THE NORMAL EQUATIONS OF THE METHOD OF LEAST SQUARES AND THEIR SOLUTION*

By M. HERZBERGER (*Eastman Kodak Company*)

The literature of the numerical solution of linear equations is extensive.** The fact that war work forced many mathematicians into contact with numerical work has brought with it a better insight into the mathematical problems involved. In 1947, von Neumann and Goldstine[1] discussed an analysis of the accuracy of the best-known methods of solving linear equations and an analysis of the steps involved. The present paper provides the applied mathematician with a method by which he can improve the accuracy of his solution, step by step. At the same time, an attempt is made to guide the computer by transforming the problem into one of analytic geometry.

The problem of solving n linear equations with n unknowns is equivalent to finding the components of a vector $\mathbf{y}$ with respect to n independent vectors $\mathbf{a}_i$:

$$\xi_i \mathbf{a}_i = \mathbf{y}. \tag{1}$$

Multiplication by $\mathbf{a}_j$ gives the equations

$$\xi_i \mathbf{a}_i \cdot \mathbf{a}_j = \mathbf{y} \cdot \mathbf{a}_j , \tag{2}$$

having a symmetric determinant.

The normal equations of the method of least squares can be written in the same form. The geometrical problem is as follows:

Given in $n > k$-dimensional space the k vectors $\mathbf{a}_1 , \cdots , \mathbf{a}_k$ and a vector $\mathbf{y}$, we want to determine a vector $\mathbf{d} = \xi_i \mathbf{a}_i$ such that

$$(\mathbf{y} - \mathbf{d})^2 = \text{min.}, \tag{3}$$

i.e., we want the best representation of $\mathbf{y}$ as a linear manifold of $\mathbf{a}_i$.

The solution is obviously the vector $\mathbf{d}$, which is the projection of $\mathbf{y}$ into the manifold $\mathbf{a}_i$ so that

$$(\mathbf{y} - \mathbf{d}) \cdot \mathbf{a}_i = 0, \tag{4}$$

a result which also can be obtained by differentiating Eq. (3) with respect to $\mathbf{a}_i$. Inserting into Eq. (4) the value of $\mathbf{d}$, we obtain

$$\xi_i \mathbf{a}_i \cdot \mathbf{a}_j = \mathbf{y} \cdot \mathbf{a}_j , \tag{5}$$

the normal equations of the problem of least squares. These equations are identical with Eqs. (2), the only difference being that the $\mathbf{a}_i$ are vectors in a space of higher dimensions.

Equations (5) are the normal equations of the problem of least squares. Their solution, in general, presents a difficulty if, and only if, the determinant of $\mathbf{a}_i$ is small, that means, if the vectors $\mathbf{a}_i$ are "nearly linearly dependent."

In an earlier article[5] the replacement of the vectors $\mathbf{a}_i$ by unit vectors was suggested:

$$\mathbf{e}_i = \mathbf{a}_i / (\mathbf{a}_i^2)^{1/2}; \qquad \mathbf{f} = \mathbf{y} / (\mathbf{y}^2)^{1/2}. \tag{6}$$

*Received July 26, 1948. Communication No. 1207 from the Kodak Research Laboratories.

**Comprehensive summaries have been given by Hotelling[2], Dwyer,[3] and Bodewig.[4]

In this case,

$$\mathbf{e}_i^2 = \mathbf{f}^2 = 1$$
$$\mathbf{e}_i \cdot \mathbf{e}_j = \cos(\mathbf{e}_i, \mathbf{e}_j), \tag{7}$$

i.e., the matrix of Eqs. (5) has all its diagonal elements equal to unity and the other elements smaller than unity.

We now re-order $\mathbf{e}_1, \cdots, \mathbf{e}_k$ with the aim of choosing e_1 and e_2 such that $[e_1e_2]^2$ has the largest possible value, choosing e_3 such that $[e_1e_2e_3]^2$ has the largest possible value, and so on.*

For practical reasons, this criterion may be replaced by the following, which, although not rigorous, is easier to apply. We choose e_1, e_2 such that $(e_1e_2)^2$ has the smallest value, then add e_3 such that $(e_2e_3)^2 + (e_1e_3)^2$ has the smallest value, and so on. These values form, on account of Hadamard's famous theorem, a majorant of the values $[e_1e_2 \cdots e_i]^2$. The re-ordering is done to investigate whether a projection into a space of a smaller number of variables gives a sufficiently close approximation. We must keep in mind that in practical problems the data are only known to a certain number of digits. Therefore, it is only necessary to determine a solution so as to reproduce the accuracy of the data.

To solve the normal equations, two procedures seem possible: one is to change the right-hand side, i.e., to replace the vectors $\mathbf{y}$ successively by smaller vectors; the other is to replace the vectors $\mathbf{a}_i$ on the left-hand side by a series of other vectors which permit the solution of the normal equations. We shall make suggestions with respect to both methods. The determinant of the normal equations is the discriminant of the quadratic form,

$$\left(\sum_1^k \mathbf{a}_i\xi_i\right)^2 = \sum_1^k \mathbf{a}_i\mathbf{a}_j\xi_i\xi_j\,, \tag{8}$$

and as such is positive and definite. Since it is derived from a finite number of digits in machine computation, however, the number of significant figures in the calculated value of the determinant may be small. In either case, the homogeneous equations (right-hand side equal to zero) have only the solution: $\xi_1 = \cdots = \xi_k = 0$.

Let us assume that we have given a vector $\mathbf{b}$ which might be an approximation to $\mathbf{y}$. Such a vector $\mathbf{b}$ permits us to obtain a new and smaller value for the right-hand side of Eqs. (5). If we introduce

$$\mathbf{b} = \beta_i\mathbf{a}_i \tag{9}$$

into Eqs. (5), we obtain

$$\xi_i\mathbf{a}_i\mathbf{a}_j = (\mathbf{y} - \mathbf{b})\mathbf{a}_j = \mathbf{y}\mathbf{a}_j - \beta_i\mathbf{a}_i\mathbf{a}_j\,. \tag{10}$$

The sum of the squares for the new vector is given by

$$(\mathbf{y} - \beta_i\mathbf{a}_i)^2 = \mathbf{y}^2 - 2\beta_i\mathbf{y}\mathbf{a}_i + \beta_i\beta_j\mathbf{a}_i\mathbf{a}_j\,. \tag{11}$$

*The symbol $[e_1e_2 \ldots e_\rho]^2$ is an abbreviation for the determinant of the products e_ie_j for $i, j = 1, \ldots, \rho$, taken from Grassmann's symbolic notation. It is the square of the volume of the parallelepiped formed by the vectors $e_1, \ldots, e_\rho$.

The reader may notice that the computation of the coefficients in (10) and (11) presupposes only a knowledge of the β, the $\mathbf{a}_i\mathbf{a}_j$, and the $\mathbf{ba}_j$, i.e., the data of the $k+1$ by k matrix. It is not necessary to go back to the original vectors in n-dimensional space.

If a number of approximation vectors $\mathbf{b}_1, \mathbf{b}_2, \cdots, \mathbf{b}_m$ are given, then it is possible to compute the best multiple of them to be taken from the right-hand side of the equation. The best approximation is given by

$$\mathbf{B} = \eta_i \mathbf{b}_i \tag{12}$$

with

$$\eta_j \mathbf{b}_j \mathbf{b}_i = \mathbf{b}_i \mathbf{y}.$$

The reader may notice again that the expressions in Eq. (12) can be computed from the coefficients of $\mathbf{b}_i$ with respect to $\mathbf{a}_i$ without having recourse to n-dimensional space.

Of much greater importance for the solution of the normal equation is the transformation of the left-hand side.

There are two different transformations of the left-hand side of the equation which have played a role in the literature of the problem. We shall analyze these transformations in the remainder of the paper. The first method will be called "the improved Gauss-Doolittle"[9] method. The second one is the square-root method found independently by Schur,[6] Banachiewicz,[7] and Dwyer.[3] We shall study both methods theoretically for the general case in which the vectors are not reduced, but we recommend the second method particularly, for the reduction to unit vectors.

Strangely enough, the best way to discuss these methods seems not to be found in the literature.* Both methods can be described as using orthogonal vectors. Let us discuss first the Gauss-Doolittle method. We find a system of orthogonal vectors $\mathbf{o}_1, \cdots, \mathbf{o}_k$ such that $\mathbf{o}_1 = \mathbf{a}_1$, that $\mathbf{o}_2$ lies in the plane of $\mathbf{a}_1$ and $\mathbf{a}_2$, that $\mathbf{o}_3$ is a linear combination of $\mathbf{a}_1, \mathbf{a}_2, \mathbf{a}_3$ and such that in all cases

$$[\mathbf{a}_1\mathbf{a}_2\mathbf{a}_j]^2 = [\mathbf{o}_1\mathbf{o}_2\mathbf{o}_j]^2. \tag{13}$$

This leads to

$$\begin{aligned} \mathbf{a}_1 &= \mathbf{o}_1, \\ \mathbf{a}_2 &= \delta_{12}\mathbf{o}_1 + \mathbf{o}_2, \\ \mathbf{a}_3 &= \delta_{13}\mathbf{o}_1 + \delta_{23}\mathbf{o}_2 + \mathbf{o}_3, \\ &\cdots \end{aligned} \tag{14}$$

The matrix Δ of the δ_{ij} is a triangular matrix which we can calculate as follows. From Eqs. (14) we find that

$$\begin{aligned} \mathbf{a}_1^2 &= \mathbf{o}_1^2, & \mathbf{a}_1\mathbf{a}_j &= \delta_{1j}\mathbf{o}_1^2, \\ \mathbf{a}_2^2 &= \delta_{12}\mathbf{o}_1^2 + \mathbf{o}_2^2, & \mathbf{a}_2\mathbf{a}_j &= \delta_{12}\delta_{1j}\mathbf{o}_1^2 + \delta_{2j}\delta_2^2, \\ \mathbf{a}_3^2 &= \delta_{13}\mathbf{o}_1^2 + \delta_{23}\mathbf{o}_2^2 + \mathbf{o}_3^2, & \mathbf{a}_3\mathbf{a}_j &= \delta_{13}\delta_{1j}\mathbf{o}_1^2 + \delta_{23}\delta_{2j}\mathbf{o}_2^2 + \delta_{3j}\mathbf{o}_3^2. \end{aligned} \tag{15}$$

*As far as we know, Kolmogoroff[10] is the only other author who has used vector notation for the least-squares method. His paper, published shortly before our first communication,[5] reached us only recently.

If we abbreviate the *diagonal matrix*,

$$\begin{pmatrix} \mathbf{o}_1^2 & 0 & 0 & 0 & 0 \\ 0 & \mathbf{o}_2^2 & 0 & 0 & 0 \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ 0 & 0 & 0 & 0 & \mathbf{o}_k^2 \end{pmatrix} = O \tag{16}$$

these equations can be written in the matrix form

$$A = \Delta' O \Delta. \tag{17}$$

Since $\mathbf{a}_i$ and $\mathbf{a}_j$ are given, the computation of δ_{ij} can now be done in the following way. Equations (15) give the relations

$$\begin{aligned} \mathbf{o}_1^2 &= \mathbf{a}_1^2 , & \delta_{1j} &= (\mathbf{a}_1\mathbf{a}_j)/\mathbf{o}_1^2 , \\ \mathbf{o}_2^2 &= \mathbf{a}_2^2 - \delta_{12}^2\mathbf{o}_1^2 , & \delta_{2j} &= (\mathbf{a}_2\mathbf{a}_j - \delta_{12}\delta_{1j}\mathbf{o}_1^2)/\mathbf{o}_2^2 , \\ \mathbf{o}_3^2 &= \mathbf{a}_3^2 - \delta_{13}^2\mathbf{o}_1^2 - \delta_{23}^2\mathbf{o}_2^2 , & \delta_{3j} &= (\mathbf{a}_3\mathbf{a}_j - \delta_{13}\delta_{1j}\mathbf{o}_1^2 - \delta_{23}\delta_{2j}\mathbf{o}_2^2)/\mathbf{o}_3^2 . \end{aligned} \tag{18}$$

Having obtained $\mathbf{o}_i^2$ and the triangular matrix, δ_{ik}, we can solve the least-squares equations in two simple steps. The equations $A\xi_j = \eta_i$ are equivalent to $\Delta' O \Delta \xi_j = \eta_i$. We write

$$\Delta\xi_j = \zeta_\rho \tag{19}$$

and

$$O\Delta'\zeta_\rho = \eta_i . \tag{20}$$

These equations are easily solved because of the triangularity of the matrix Δ. Written in full, Eqs. (20) are

$$\begin{aligned} \mathbf{o}_1^2\zeta_1 &= \eta_1 , \\ \mathbf{o}_1^2\delta_{12}\zeta_1 + \mathbf{o}_2^2\zeta_2 &= \eta_2 , \\ \mathbf{o}_1^2\delta_{13}\zeta_1 + \mathbf{o}_2^2\delta_{23}\zeta_2 + \mathbf{o}_3^2\zeta_3 &= \eta_3 , \end{aligned} \tag{21}$$

or, solved

$$\begin{aligned} \zeta_1 &= \eta_1/\mathbf{o}_1^2 , \\ \zeta_2 &= (\eta_2 - \mathbf{o}_1^2\delta_{12}\zeta_1)/\mathbf{o}_2^2 , \\ \zeta_3 &= (\eta_3 - \mathbf{o}_1^2\delta_{13}\zeta_1 - \mathbf{o}_2^2\delta_{23}\zeta_2)/\mathbf{o}_3^2 . \end{aligned} \tag{22}$$

The computation of ξ_i is then obtained by the equations

$$\begin{aligned} \xi_k &= \zeta_k , \\ \xi_{k-1} &= \zeta_{k-1} - \delta_{k,k+1}\xi_k . \end{aligned} \tag{23}$$

It may be noted that the ζ_i are the coefficients of the vector $\mathbf{x} = \xi_i \mathbf{a}_i$ with respect to $\mathbf{o}_i$, for

$$\mathbf{x} = \xi_i \mathbf{a}_i = \xi_i \delta_{ki} \mathbf{o}_j = \zeta_k \mathbf{o}_j . \tag{24}$$

The discussion of the *square-root method* is similar, except that, instead of choosing orthogonal vectors which fulfill Eq. (13), we choose orthogonal unit vectors $\mathbf{o}_\nu$ in the direction of $\mathbf{o}_i$. This leads again to a triangular matrix of coefficients:

$$\begin{aligned} \mathbf{a}_1 &= \gamma_{11}\mathbf{o}_1 , \\ \mathbf{a}_2 &= \gamma_{12}\mathbf{o}_1 + \gamma_{22}\mathbf{o}_2 , \\ \mathbf{a}_3 &= \gamma_{13}\mathbf{o}_1 + \gamma_{23}\mathbf{o}_2 + \gamma_{33}\mathbf{o}_3 , \end{aligned} \tag{25}$$

with the distinction that

$$A = \Gamma'\Gamma, \tag{26}$$

since

$$\mathbf{a}_i \mathbf{a}_j = \gamma_{\nu i}\gamma_{\nu j} . \tag{27}$$

Hence, by this procedure the original matrix becomes the product of two triangular matrices.

The computation of γ_{ij} proceeds similarly to that of δ_{ij}. Equations (25) give

$$\begin{aligned} \gamma_{11}^2 &= \mathbf{a}_1^2 , & \gamma_{11}\gamma_{1j} &= \mathbf{a}_1\mathbf{a}_j , \\ \gamma_{12}^2 + \gamma_{22}^2 &= \mathbf{a}_2^2 , & \gamma_{12}\gamma_{1j} + \gamma_{22}\gamma_{2j} &= \mathbf{a}_2\mathbf{a}_j . \end{aligned} \tag{28}$$

Thus, we find that

$$\begin{aligned} \gamma_{11} &= (\mathbf{a}_1^2)^{1/2}, & \gamma_{1j}\gamma_{11} &= \mathbf{a}_1\mathbf{a}_j , \\ \gamma_{22} &= (\mathbf{a}_2^2 - \gamma_{12}^2)^{1/2}, & \gamma_{2j}\gamma_{22} &= \mathbf{a}_2\mathbf{a}_j - \gamma_{1j}\gamma_{12} , \\ \gamma_{33} &= (\mathbf{a}_3^2 - \gamma_{13}^2 - \gamma_{23}^2)^{1/2}, & \gamma_{3j}\gamma_{33} &= \mathbf{a}_3\mathbf{a}_j - \gamma_{2j}\gamma_{22} - \gamma_{1j}\gamma_{11} . \end{aligned} \tag{29}$$

The computation of square roots on a calculating machine can be reduced to simple division if an approximate value is known, because, for any value x and a small increase dx, we have

$$x^2/(x + dx) = x - dx. \tag{30}$$

Having found the value of γ_{ij}, we can proceed with the solution of Eqs. (2), again in two steps:

$$\Gamma'\Gamma\xi_j = A\xi_j = \eta_i \tag{31}$$

which splits into

$$\begin{aligned} \Gamma\xi_j &= \tilde{\zeta}_\rho \\ \text{and} \qquad & \\ \Gamma'\tilde{\zeta}_\rho &= \eta_i . \end{aligned} \tag{32}$$

Again we find that because of Eqs. (25) and (32),

$$\mathbf{x} = \xi_i \mathbf{a}_i = \xi_i \mathbf{o}_i , \tag{33}$$

so that the ζ are components of $\mathbf{x}$ with respect to $\mathbf{o}_i$.

The advantage of the second method in connection with making the vectors $\mathbf{a}$ and $\mathbf{y}$ unit vectors is that all the numbers calculated, with the exception of the final ξ_i , are projections of vectors of unit length into a Cartesian coordinate system; they, therefore, are all smaller than one. The accuracy of all the operations can hence be preserved to the last significant figure, since division does not lead to a loss of significant figures.

A check consists in forming

$$\Gamma'\Gamma = A \tag{34}$$

and making sure that the elements of A are reproduced to the last decimal.

These remarks hold up to and including the computation of the ζ. If the values of ξ_i , when inserted into Eqs. (5), leave small residuals, a correction can be obtained by repeating the solution of Eqs. (5) with the residuals at the right-hand side. A small number of digits will be sufficient to carry through the computation in this case.

The author suggests the following procedure in solving a system of linear equations having a small determinant and a large number of variables. The first step is to change the vectors $\mathbf{a}_i$ and $\mathbf{b}_i$ into unit vectors and to rearrange the vectors in the manner described previously. We then solve part of the equations according to the square-root method, until $\gamma_{\rho\rho}$ has sufficient significant digits, using the number of decimals which the calculating machine permits. This gives a solution for $\rho < k$ unknowns and reduces the first ρ numbers on the right-hand side to zero or nearly zero. These values are put into the remaining $k - \rho$ equations, part of which are solved, and the same procedure is followed until values of all the unknowns are determined. If the sum of the squares of the residuals is small enough, we stop. If not, we repeat the procedure.

Going through it for a second time, we look for the best multiple of the last solution to diminish further the sum of the squares. The procedure is repeated until the sum of the squares becomes compatible with the error of the original data.

Acknowledgment. I should like to express here my gratitude to Mr. R. Morris, who worked with me on many theoretical details of the paper, and to Dr. F. Kottler, of these Laboratories, for his helpful criticisms.

References

1. J. von Neumann and H. Goldstine, *Numerical inverting of matrices of high order*, Bull. Amer. Math. Soc. **53**, 1021-1099 (1947).

2. H. Hotelling, *Some new methods in matrix calculation*, Ann. Math. Stat. **14**, 1-34 (1943).

3. P. S. Dwyer, *Recent development in correlation technique*, Amer. Stat. Assn. **37**, 441-460 (1941).

4. E. Bodewig, *Comparison of some direct methods for computing determinants and inverse matrices*, Nederl. Akad. Wetens. **50**, 49-57 (1947).

5. M. Herzberger and R. Morris, *A contribution to the method of least squares*, Q. Appl. Math. **5**, 355-357 (1947).

6. J. Schur, *Uber Potenzreihen die im Innern des Einheitskreises beschrankt sind*, Crelle's Journal f. reine u. ange. Math. **147**, 205-232 (1917).

7. T. Banachiewicz, *Calcul des déterminants par la méthode des cracoviens*, Acad. Polon. Sci. (Bull. Int.), A. 1937, pp. 109-120.

8. C. F. Gauss, *Supplementum theoriae combinationis observationum erroribus minimis obnoxiae*, C. F. Gauss, Werke, Göttingen, 1870, vol. 4, pp. 55-93.

9. M. H. Doolittle, *Method employed in the solution of normal equations and the adjustment of a triangulation*, U. S. Coast Survey Rep., 1878, pp. 115-120.

10. A. N. Kolmogoroff, *On the proof of the method of least squares* (Russian), Uspekhi mat. Nauk (new series) 1, 1946, pp. 57-70.

ON REISSNER'S THEORY OF BENDING OF ELASTIC PLATES*

By A. E. GREEN (*Durham, England*)

1. Introduction. The classical theory of bending of elastic plates has recently been extended and improved by Reissner.[1,2,3] His theory takes into account the transverse-shear deformations of the plate and the equations of the theory are obtained by an application of Castigliano's theorem of minimum energy. The object of the present note is to show that Reissner's equations can be obtained directly from the stress equations of equilibrium and the stress-strain relations. Moreover, by consistent use of complex variable notation, the form of the results is simplified. The equations are first obtained for an isotropic material and are then extended to an aeolotropic material which is transversely isotropic in planes parallel to the faces of the plates.

2. Fundamental equations for isotropic plates. Consider Cartesian coordinates x, y, Z and let $z = x + iy$ denote the complex variable with $\bar{z} = x - iy$ the complex conjugate of z. Stresses connected with the coordinate Z are denoted by τ_{zz} , τ_{yz} , σ_z since there is no need to confuse the z in this notation with the complex variable. Attention is directed to stresses in plates bounded by the planes $Z = \pm h$.

When body forces are absent, Stevenson[4] has shown that the stress equations of equilibrium can be expressed in the form

$$\frac{\partial \Phi}{\partial z} + \frac{\partial \Theta}{\partial \bar{z}} + \frac{\partial \Psi}{\partial Z} = 0, \tag{1a}$$

$$\frac{\partial \Psi}{\partial z} + \frac{\partial \bar{\Psi}}{\partial \bar{z}} + \frac{\partial \sigma_z}{\partial Z} = 0, \tag{1b}$$

where a bar placed over a quantity denotes the complex conjugate of that quantity and where

$$\Theta = \sigma_x + \sigma_y , \qquad \Phi = \sigma_x - \sigma_y + 2i\tau_{xy} , \qquad \Psi = \tau_{xz} + i\tau_{yz} . \tag{2}$$

If u, v, w denote the Cartesian components of displacement and if $D = u + iv$, the complex form of the stress-strain relations is

$$(1 - 2\eta)\Theta = 2\mu\left\{\nabla_1^2(F + \bar{F}) + 2\eta \frac{\partial w}{\partial Z}\right\}, \tag{3a}$$

*Received Aug. 13, 1948.

[1] E. Reissner, J. Math. Phys. 23, 184-191 (1944).

[2] E. Reissner, J. Appl. Mech. 12, A68-A77 (1945).

[3] E. Reissner, Q. Appl. Math. 5, 55-68 (1947).

[4] A. C. Stevenson, Phil. Mag. (7) 33, 639-661 (1942).