ satisfying (11), and distributed symmetrically about zero. (It would be reasonable to assume further that they possess a frequency function, which is nowhere concave upward.) Writing $\mathfrak{E}(x)$ for "expectation of the random variable x," we have

$$\mathfrak{E}(\alpha_{ij}) = \mathfrak{E}(\gamma_i) = 0, \qquad \mathfrak{E}(\alpha_{ij}^2) = \mathfrak{E}(\gamma_i^2) = \sigma^2 < \delta^2.$$

On account of independence and symmetry, the expectation of any power-product of α's and γ's containing an odd power must be zero. To first order, the mean a_j of the solution-error ξ_j is approximated by

$$a_j^{(1)} = \mathfrak{E}(\xi_j^{(1)}) = 0; \tag{12}$$

and the standard deviation S_j by

$$S_j^{(1)} = \sqrt{\mathfrak{E}\{(\xi_j^{(1)})^2\}}\,\frac{\sigma}{|A|}\,\{(1 + \sum_{\mu} x_\mu^2)(\sum_{\mu} A_{\mu j}^2)\}^{\frac{1}{2}}. \tag{13}$$

The second approximation to a_j is also easily obtained:

$$a_j^{(2)} = \mathfrak{E}(\xi_j^{(2)}) = \frac{\sigma^2}{A^2}(\sum_{\nu}\sum_{\mu} x_\nu A_{\mu\nu} A_{\mu j}). \tag{14}$$

Both (13) and (14) were given by Etherington [1], though in a less symmetric form. Higher approximations, as he remarks, involve complicated summations; but if they should ever be required, the machinery exists in (6) and (7) for their systematic computation. As to the errors in using (13) for the standard deviation S_j and (14) for the mean, we know only that

$$a_j = a_j^{(2)} + o(\delta^4), \qquad S_j^2 = (S_j^{(1)})^2 + o(\delta^4).$$

Etherington ([1], p. 111) considers the important special case of "rounding off" decimal expressions. Each a and c is supposed correct in the qth decimal place, the $(q + 1)$th figure being "forced," i.e., increased by one when the $(q + 2)$th figure is dropped, if the $(q + 2)$th is 5, 6, 7, 8, or 9. Assuming constant frequency 10^{-q} in the interval $(-\frac{1}{2}10^{-q}, \frac{1}{2}10^{-q})$, we may use (13) and (14) with $\sigma^2 = 10^{-2q}/12$.

Errors of observation are often assumed to be normally distributed. There is nothing against such an assumption with regard to the γ's, but the α's must not make (3) singular, and must accordingly be suitably bounded, e.g. by (11).

6. Conclusion. The formulas and bounds of this paper involve only these quantities: the determinant A, its first order minors, and the solutions of (1). They can be found in the course of solving (1) by orthodox methods.

Inequality (10) definitely limits the maximum solution-errors, in terms of the maximum coefficient-error δ, provided δ satisfies (11). But it may be that (8), either alone or in conjunction with the second-order bound from (9), will give a better approximation.

The ratio $\Sigma\Sigma\,|\,A_{\mu\nu}\,|/|\,A\,|$ may be taken as a "measure of sensitivity" of (1) to error.

The fundamental formulas (6) and (7) are capable of solving other problems than those studied here. For example, it may happen that only certain elements (such as those of a single column) are in error, in which case better inequalities can be found. Or the α's and γ's may not be independently and identically distributed.

REFERENCES

[1] I. M. H. ETHERINGTON, "On errors in determinants," *Proc. Edinburgh Math. Soc.*, Ser. 2, Vol. 3 (1932), pp. 107–117.

[2] F. R. MOULTON, "On the solutions of linear equations having small determinants," *Amer. Math. Monthly*, Vol. 20 (1913), pp. 242–249.

[3] L. B. TUCKERMAN, "On the mathematically significant figures in the solution of simultaneous linear equations," *Annals of Math. Stat.*, Vol. 12 (1941), pp. 307–316.

[4] P. G. HOEL, "The errors involved in evaluating correlation determinants," *Annals of Math. Stat.*, Vol. 11 (1940), pp. 58–65.