—NOTES—

A CONTRIBUTION TO THE METHOD OF LEAST SQUARES*

BY M. HERZBERGER AND R. H. MORRIS (*Eastman Kodak Company*)

In trying to calculate the image figures of an optical problem the authors were led to consider a problem of least squares which proved unusually difficult, since the determinant of the normal equations to be solved had a very small value. The following attempt to make the problem a geometrical one should be of value in other applications, and is presented here for what it is worth.

Most problems of least squares can be reduced to the following: Suppose that we are given a sequence of k known functions $f_1, \cdots, f_k$ of one or more variables. Our problem is to determine k constants, C_i, so that an unknown function, F, whose values are known for $n > k$ values of the coordinates will be approximated as closely as possible by a linear combination of the functions f_i:

$$F = C_1 f_1 + \cdots + C_k f_k. \tag{1}$$

Since F is given for only $n > k$ values, we can consider these values as the coordinates of a vector $\bar{b}$ in an n-dimensional space. The values of the f_i for the n values of the coordinates may also be considered as the coordinates of k vectors $\bar{a}_i$ in the n-dimensional space. It is a simple geometrical problem to find the best approximation of the vector $\bar{b}$ by a linear combination of the vectors $\bar{a}_i$. We merely project $\bar{b}$ in to the space of the vectors $\bar{a}_i$, i.e., we minimize

$$\left(\bar{b} - \sum_{i=1}^{k} C_i \bar{a}_i\right)^2. \tag{2}$$

This leads to the k normal equations for the constant C_i

$$\sum_i C_i a_i d_\lambda = b a_\lambda, \tag{3}$$

which must be solved.

We can simplify the problem considerably if we order the vectors in a suitable way, which in this case will be the transformation of (3) into a form which contains only unit vectors, i.e., the projection of our problem on the unit sphere. Let

$$\bar{a}_i = \alpha_i \bar{s}_i, \qquad \bar{b} = \beta \bar{t}, \tag{4}$$

where $s_i^2 = t^2 = 1$. Equation (3) becomes

$$\sum \gamma_i s_i s_\lambda = t s_\lambda. \tag{5}$$

A solution of (5) which yields γ_i will also yield C_i for

$$C_i = \frac{\beta \gamma_i}{\alpha_i}, \qquad \gamma_i = \frac{\alpha_i C_i}{\beta}. \tag{6}$$

From (4) and (5) we see that the diagonal terms of the matrix of (5) are all 1, whereas the members not on the diagonal are smaller than 1.

* Received Nov. 1, 1946. Communication No. 1123 from the Kodak Research Laboratories.

The senior author now suggests an ordering of the vectors in such a way that the principal subdeterminants* have a maximum value (which will always be less than 1, since k independent edges of the parallelopiped formed by the unit vectors $\bar{s}_i$ lie inside the unit sphere). If the determinant of the problem is small, this means that the volume of the parallelopiped is small or that the vectors $\bar{a}_i$ are not "too independent," i.e., that one or more of the vectors can be approximated by a linear combination of the others.

If we denote by a_{ij} the quantity, $s_i s_j$, by $c_i^{(0)}$ the quantity ts_i, we may write (5) in the form

$$\begin{aligned} a_{11}\gamma_1 + a_{12}\gamma_2 + \cdots + a_{1k}\gamma_k &= c_1^{(0)}, \\ \cdots\cdots\cdots\cdots\cdots\cdots & \\ a_{k1}\gamma_1 + a_{k2}\gamma_2 + \cdots + a_{kk}\gamma_k &= c_k^{(0)}. \end{aligned} \tag{7}$$

The method used will be this: We choose an arbitrary number, say r, where $r<k$, and find the r-rowed principal subdeterminant which has the largest value. We set the $(k-r)$ variables not involved equal to zero, and solve the r equations for the r unknown variables, obtaining the values $\gamma_{j,1}=\bar{\gamma}_{1,1}\cdots\bar{\gamma}_{r,1},\gamma_{j,1}=0$, $j>r$. Substituting these for the variables $\gamma_{1,1}\cdots\gamma_{r,1}$ in the $k-r$ remaining equations, we have

$$\begin{aligned} a_{1,r+1}\gamma_{r+1} + \cdots + a_{1,k}\gamma_k &= c_1^{(1)}, \\ a_{k,r+1}\gamma_{r+1} + \cdots + a_{k,k}\gamma_k &= c_k^{(1)} \end{aligned} \tag{8}$$

where $c_i^{(1)}=c_i^{(0)}-\sum_{j=1}^{r}a_{i,j}\bar{\gamma}_{j,i}$. At this point we have two possibilities. Either the constants $c_i^{(1)}$ are very near to zero, in which case we have that the solution of (8) is given by $\gamma_j=0$, $j>r$, so that a reasonably accurate solution of (7) and thus of (5) is given by

$$\gamma_{i,1} = \bar{\gamma}_{i,1} \quad (i = 1, 2, \cdots, r); \qquad \gamma_{i,1} = 0 \quad (i = r+1, \cdots, k). \tag{9}$$

In the event that they are not sufficiently small, we take the largest $(r+1)$-rowed principal subdeterminant of (7), equate the remaining $(k-r-1)$ variables not involved to zero, and proceed as before, obtaining $(k-r-1)$ equations in $(k-r-1)$ unknowns:

$$\sum_{\lambda=1}^{k-r-1} \mu_{i\lambda}\gamma_{j\lambda} = c_i^{(2)}, \qquad (i = 1, 2, \cdots, k-r-1), \tag{10}$$

where the meaning of the terms $\mu_{i\lambda}$ and $c_i^{(2)}$ is clear. Should the constants $c_i^{(2)}$ be sufficiently small, the approximate solution is given by

$$\gamma_i = \bar{\gamma}_i \quad (i = 1, 2, \cdots, r+1); \qquad \gamma_i = 0, \quad (i = r+2, \cdots, k). \tag{11}$$

One must keep in mind that as larger subdeterminants are taken, their values get smaller, and that, of course, as r gets nearer k the $c_i^{(n)}$ become smaller, and vanish when $r=k$. There will usually be a place somewhere in this process, however, where sufficiently good accuracy will be achieved. If greater accuracy is desired, one must

* Subdeterminants whose principal diagonals form part of the principal diagonal of the determinant of the system (5).

attempt one of the ordinary approximation methods, in particular, an iteration procedure.

We present as an example a set of six equations with six unknowns. Since the calculation of the maximum subdeterminant is very laborious, we have replaced it by the less laborious, but theoretically less satisfying, method of ordering the vector so that the sum of the squares of the elements outside the principal subdeterminant is as small as possible.

The system is represented by the following matrix.

I

6.35083641	3.38272218	1.83936868	1.02023449	0.57681127	0.33173876	2.27198287
	1.83271729	1.01325021	0.57097916	0.32757965	0.19093008	1.26347780
		0.56915637	0.32551164	0.18928551	0.11166820	0.71717786
			0.18871015	0.11107961	0.06624255	0.41473305
				0.06609478	0.03979216	0.24391707
					0.02415650	0.14552476

This matrix is reduced to unit diagonal and reordered as described above (in this example, the ordering remains the same). The elements of the resulting matrix are the light-face elements of the following matrix.

II

1.00000000	0.99152377	0.96747004	0.93193774	0.89029657	0.84696167	0.90154952	**0.90154952**
	1.00000000	0.99209478	0.97090085	0.94120796	0.90742269	0.93329702	
−0.99152377	**0.01688061**	**0.03282524**	**0.04686243**	**0.05845775**	**0.06764006**	**0.03938924**	**2.33340146**
		1.00000000	0.99323746	0.97592846	0.95235029	0.95062903	
−0.96747004	**−1.94455295**	**0.00017130**	**0.00048914**	**0.00091901**	**0.00141057**	**0.00181242**	**10.58038529**
			1.00000000	0.99461070	0.98112063	0.95470899	
−0.93193774	**−2.77610999**	**−2.85545826**	**0.00000007**				
				1.00000000	0.99585694	0.94876491	
−0.89029657	**−3.46301170**	**−5.36491535**					
					1.00000000	0.93631051	
−0.84696167	**−4.00696776**	**−8.23450088**					

The bold-face elements are found as follows: The first bold-face row contains the determinants of the second order found from the first and second rows. Any other bold-face elements above the stairway dividing line are formed by adding to the element immediately above it the product of the rest of the column above it and the row to its left, e.g., $0.99323748-0.93193775\times0.96747003-0.04686244\times1.94455183 = 0.00048919$. The first column to the left of the stairway dividing is the negative of first row, the second column is the negative of the first bold-face row divided by its first number, and so on. The process is continued until one of the diagonal elements becomes very small. The solution of the equations, omitting the last and successive variables, is found in the usual manner from the bold-face column at the extreme right. This column is the ratio of the last number of each row to the first number of each row.

We find in our case·

$$\gamma_1 = 8.75144919, \qquad \gamma_2 = -18.24071797, \qquad \gamma_3 = 10.58038529.$$

The three remaining unknowns are found by distributing the residuals among all six equations by least squares. When this is done, we find for our original unknowns:

$$C_1 = 3.4727, \qquad C_2 = -13.4739, \qquad C_3 = 14.0244,$$
$$C_4 = -0.0003, \qquad C_5 = 0.0002, \qquad C_6 = 0.0001.$$

This method has been applied successfully to least-squares solutions in geometrical optics and to the colorimetric problem of finding polynomial reflection curves which will yield a prescribed set of tristimulus values under a given illuminant and fit other prescribed conditions.

In the latter problem, the small unknowns must be neglected prior to symmetrization and reduction to unit diagonal.

NEW FORMULATIONS OF THE EQUATIONS FOR COMPRESSIBLE FLOW*

By B. L. HICKS (*Ballistic Research Laboratories, Aberdeen Proving Ground*)
P. E. GUENTHER (*Case School of Applied Science*) and R. H. WASSERMAN (*University of Chicago*)

Introduction. A prominent aerodynamic effect of combustion in a moving gas stream is an alteration of the flow pattern owing to heat release within the fluid. This alteration occurs not only in the immediate neighborhood of heat sources but also downstream where the entropy and stagnation temperature vary from one streamline to another. As a background for combustion research, appropriate descriptions of these altered flow patterns have been investigated. This paper considers the downstream patterns, which are restricted to be the adiabatic and steady flows of an inviscid fluid. In a second paper,[1] diabatic (i.e., non-adiabatic) flows will be discussed.

Since one-dimensional flow theory[2,3] can be considerably condensed by use of the local Mach number M, it was natural to seek a corresponding condensation with the help of the Mach vector

$$\mathbf{M} = M\mathbf{V}/V$$

and the Crocco vector

$$\mathbf{W} = \mathbf{V}/V_t$$

in which V_t is the variable limiting velocity at each point of the fluid. The introduction of Mach and Crocco vectors into the compressible flow equations sufficiently simplified or altered their form that a number of further investigations were suggested including those of diabatic flow.

* Received Feb. 12, 1947. This paper is a revised report of theoretical work performed by the authors at the Cleveland Laboratory of the National Advisory Committee for Aeronautics in 1943–45.

[1] B. L. Hicks, *Diabatic flow of a compressible fluid*, submitted to Quarterly of Applied Mathematics.

[2] Neil P. Bailey, *The thermodynamics of air at high velocities*, Journ. Aero. Sci. **11**, 227–238 (1944).

[3] B. L. Hicks, D. J. Montgomery, and R. H. Wasserman, *The one dimensional theory of steady compressible fluid flow in ducts with friction and heat addition*, NACA TN, 1947.