On the correlation of physical properties with chemical composition in multivariate systems

By Max H. Hey, M.A., D.Sc.

Department of Mineralogy, British Museum (Natural History).

[Read 27 January 1955.]

Summary.—A systematic and time-saving procedure for the correlation of optical or other physical properties with chemical composition is outlined, and is applicable even where the composition is complex and involves several variables. The procedure is applied to anthophyllite, for which the following partial regression equations are derived:

$$\gamma = 1{\cdot}7249 - 0{\cdot}0130\mathrm{Si} + 0{\cdot}0140(\mathrm{Ti} + \mathrm{Fe}''' + \mathrm{Fe}'' + \mathrm{Mn}) + 0{\cdot}0012,$$
$$\beta = 1{\cdot}7275 - 0{\cdot}0142\mathrm{Si} + 0{\cdot}024(\mathrm{Ti} + \mathrm{Fe}''') + 0{\cdot}0110(\mathrm{Fe}'' + \mathrm{Mn}) \pm 0{\cdot}0015,$$
$$\alpha = 1{\cdot}6951 - 0{\cdot}0117\mathrm{Si} - 0{\cdot}040(\mathrm{Ti} + \mathrm{Fe}''') + 0{\cdot}0133(\mathrm{Fe}'' + \mathrm{Mn}) \pm 0{\cdot}0025,$$
$$b\,(\text{Å}.) = 16{\cdot}44 + 0{\cdot}28\mathrm{Si} - 0{\cdot}13\ \mathrm{Mg} + 0{\cdot}40(\mathrm{Ca} + \mathrm{Na} + \mathrm{K}) + 0{\cdot}04.$$

The a and c cell-dimensions appear to be constant, within the experimental error of the available data.

BY far the most popular as well as the simplest means of deriving and displaying the correlation of the physical properties of a mineral with its chemical composition is the graphical method, either in the form of a linear graph or a triangular or square contoured correlation diagram; but these methods are necessarily limited to minerals whose variations in composition can be expressed in terms of one or two parameters,[1] and only the simple linear graph can indicate efficiently the probable accuracy of the correlation.

The method of least squares is applicable to both linear and non-linear correlations with any number of independent parameters, and is widely used in many branches of the physical and biological sciences, but has rarely been applied in mineralogy. Probably one reason for this neglect is the lack of any readily available description of the method as applied to multivariate systems, for such well-known textbooks as J. W. Mellor's 'Higher Mathematics for Students of Chemistry and Physics' (London, 1931) and R. A. Fisher's 'Statistical Methods for Research

[1] The triangular diagram involves three variables, the square one four, but in each case only two are independent.

Workers' (11th edn., Edinburgh, 1950) only devote a few short paragraphs to it, and many only deal with its application to simple correlations with one independent parameter ($y = ax+b$), for which the familiar graphical method is usually adequate. Another reason for this neglect is the assumption that the procedure, as applied to a multivariate system, is necessarily very time-consuming; the evaluation of a determinant of high order is at best a slow process and needs careful checks, and at first glance a set of regression equations in n independent variables would appear to involve the evaluation of $n+1$ determinants of nth order for each dependent variable, plus ancillary calculations: the procedure outlined below calls for little more labour than is involved in the calculation of 2 to 4 determinants of the nth order, whatever the number of variables.

In many minerals the principal physical properties, including unit-cell dimensions, density, refractive indices, and birefringences, appear to be reasonably well reproduced by linear partial regression equations in which the independent parameters define the chemical composition; they may be weight percentages, empirical unit-cell contents, or atomic ratios (the last will usually be most convenient). The possibility of a non-linear regression must always be borne in mind; perhaps the simplest check is, after deriving the best linear correlation, to plot the residuals (that is, the differences between the observed values of the dependent variable and the value calculated from the regression equation) against each of the compositional parameters in turn. The points should lie evenly about a straight line parallel to the composition axis; if they appear to lie evenly about a straight line not parallel to the composition axis, the regression coefficient for that constituent has been incorrectly estimated, while if they suggest a curve, the correlation is probably non-linear in that constituent.[1]

Before proceeding to discuss methods for deriving regression equations, certain inherent limitations to their use should, perhaps, be emphasized. A regression equation can only be utilized for one particular purpose—an equation or set of equations derived to predict the refractive index or indices of a mineral, given the chemical composition, cannot properly be used to predict the composition, given the refractive indices. The reason for this restriction is clearly explained and illustrated by M. J. Moroney.[2] If it is desired to predict the chemical composition, given an

[1] M. H. Hey, Min. Mag., 1954, vol. 30, pp. 281–4.

[2] M. J. Moroney, Facts from Figures, 2nd edn., Harmondsworth, 1953, pp. 293–4.

adequate number of physical data, a suitable set of regression equations must be calculated for that purpose.[1]

The derivation of a partial regression equation; selection of data

The first steps in the investigation of any correlation are to decide on the relevant variables and to review all the available data. As a rule, the independent variables will be parameters defining the chemical composition, and the only question will be how many such parameters are really essential to the correlation: every additional variable increases the order of the determinants involved and adds greatly to the labour of calculation; on the other hand, whereas it is a relatively simple matter to discard a superfluous variable, the whole investigation has to be repeated if it should be found later that some neglected variable is really essential to the correlation (e.g. if some supposedly unimportant constituent such as fluorine proves to have an important effect on a physical constant such as the refractive index).

Having decided on the relevant variables and collected all the available data, we have to decide whether all the data should or can be used, and whether all should be assigned equal weight. Generally speaking, it is desirable to make the fullest use of all available data, but it will be obvious that if part of the data is more accurate than the rest, it should be given greater weight.[2] Occasionally it may happen that a part of the data is much more accurate than the rest; then it may be desirable to derive the regression equation from the more accurate data alone, and check the result by comparison of predictions made using this equation with the remaining, less accurate data. Similarly, a few sets of data of doubtful accuracy are best rejected for the derivation of the regression equation.

If several physical quantities, such as refractive indices, density,

[1] This has been done for the mineral here taken as an example, anthophyllite, but owing to the restricted range of variation of the birefringences the results will always be rather inaccurate. The equations derived for the composition, in atoms per 24(O,OH,F), and assuming an accuracy of $\pm 0{\cdot}001$ in the refractive indices and $\pm 0{\cdot}03$ in the density, are:

$$\begin{aligned}
\mathrm{Si} &= 165\,(\gamma-\beta)-102\,(\gamma-\alpha)-30\gamma-3{\cdot}3D+60{\cdot}13\pm 0{\cdot}5,\\
\mathrm{Al} &= -346\,(\gamma-\beta)+229\,(\gamma-\alpha)+29\gamma+10{\cdot}8D-81{\cdot}70\pm 1{\cdot}0,\\
\mathrm{Mg} &= -33\,(\gamma-\beta)+37\,(\gamma-\alpha)-61\gamma+0{\cdot}2\,D+103{\cdot}87\pm 0{\cdot}5,\\
\mathrm{Fe''} &= 103\,(\gamma-\beta)-10\,(\gamma-\alpha)-50\gamma+18{\cdot}3\,D+24{\cdot}59\pm 0{\cdot}4.
\end{aligned}$$

These equations give a fairly satisfactory representation of the ten superior sets of data by J. C. Rabbitt (see below).

[2] The procedure for giving greater or less weight to a part of the data is discussed below, p. 91.

unit-cell dimensions, &c., are to be correlated with the same group of independent variables, such as parameters defining the composition, it will almost always be advantageous to reject any sets of observations in which only some of the physical quantities were determined (provided there are not too many such sets). The reason for this is that if the independent variables are the same for all the sets of observations utilized, a single covariance matrix can be calculated and applied to derive the regression equations for all the physical quantities; and, as will be seen below, the calculation of the covariance matrix is the most laborious stage in the calculations. But calculated values for any rejected data should always be derived and compared with the observations, as a check.

It is perhaps preferable, in discussing the derivation of a set of partial regression equations, to consider a concrete example, and the correlation of the refractive indices and unit-cell dimensions of anthophyllite with its composition is a useful one, because it is not too simple, but illustrates well many of the possible complications. The available optical data consist of some thirty-eight sets of data obtained on analysed specimens,[1] but in some cases only one refractive index or only the optic axial angle was determined; moreover, ten sets of data, all determined by one worker and therefore more strictly comparable, are of decidedly superior accuracy. The X-ray data consist of ten sets of cell-dimensions, on a group of analysed specimens[2] overlapping the group with optical data of superior accuracy. The variations in the *a* and *c* dimensions are within the accuracy of the measurements, but the *b* dimension shows significant variation.

The first step was to select suitable parameters to define the composition. Anthophyllite may have at least fourteen composition variables: Si, Ti, Al, Fe‴, Fe″, Mn, Mg, Ca, Na, K, F′, OH′, O″, and interstitial H_2O (as in the Glen Urquhart gedrite); but only thirteen of these will be independent, on account of the valency balance.

The composition would be best expressed in empirical unit-cell contents, but the regression equations could then only be used where the necessary density and cell-dimensions were available. If empirical unit-

[1] Analyses 1, 2, 3, 4, 6, 7, 8, 9, 10, 13, 14, 15, 16, 17, 20, 22, 24, 25, 26, 29, 30, 33, 34, 35, 38, 39, 40, 41, 43, 44, 45, 72, 79a, and 85 of J. C. Rabbitt (Amer. Min., 1948, vol. 33, p. 263 [M.A. **10**–416]); R. Pirani, Atti (Rend.) Accad. Naz. Lincei, cl. fis. mat. nat., 1952, ser. 8, vol. 13, sem. 2, p. 83 [M.A. **12**–30], and p. 170 [M.A. **12**–140], and 1953, vol. 15, sem. 2, p. 422 [M.A. **12**–374]; G. H. Francis, Min. Mag., 1955, vol. 30, p. 709.

[2] J. C. Rabbitt's nos. 1, 8, 9, 14, 17, 20, 26, 29, 30, and 43.

cell contents are not used, some assumption must be made, and it is convenient to assume that $\sum$ (O,OH,F) = 96, neglecting the possibility of interstitial water; in fact, none of the ten selected analyses shows any interstitial H_2O, so this parameter is not required, though, as will be seen below, interstitial H_2O appears to have a very marked effect on the refractive indices. The number of parameters has thus been reduced to eleven. Owing to the small number of analyses in which it was sought, F′ was necessarily neglected; this is unfortunate in view of the large effect it usually has on the refractive indices. Variations in OH′ were also neglected in view of the uncertainties relating to the water determinations. Further, in view of the small amounts of Ti and Fe‴ present, and their similar optical effects, (Ti+Fe‴) was taken as one variable for the optical data; similarly, (Ca+Na+K) was taken as one variable, and Mn was included with Fe″. The reduced number of independent parameters is six: Si, Al, (Ti+Fe‴), (Fe″+Mn), Mg, and (Ca+Na+K). Finally, it has been assumed that the replacements Ti $\rightleftharpoons$ Fe, Ca $\rightleftharpoons$ (Na, K), and O $\rightleftharpoons$ (OH,F) can be set against one another and balanced out within the probable accuracy of the chemical analyses; with this restriction, the requirement of valency balance reduces the number of independent parameters to five. It is convenient to discard Al as an independent parameter, leaving for the correlation with the optical data: Si, (Ti+Fe‴), (Fe″+Mn), Mg, and (Ca+Na+K), all expressed in atoms per 24(O,OH,F); the unit of 24(O,OH,F) was chosen to facilitate comparison with the clinoamphiboles and the pyroxenes.

For the correlation with the unit-cell dimensions it may reasonably be assumed that in view of their similar ionic radii the small amounts of Mn may be included with Fe″, and the Ti and Fe‴ with Al, while Ca, Na, and K may be taken together. And in view of the uncertainty of some of the water contents, and the similarity in the radii of F′, OH′, and O″, the variations in H_2O and F′ have been disregarded. The number of variables is thus reduced to five, only four of which are independent; these are the same as for the correlation with the refractive indices, excluding Ti+Fe‴.

The next step is to decide whether to use the whole or a selected part of the available optical data (all the X-ray data are useful and of about equal weight). The use of incomplete sets of data would have the disadvantage that a complete recalculation of the covariance matrix (see below, p. 78) would be needed in respect of that optical constant for which additional data was available, and the incomplete data were not numerous, nor did they extend the composition field notably; they were

TABLE I. X-ray and selected optical data for analysed specimens of anthophyllite. Chemical data are given in atoms per 24(O,OH,F). The inferior optical data in square brackets are not included in the means. M_o gives the means of the ten sets of data used for the optical correlations, nos. 1, 8, 9, 14, 17, 22, 26, 29, 30, and 35, while M_x gives the means of the ten sets used for the X-ray correlation, nos. 1, 8, 9, 14, 17, 20, 26, 29, 30 and 43.

No.*	Si.	Ti+Fe‴.	Fe″+Mn.	Mg.	Ca+Na+K.	γ.	β.	α.	b(Å.)
1	6·20	0·15	2·23	3·37	0·41	1·6781	1·6670	1·6566	17·96
8	6·50	0·35	1·83	3·78	0·08	1·6718	1·6630	1·6553	17·86
9	6·60	0·12	2·08	3·93	0·12	1·6695	1·6603	1·6520	17·84
14	6·82	0·17	1·71	4·34	0·02	1·6619	1·6545	1·6476	17·70
17	7·01	0·28	2·11	3·67	0·29	1·6671	1·6595	1·6540	17·99
20	7·42	0·69	2·02	2·67	0·00	[1·672]	[1·6667]	[1·656]	18·14
22	7·77	0·19	2·48	4·09	0·04	1·6605	1·6490	1·6454	—
26	7·83	0·15	1·85	4·83	0·08	1·6517	1·6384	1·6329	18·08
29	7·85	0·00	1·00	5·95	0·37	1·6350	1·6270	1·6180	17·94
30	7·84	0·00	1·27	5·51	0·16	1·6410	1·6280	1·6162	18·02
35	7·96	0·00	1·22	5·91	0·03	1·6404	1·6301	1·6195	—
43	7·76	0·03	0·32	6·04	0·31	—	[1·62]	—	17·94
M_o	7·24	0·14	1·78	4·54	0·16	1·6577	1·6477	1·6397	—
M_x	7·18	0·19	1·64	4·41	0·19	—	—	—	17·95

* The numbers are those assigned by J. C. Rabbitt, Amer. Min., 1948, vol. 33, p. 263 [M.A. **10**–416]; reference should be made to this paper for localities and other information.

therefore not used for the calculation of the partial regression equations. An attempt was made to derive equations using all the remaining data, and assigning a weight of 2 to J. C. Rabbitt's data in view of their superior accuracy; but the equations so deduced gave residuals[1] for the superior data that were unexpectedly high and distinctly biased, suggesting that not enough weight had been given to the superior data.

It would indeed appear that the difference between the accuracy and coherence of J. C. Rabbitt's optical data and that of the rest cannot adequately be met by a weighting factor, and as Rabbitt's data appear to cover the composition field reasonably well, a fresh start was made, using them alone (analyses 1, 8, 9, 14, 17, 22, 26, 29, 30, and 35); their number is uncomfortably small, and it is probable that the equations now deduced will require considerable amendment and extension when more data are available.

Preliminary preparation of the data; formation of the matrix equations.

Having ascertained the relevant variables and selected the data, the next step is to reduce the data to a form suitable for computation. First, the mean of each physical quantity and compositional parameter over all the sets of observations is calculated; this is done for our example in

TABLE IIA. Selected optical data for analysed specimens of anthophyllite, expressed as differences from the means (M_0) of table I. $S = \mathrm{Si}-\overline{\mathrm{Si}}$, $T = \mathrm{Ti}+\mathrm{Fe}'''-\overline{\mathrm{Ti}+\mathrm{Fe}'''}$, $F = \mathrm{Fe}''+\mathrm{Mn}-\overline{\mathrm{Fe}''+\mathrm{Mn}}$, $M = \mathrm{Mg}-\overline{\mathrm{Mg}}$, $C = \mathrm{Ca}+\mathrm{Na}+\mathrm{K}-\overline{\mathrm{Ca}+\mathrm{Na}+\mathrm{K}}$, $\Gamma = \gamma-\bar{\gamma}$, $\mathrm{B} = \beta-\bar{\beta}$, $\mathrm{A} = \alpha-\bar{\alpha}$ where the bar indicates a mean.

No.	$S\times10^2$.	$T\times10^2$.	$F\times10^2$.	$M\times10^2$.	$C\times10^2$.	$\Gamma\times10^4$.	$\mathrm{B}\times10^4$.	$\mathrm{A}\times10^4$.
1	104	1	45	−117	25	204	193	169
8	−74	21	5	76	−8	141	153	156
9	−64	−2	30	−61	4	118	126	123
14	−42	3	−7	−20	14	42	68	79
17	−23	14	33	−87	13	94	118	143
22	53	5	70	−45	−12	28	13	56
26	59	1	7	29	−8	−60	−93	−68
29	61	−14	−78	141	21	−227	−207	−217
30	60	−14	−51	97	0	−167	−197	−235
35	72	−14	−56	137	−13	−173	−176	−209

table I; it will usually be sufficient if the several means are taken to the same number of decimal places as the quantities being averaged. Next, all the variables are expressed in the form of differences from their respective means, as in tables IIA and IIB.

[1] That is, differences between the observed optical data and the values calculated from the regression equation.

TABLE IIB. X-ray data for analysed specimens of anthophyllite, expressed as differences from the means (M_x) of table I, together with the squares and products required for the derivation of a regression equation, and the calculated values and residuals obtained from the resulting equation. $X = b - \bar{b}$; $\hat{X}$ is the calculated value of X; other symbols as in table I.

No.	$S\times10^2$.	$F\times10^2$.	$M\times10^2$.	$C\times10^2$.	$X\times10^2$.	$S^2\times10^4$.	$SF\times10^4$.	$SM\times10^4$.	$SC\times10^4$.	$F^2\times10^4$.	$FM\times10^4$.
1	−98	59	−104	23	1	9604	−5782	10192	−2254	3481	−6136
8	−68	19	−63	−11	−9	4624	−1292	4284	748	361	−1197
9	−58	44	−48	−7	−11	3364	−2552	2784	406	1936	−2112
14	−36	7	−7	−17	−25	1296	−252	252	612	49	−49
17	−17	47	−70	10	4	289	−799	1190	−170	2209	−3290
20	24	38	−174	−19	19	576	912	−4176	−456	1444	−6612
26	65	21	42	−11	13	4225	1365	2730	−715	441	882
29	67	−64	154	18	−1	4489	−4288	10318	1206	4096	9856
30	66	−37	110	−3	7	4356	−2442	7260	−198	1369	−4070
43	58	−132	163	12	−1	3364	−7656	9454	696	17424	−21516
Σ	—	—	—	—	—	36187	−22786	44288	−125	32810	−53956

No.	$FC\times10^4$.	$M^2\times10^4$.	$MC\times10^4$.	$C^2\times10^4$.	$XS\times10^4$.	$XF\times10^4$.	$XM\times10^4$.	$XC\times10^4$.	$X^2\times10^4$.	$\hat{X}\times10^2$.	$(X-\hat{X})\times10^2$.
1	1359	10816	−2392	529	−98	59	−104	23	1	−5	6
8	−209	3969	693	121	612	−171	567	99	81	−15	6
9	−308	2304	336	49	638	−484	528	77	121	−13	2
14	−119	49	119	289	900	−175	175	425	625	−16	−9
17	470	4900	−700	100	−68	188	−280	40	16	8	−4
20	−722	30276	3306	361	456	722	−3306	−361	361	22	−3
26	−231	1764	−462	121	845	273	546	−143	169	8	5
29	−1152	23716	2772	324	−67	64	−154	−18	1	5	−6
30	111	12100	−330	9	462	−259	770	−21	49	3	4
43	−1584	26569	1956	144	−58	132	−163	−12	1	0	−1
Σ	−2385	116463	5298	2047	3622	349	−1421	109	1425	—	—

Now in the general case we may have N sets of observations, each set referring to m specified physical properties (X, Y, Z,....) which are dependent variables and each of which is to be correlated with n independent variables such as the parameters of composition (A, B, C,...; or in practice SiO_2, Al_2O_3, &c.). And all these variables have been expressed in the form of differences from their respective means. We can therefore set up for each physical quantity N equations of the type:[1]

$$X_i = a_x A_i + b_x B_i + c_x C_i ... + p_x P_i \ (i = 1,2,..., N),$$

where a_x, b_x, c_x,... are constants, n in number.

Considering in the first place only the N equations for the one physical quantity X, the squares of the coefficients A_i, B_i, C_i,... and their products in pairs A_iB_i, A_iC_i, A_iD_i,..., B_iC_i, B_iD_i,..., C_iD_i,... are evaluated and tabulated as in table IIB (there will be $n(n+1)/2$ such terms). The products X_iA_i, X_iB_i, X_iC_i,... are also calculated (n terms). Corresponding products are now summed over the N equations, and from the sums a matrix equation in a_x, b_x, c_x,... can be set up:

$$\begin{Vmatrix} \sum A_i^2 & \sum A_iB_i & \sum A_iC_i & \dots & \sum A_iP_i \\ \sum A_iB_i & \sum B_i^2 & \sum B_iC_i & \dots & \sum B_iP_i \\ \sum A_iC_i & \sum B_iC_i & \sum C_i^2 & \dots & \sum C_iP_i \\ \cdot & \cdot & \cdot & \dots & \cdot \\ \cdot & \cdot & \cdot & \dots & \cdot \\ \cdot & \cdot & \cdot & \dots & \cdot \\ \sum A_iP_i & \sum B_iP_i & \sum C_iP_i & \dots & \sum P_i^2 \end{Vmatrix} \cdot \begin{Vmatrix} a_x \\ b_x \\ c_x \\ \cdot \\ \cdot \\ \cdot \\ P_x \end{Vmatrix} = \begin{Vmatrix} \sum X_iA_i \\ \sum X_iB_i \\ \sum X_iC_i \\ \cdot \\ \cdot \\ \cdot \\ \sum X_iP_i \end{Vmatrix} (i = 1,2,3,..., N).$$

A similar equation can be set up for each of the other physical

[1] It will be noticed that the equations are homogeneous, containing no constant term. In general, a regression equation correlating one dependent variable with n independent variables will contain a constant term, making $n+1$ constants, and in the derivation of the equation determinants of order $n+1$ will be involved. But if all the variables are expressed as differences from the mean, the constant term becomes zero. For if we assume that the constant term is ζ_x, we have $X_i = \zeta_x + a_xA_i + b_xB_i + c_xC_i + ...$; and summing, $\sum X_i = N\zeta_x + a_x \sum A_i + b_x \sum B_i + c_x \sum C_i + ...$ ($i = 1,2,..., N$); but if X_i, A_i, &c. are measured from their several means, the sums $\sum X_i$, $\sum A_i$,... are all zero; hence $\zeta_x = 0$. This elimination of the constant term reduces the order of the determinants involved in the derivation of the regression equations from $n+1$ to n, which amply repays the labour of expressing all the variables as differences from their means. It will also be obvious that if the number of sets of observations, N, is less than the number of independent variables, n, the system of N equations has no definite solution; if $N = n$ a solution is possible; and if, as will normally be the case, $N > n$, the equations will form an inconsistent system, from which, however, an optimum solution can be derived by the method of least squares; in what follows, we assume that $N > n$ and apply the method of least squares.

quantities, Y, Z,...; but provided[1] that all the physical quantities relate to the same N sets of values of the independent variables A, B, C,..., the left-hand matrix will be the same in all the equations, which can therefore all be combined into a single equation:

$$\begin{Vmatrix} \Sigma A_i^2 & \Sigma A_iB_i & \Sigma A_iC_i \dots & \Sigma A_iP_i \\ \Sigma A_iB_i & \Sigma B_i^2 & \Sigma B_iC_i \dots & \Sigma B_iP_i \\ \Sigma A_iC_i & \Sigma B_iC_i & \Sigma C^2 \dots & \Sigma C_iP_i \\ \cdot & \cdot & \cdot \dots & \cdot \\ \cdot & \cdot & \cdot \dots & \cdot \\ \cdot & \cdot & \cdot \dots & \cdot \\ \Sigma A_iP_i & \Sigma B_iP_i & \Sigma C_iP_i \dots & \Sigma P_i^2 \end{Vmatrix} \cdot \begin{Vmatrix} a_x & a_y & a_z \dots \\ b_x & b_y & b_z \dots \\ c_x & c_y & c_z \dots \\ \cdot & \cdot & \cdot \dots \\ \cdot & \cdot & \cdot \dots \\ \cdot & \cdot & \cdot \dots \\ p_x & p_y & p_z \dots \end{Vmatrix} = \begin{Vmatrix} \Sigma X_iA_i & \Sigma Y_iA_i & \Sigma Z_iA_i \dots \\ \Sigma X_iB_i & \Sigma Y_iB_i & \Sigma Z_iB_i \dots \\ \Sigma X_iC_i & \Sigma Y_iC_i & \Sigma Z_iC_i \dots \\ \cdot & \cdot & \cdot \dots \\ \cdot & \cdot & \cdot \dots \\ \cdot & \cdot & \cdot \dots \\ \Sigma X_iP_i & \Sigma Y_iP_i & \Sigma Z_iP_i \dots \end{Vmatrix}.$$

In this equation, it will be seen that the left-hand matrix (which is a symmetrical square matrix of order $n \times n$) contains only the sums of squares and products of the independent variables; the central matrix contains all the desired coefficients of the regression equations; and the dependent variables, the m quantities X, Y, Z,..., only occur in the right-hand matrix; the central and right-hand matrices are both of order $n \times m$.

To solve the equation (that is, to evaluate the central matrix of unknown constants), the most satisfactory procedure is to premultiply the right-hand matrix by the covariance matrix, which is the reciprocal of the left-hand matrix and, like it, is a symmetrical square matrix of order $n \times n$. To evaluate the covariance matrix, we must solve the equation:

$$\begin{Vmatrix} \Sigma A_i^2 & \Sigma A_iB_i & \Sigma A_iC_i & \dots & \Sigma A_iP_i \\ \Sigma A_iB_i & \Sigma B_i^2 & \Sigma B_iC_i & \dots & \Sigma B_iP_i \\ \Sigma A_iC_i & \Sigma B_iC_i & \Sigma C_i^2 & \dots & \Sigma C_iP_i \\ \cdot & \cdot & \cdot & \dots & \cdot \\ \cdot & \cdot & \cdot & \dots & \cdot \\ \cdot & \cdot & \cdot & \dots & \cdot \\ \Sigma A_iP_i & \Sigma B_iP_i & \Sigma C_iP_i & \dots & \Sigma P_i^2 \end{Vmatrix} \cdot \begin{Vmatrix} \xi_{11} & \xi_{12} & \xi_{13} & \dots & \xi_{1p} \\ \xi_{21} & \xi_{22} & \xi_{23} & \dots & \xi_{2p} \\ \xi_{31} & \xi_{32} & \xi_{33} & \dots & \xi_{3p} \\ \cdot & \cdot & \cdot & \dots & \cdot \\ \cdot & \cdot & \cdot & \dots & \cdot \\ \cdot & \cdot & \cdot & \dots & \cdot \\ \xi_{p1} & \xi_{p2} & \xi_{p3} & \dots & \xi_{pp} \end{Vmatrix} = \begin{Vmatrix} 1 & 0 & 0 & \dots & 0 \\ 0 & 1 & 0 & \dots & 0 \\ 0 & 0 & 1 & \dots & 0 \\ \cdot & \cdot & \cdot & \dots & \cdot \\ \cdot & \cdot & \cdot & \dots & \cdot \\ \cdot & \cdot & \cdot & \dots & \cdot \\ 0 & 0 & 0 & \dots & 1 \end{Vmatrix}.$$

The solution of this equation is $\xi_{ij} = D_{ij}/D_0$, where ξ_{ij} is the element appearing in the ith row and jth column of the covariance matrix, D_0 is the determinant of the left-hand matrix (the matrix of sum of squares and products of the independent variables), and D_{ij} is the determinant obtained by replacing the ith column of D_0 by the jth column of the right-hand unit matrix.

[1] This proviso will usually have been met during the preliminary selection and preparation of the data. If for any reason a different group of independent variables must be used for any particular physical quantity, the whole procedure, including the preparatory expression of the data as differences from their means, will have to be carried out separately for that physical quantity. This is exemplified in the case of anthophyllite by the data for the unit-cell dimension, b (tables I and IIB; compare table IIA).

A systematic procedure for the evaluation of the covariance matrix will be considered below; for the moment, we will assume that it has been evaluated. Then we can write:

$$\begin{Vmatrix} a_x & a_y & a_z \dots \\ b_x & b_y & b_z \dots \\ c_x & c_y & c_z \dots \\ \cdot & \cdot & \cdot \dots \\ \cdot & \cdot & \cdot \dots \\ \cdot & \cdot & \cdot \dots \\ p_x & p_y & p_z \dots \end{Vmatrix} = \begin{Vmatrix} \xi_{11} & \xi_{12} & \xi_{13} & \dots & \xi_{1p} \\ \xi_{21} & \xi_{22} & \xi_{23} & \dots & \xi_{2p} \\ \xi_{31} & \xi_{32} & \xi_{33} & \dots & \xi_{3p} \\ \cdot & \cdot & \cdot & \dots & \cdot \\ \cdot & \cdot & \cdot & \dots & \cdot \\ \cdot & \cdot & \cdot & \dots & \cdot \\ \xi_{p1} & \xi_{p2} & \xi_{p3} & \dots & \xi_{pp} \end{Vmatrix} \cdot \begin{Vmatrix} \sum X_i A_i & \sum Y_i A_i & \sum Z_i A_i \dots \\ \sum X_i B_i & \sum Y_i B_i & \sum Z_i B_i \dots \\ \sum X_i C_i & \sum Y_i C_i & \sum Z_i C_i \dots \\ \cdot & \cdot & \cdot \dots \\ \cdot & \cdot & \cdot \dots \\ \cdot & \cdot & \cdot \dots \\ \sum X_i P_i & \sum Y_i P_i & \sum Z_i P_i \dots \end{Vmatrix}.$$

Multiplying out, any element of the left-hand matrix of regression coefficients, say that in the ith row and jth column, is obtained by multiplying the several elements of the corresponding column (the jth) of the right-hand matrix by the elements of the corresponding row (the ith) of the covariance matrix and adding the products. Thus in the second row and third column:

$$b_z = \xi_{21} \sum Z_i A_i + \xi_{22} \cdot \sum Z_i B_i + \xi_{23} \cdot \sum Z_i C_i + \dots + \xi_{2p} \cdot \sum Z_i P_i.$$

Finally, we arrive at the desired regression equations:

$$\hat{X}_i = A_i a_x + B_i b_x + C_i c_x + \dots + P_i p_x \pm \hat{\sigma}_x;$$

$$\hat{Y}_i = A_i a_y + B_i b_y + C_i c_y + \dots + P_i p_y \pm \hat{\sigma}_y;$$

&c., or in matrix form:

$$\begin{Vmatrix} \hat{X}_i \\ \hat{Y}_i \\ \hat{Z}_i \\ \cdot \\ \cdot \end{Vmatrix} = \begin{Vmatrix} a_x & b_x & c_x & \dots & p_x \\ a_y & b_y & c_y & \dots & p_y \\ a_z & b_z & c_z & \dots & p_z \\ \cdot & \cdot & \cdot & \dots & \cdot \\ \cdot & \cdot & \cdot & \dots & \cdot \end{Vmatrix} \cdot \begin{Vmatrix} A_i \\ B_i \\ C_i \\ \cdot \\ P_i \end{Vmatrix} \pm \begin{Vmatrix} \hat{\sigma}_x \\ \hat{\sigma}_y \\ \hat{\sigma}_z \\ \cdot \\ \cdot \end{Vmatrix}.$$

It will be noticed that $\hat{X}_i$, $\hat{Y}_i$, &c., are used for the estimated values of the dependent variables, derived from the regression equations, and that each equation includes a standard deviation term, $\hat{\sigma}_x$, $\hat{\sigma}_y$, &c. The standard deviation terms may be derived in two ways, and it may be thought desirable to use both as a check, though exact agreement will not normally be obtained on account of the approximations used in the course of the calculations. One procedure is to compare the observed and estimated values of the dependent variables, forming the residuals

$X_i - \hat{X}_i$, $Y_i - \hat{Y}_i$, &c.; then $\hat{\sigma}_x^2 = \sum (X_i - \hat{X}_i)^2/N$, &c. The other procedure is to use the relation

$$N\hat{\sigma}_x^2 = \sum X_i^2 - a_x \sum X_i A_i - b_x \sum X_i B_i - c_x \sum X_i C_i - \ldots - p_x \sum X_i P_i.$$

For our example of anthophyllite, it would take too much space to follow through the derivation of the regression equations for all the dependent variables in full detail, but the first stage of the procedure for the cell-side b is illustrated in table IIB; the sums of squares and products in this table, and the corresponding sums for table IIA (the derivation of which is not shown) lead to the two matrix equations:

$$\begin{Vmatrix} 3{\cdot}6187 & -2{\cdot}2786 & 4{\cdot}4288 & -0{\cdot}0125 \\ -2{\cdot}2786 & 3{\cdot}2810 & -5{\cdot}3956 & -0{\cdot}2385 \\ 4{\cdot}4288 & -5{\cdot}3956 & 11{\cdot}6463 & 0{\cdot}5298 \\ -0{\cdot}0125 & -0{\cdot}2385 & 0{\cdot}5298 & 0{\cdot}2047 \end{Vmatrix} \cdot \begin{Vmatrix} s_x \\ f_x \\ m_x \\ c_x \end{Vmatrix} = \begin{Vmatrix} 0{\cdot}3622 \\ 0{\cdot}0349 \\ 0{\cdot}1421 \\ 0{\cdot}0109 \end{Vmatrix}, \text{ and}$$

$$\times 10^2$$

$$\begin{Vmatrix} 4{\cdot}1476 & -0{\cdot}4356 & -1{\cdot}5162 & 4{\cdot}8148 & -0{\cdot}2236 \\ -0{\cdot}4356 & 0{\cdot}1265 & 0{\cdot}3478 & -0{\cdot}8315 & -0{\cdot}0175 \\ -1{\cdot}5162 & 0{\cdot}3478 & 2{\cdot}0858 & -3{\cdot}6770 & -0{\cdot}0314 \\ 4{\cdot}8148 & -0{\cdot}8315 & -3{\cdot}6770 & 8{\cdot}2080 & -0{\cdot}1436 \\ -0{\cdot}2226 & -0{\cdot}0175 & -0{\cdot}0314 & -0{\cdot}1436 & 0{\cdot}1888 \end{Vmatrix} \cdot \begin{Vmatrix} s_\Gamma & s_B & s_A \\ t_\Gamma & t_B & t_A \\ f_\Gamma & f_B & f_A \\ m_\Gamma & m_B & m_A \\ c_\Gamma & c_B & c_A \end{Vmatrix} = \begin{Vmatrix} -8{\cdot}1507 & -8{\cdot}6495 & -8{\cdot}6471 \\ 1{\cdot}2389 & 1{\cdot}3102 & 1{\cdot}4811 \\ 5{\cdot}3684 & 5{\cdot}2956 & 5{\cdot}9468 \\ -12{\cdot}5707 & -12{\cdot}9211 & -13{\cdot}8756 \\ 0{\cdot}1760 & 0{\cdot}2208 & 0{\cdot}1167 \end{Vmatrix}$$

In both these equations the coefficients s_x, s_Γ, s_B, s_A are the coefficients of S in the desired regression equations for X, Γ, B, and A respectively (these symbols are defined in tables I and II), while t, f, m, and c are the corresponding coefficients of T, F, M, and C; and the power of 10 written over the first term of a matrix is to be understood, here and later, as multiplying every term of the matrix.

The two covariance matrices, which are the reciprocals of the left-hand matrices of these two equations, are next found by the procedure outlined below (p. 93):

$$\begin{Vmatrix} \xi_{ss} & \xi_{sf} & \xi_{sm} & \xi_{sc} \\ \xi_{fs} & \xi_{ff} & \xi_{fm} & \xi_{fc} \\ \xi_{ms} & \xi_{mf} & \xi_{mm} & \xi_{mc} \\ \xi_{cs} & \xi_{cf} & \xi_{cm} & \xi_{cc} \end{Vmatrix} = \begin{Vmatrix} 0{\cdot}6759 & 0{\cdot}1898 & -0{\cdot}1875 & 0{\cdot}7248 \\ 0{\cdot}1898 & 1{\cdot}2858 & 0{\cdot}5423 & 0{\cdot}1558 \\ -0{\cdot}1875 & 0{\cdot}5423 & 0{\cdot}4276 & -0{\cdot}4945 \\ 0{\cdot}7248 & 0{\cdot}1556 & -0{\cdot}4945 & 6{\cdot}3390 \end{Vmatrix} \text{ (for } X\text{) and}$$

$$\begin{Vmatrix} \xi_{ss} & \xi_{st} & \xi_{sf} & \xi_{sm} & \xi_{sc} \\ \xi_{ts} & \xi_{tt} & \xi_{tf} & \xi_{tm} & \xi_{tc} \\ \xi_{fs} & \xi_{ft} & \xi_{ff} & \xi_{fm} & \xi_{fc} \\ \xi_{ms} & \xi_{mt} & \xi_{mf} & \xi_{mm} & \xi_{mc} \\ \xi_{cs} & \xi_{ct} & \xi_{cf} & \xi_{cm} & \xi_{cc} \end{Vmatrix} = \begin{Vmatrix} 6{\cdot}891 & -18{\cdot}836 & -12{\cdot}072 & -11{\cdot}434 & -4{\cdot}325 \\ -18{\cdot}836 & 82{\cdot}349 & 36{\cdot}260 & 35{\cdot}962 & 18{\cdot}808 \\ -12{\cdot}072 & 36{\cdot}260 & 24{\cdot}058 & 21{\cdot}699 & 9{\cdot}634 \\ -11{\cdot}434 & 35{\cdot}962 & 21{\cdot}699 & 20{\cdot}348 & 8{\cdot}939 \\ -4{\cdot}325 & 18{\cdot}808 & 9{\cdot}634 & 8{\cdot}939 & 10{\cdot}342 \end{Vmatrix} \text{ (for } \Gamma\text{, B, and A).}$$

Applying the covariance matrix for X:

$$\begin{Vmatrix} s_x \\ f_x \\ m_x \\ c_x \end{Vmatrix} = \begin{Vmatrix} 0{\cdot}6759 & 0{\cdot}1898 & -0{\cdot}1875 & 0{\cdot}7248 \\ 0{\cdot}1898 & 1{\cdot}2858 & 0{\cdot}5423 & 0{\cdot}1558 \\ -0{\cdot}1875 & 0{\cdot}5423 & 0{\cdot}4276 & -0{\cdot}4945 \\ 0{\cdot}7248 & 0{\cdot}1558 & -0{\cdot}4945 & 6{\cdot}3390 \end{Vmatrix} \cdot \begin{Vmatrix} 0{\cdot}3622 \\ 0{\cdot}0349 \\ -0{\cdot}1421 \\ 0{\cdot}0109 \end{Vmatrix}$$

$$= \begin{Vmatrix} 0{\cdot}6759 \times 0{\cdot}3622 + 0{\cdot}1898 \times 0{\cdot}0349 + 0{\cdot}1875 \times 0{\cdot}1421 + 0{\cdot}7248 \times 0{\cdot}0109 \\ 0{\cdot}1898 \times 0{\cdot}3622 + 1{\cdot}2858 \times 0{\cdot}0349 - 0{\cdot}5423 \times 0{\cdot}1421 + 0{\cdot}1558 \times 0{\cdot}0109 \\ -0{\cdot}1875 \times 0{\cdot}3622 + 0{\cdot}5423 \times 0{\cdot}0349 - 0{\cdot}4276 \times 0{\cdot}1421 - 0{\cdot}4945 \times 0{\cdot}0109 \\ 0{\cdot}7248 \times 0{\cdot}3622 + 0{\cdot}1558 \times 0{\cdot}0349 + 0{\cdot}4945 \times 0{\cdot}1421 + 6{\cdot}3390 \times 0{\cdot}0109 \end{Vmatrix} = \begin{Vmatrix} 0{\cdot}286 \\ 0{\cdot}038 \\ -0{\cdot}115 \\ 0{\cdot}407 \end{Vmatrix}.$$

Hence we derive the regression equation for X:

$$\hat{X} = 0{\cdot}286\,S + 0{\cdot}038\,F - 0{\cdot}115\,M + 0{\cdot}407\,C,$$

with a variance

$$\begin{aligned}\hat{\sigma}_x^2 &= (\Sigma X^2 - s_x \Sigma XS - f_x \Sigma XF - m_x \Sigma XM - c_x \Sigma XC)/N \\ &= (0{\cdot}1425 - 0{\cdot}286 \times 0{\cdot}3622 - 0{\cdot}038 \times 0{\cdot}0349 - 0{\cdot}115 \times 0{\cdot}1421 - \\ &\qquad - 0{\cdot}407 \times 0{\cdot}0109)/10 \\ &= 0{\cdot}0017,\end{aligned}$$

and standard deviation $\hat{\sigma}_x = 0{\cdot}04$. Lastly, we replace the differences from the mean, S, F, M, C, and X, by the measured quantities Si, Fe″+Mn, Mg, Ca+Na+K, and b, arriving at the final form of our regression equation for b:

$$\begin{aligned}\hat{b} - 17{\cdot}95 &= 0{\cdot}286\,(\mathrm{Si} - 7{\cdot}18) + 0{\cdot}038\,(\mathrm{Fe''} + \mathrm{Mn} - 1{\cdot}64) - \\ &\qquad - 0{\cdot}165\,(\mathrm{Mg} - 4{\cdot}41) + 0{\cdot}407\,(\mathrm{Ca} + \mathrm{Na} + \mathrm{K} - 0{\cdot}19) \pm 0{\cdot}04,\end{aligned}$$

$$\begin{aligned}\hat{b} &= 16{\cdot}48 + 0{\cdot}286\,\mathrm{Si} + 0{\cdot}038\,(\mathrm{Fe''} + \mathrm{Mn}) - 0{\cdot}165\,\mathrm{Mg} + \\ &\qquad + 0{\cdot}407\,(\mathrm{Ca} + \mathrm{Na} + \mathrm{K}) \pm 0{\cdot}04.\end{aligned}$$

Similarly, we arrive at regression equations for γ, β, and α:

$$\begin{aligned}\hat{\gamma} &= 1{\cdot}7273 - 0{\cdot}0134\,\mathrm{Si} + 0{\cdot}0145\,(\mathrm{Ti} + \mathrm{Fe'''}) + 0{\cdot}0140\,(\mathrm{Fe''} + \mathrm{Mn}) + \\ &\qquad + 0{\cdot}0002\,\mathrm{Mg} - 0{\cdot}0030\,(\mathrm{Ca} + \mathrm{Na} + \mathrm{K}) \pm 0{\cdot}0009,\end{aligned}$$

$$\begin{aligned}\hat{\beta} &= 1{\cdot}7299 - 0{\cdot}0143\,\mathrm{Si} + 0{\cdot}0232\,(\mathrm{Ti} + \mathrm{Fe'''}) + 0{\cdot}0108\,(\mathrm{Fe''} + \mathrm{Mn}) - \\ &\qquad - 0{\cdot}0002\,\mathrm{Mg} - 0{\cdot}0015\,(\mathrm{Ca} + \mathrm{Na} + \mathrm{K}) \pm 0{\cdot}0014,\end{aligned}$$

$$\begin{aligned}\hat{\alpha} &= 1{\cdot}7009 - 0{\cdot}0113\,\mathrm{Si} + 0{\cdot}0368\,(\mathrm{Ti} + \mathrm{Fe'''}) + 0{\cdot}0120\,(\mathrm{Fe''} + \mathrm{Mn}) - \\ &\qquad - 0{\cdot}0012\,\mathrm{Mg} - 0{\cdot}0028\,(\mathrm{Ca} + \mathrm{Na} + \mathrm{K}) \pm 0{\cdot}0021.\end{aligned}$$

This completes the derivation of the desired regression equations. But we notice that they contain several very small coefficients, and we may reasonably doubt whether these differ significantly from zero. It is not difficult to test whether this is so, to test whether two coefficients that are nearly equal differ significantly from one another, and to decide how many significant figures are justifiable in each coefficient.

The standard deviations of the regression coefficients; tests of the significance of the regression coefficients.

An estimate of the standard deviations of the regression coefficients can very readily be made by using the covariance matrix. For if the estimated standard deviation of the dependent variable, X, is $\hat{\sigma}_x$, and

the estimated standard deviations of the regression coefficients a_x, b_x, c_x,... are $\hat{\sigma}_{x,a}$, $\hat{\sigma}_{x,b}$, &c., then $\hat{\sigma}^2_{x,a} = \sigma^2_x \xi_{11} N/(N-n-1)$, where N is the number of observed values of X and n the number of independent variables a, b, &c.; or in general for the kth in the series $\hat{\sigma}_{x,a}$, $\hat{\sigma}_{x,b}$,..., we have $\hat{\sigma}^2_{x,k} = \hat{\sigma}^2_x \xi_{kk} N/(N-n-1)$.

For the coefficients of the regression equation derived above for the b cell-dimension of anthophyllite, we have:

$$\hat{\sigma}^2_{b,s} = \hat{\sigma}^2_b \xi_{ss} N/(N-n-1) = 0{\cdot}0017 \times 0{\cdot}6759 \times 10/5 = 0{\cdot}0023;$$
$$\hat{\sigma}^2_{b,f} = 0{\cdot}0017 \times 1{\cdot}2858 \times 10/5 = 0{\cdot}0044;$$
$$\hat{\sigma}^2_{b,m} = 0{\cdot}0017 \times 0{\cdot}4276 \times 10/5 = 0{\cdot}0014;$$
$$\hat{\sigma}^2_{b,c} = 0{\cdot}0017 \times 6{\cdot}3390 \times 10/5 = 0{\cdot}0215$$

(note that the second factor in each is an element of the principal diagonal of the covariance matrix). Hence we have

$$\hat{\sigma}_{b,s} = 0{\cdot}048, \quad \hat{\sigma}_{b,f} = 0{\cdot}066, \quad \hat{\sigma}_{b,m} = 0{\cdot}037, \quad \hat{\sigma}_{b,c} = 0{\cdot}147,$$

and

$$s_b = 0{\cdot}286 \pm 0{\cdot}048, \quad f_b = 0{\cdot}038 \pm 0{\cdot}066, \quad m_b = -0{\cdot}115 \pm 0{\cdot}037,$$
$$c_b = 0{\cdot}407 \pm 0{\cdot}147.$$

It is obvious that the coefficient f_b, with a standard deviation much greater than the coefficient itself, is not significantly different from zero; but some quantitative test for the significance of the coefficients is desirable, and is, in fact, readily available. Most collections of tabulated statistical functions include a table of 'Student's ratio', t. This is the ratio of a quantity to its standard deviation, and the tabulated figures are the probabilities, for given values of Student's ratio and given numbers of degrees of freedom, that the quantity under consideration does not differ significantly from zero. It is not necessary for our purpose to consider the exact meaning of the term 'degrees of freedom' in statistical theory; it will suffice to say that in this connexion the number of degrees of freedom is $N-n-1$: N and n are defined above.[1] Applying this procedure to our example, the values of Student's ratio are

$$t_{b,s} = s_b/\hat{\sigma}_{b,s} = 6{\cdot}0, \quad t_{b,f} = 0{\cdot}57, \quad t_{b,m} = 3{\cdot}1, \quad t_{b,c} = 2{\cdot}8,$$

and there are 5 degrees of freedom; from the tables the corresponding

[1] In this connexion, it must not be forgotten that n is the number of independent variables *remaining*, not necessarily the original number the investigation started with; if any terms have been rejected from the regression equation as not significantly different from zero, n will be reduced accordingly.

probabilities that the several coefficients are not significantly different from zero are $P_{b,s} < 0{\cdot}01$, $P_{b,f}$ 0·6, $P_{b,m}$ 0·03, $P_{b,c}$ 0·04. The degree of probability that a coefficient is significantly different from zero is of course a matter for the judgement of the individual worker, but this simple technique does provide an objective measure of that probability. We have accepted a probability of 0·40 or less as justifying the inclusion of the term in question, and on this basis the term in F may be rejected from the equation for b as not justified by the data at present available, while the terms in S, M, and C must be retained.[1] This does not imply that b does not vary with $Fe''+Mn$, but merely that the data at present available are not adequate to prove any such variation.

When any coefficient of a regression equation has been shown not to differ significantly from zero, we proceed to eliminate that term and readjust the remaining coefficients to give the best representation of the observations. But before considering the procedure for this readjustment it is desirable to consider the possibility that two coefficients may not differ significantly from one another, when it would be proper to replace the two separate terms in two independent variables, say B and G, by a single term in their sum, $B+G$, adjusting all the coefficients accordingly.

To test whether two coefficients, say b_x and g_x, differ significantly from one another, we divide their difference by its standard deviation to obtain the appropriate Student's ratio $t_{x,b-g}$; then from the tables, with this value of t and $N-n-1$ degrees of freedom, we derive the probability $P_{x,b-g}$ that the two coefficients do not differ significantly from one another. The standard deviation $\hat{\sigma}_{x,b-g}$ of the difference $b-g$ is given, in terms of the standard deviation $\hat{\sigma}_x$ of the dependent variable, by the relation: $\hat{\sigma}^2_{x,b-g} = \hat{\sigma}^2_x(\xi_{22}+\xi_{77}-2\xi_{27})\, N/(N-n-1)$; the appropriate elements of the covariance matrix are ξ_{22}, ξ_{77}, and ξ_{27} because b and g are second and seventh in the series a, b, c,....

Considering the remaining, adjusted coefficients of the regression equation for the b cell-dimension of anthophyllite, after elimination of f_b (see p. 84) it is obvious that since all the coefficients have been shown to differ significantly from zero, the negative coefficient $m_b = -0{\cdot}131$ must differ significantly from the two positive coefficients $s_b = 0{\cdot}280$ and $c_b = 0{\cdot}403$. The difference of the two positive coefficients, $c_b - s_b$,

[1] An alternative test, less rigorously based but quite adequate for most investigations, is to accept any coefficient as probably significant if it is greater than its standard deviation. This test has the advantage of not requiring tables of Student's ratio.

is 0·123, and the standard variation of this difference is given by:

$$\begin{aligned}\hat{\sigma}^2_{b,c-s} &= \hat{\sigma}_b(\xi_{cc}+\xi_{ss}-2\xi_{cs})N/(N-n-1)\\ &= 0{\cdot}00181\ (0{\cdot}6479+6{\cdot}3201-2\times 0{\cdot}7018)\ 10/(10-3-1)\\ &= 0{\cdot}0168,\end{aligned}$$

whence $\hat{\sigma}_{b,c-s} = 0{\cdot}130$; then $t_{c-s} = (c-s)/\hat{\sigma}_{b,c-s} = 0{\cdot}123/0{\cdot}130 = 0{\cdot}95$, and with six degrees of freedom, we find from the tables $P_{b,c-s} = 0{\cdot}38$; this is on the borderline of significance, but we retain the separate terms because of the chemical contrast between the independent variables involved, Si and Ca+Na+K.

Procedure for the elimination of a non-significant term from the regression equations.

If it has been established that one of the coefficients, say the kth in the series a, b, c,..., does not differ significantly from zero, it can be eliminated, and the remaining coefficients adjusted by recalculating a new covariance matrix of reduced order $(n-1)$; the general term, ξ'_{ij}, of the new matrix is derived from the corresponding term, ξ_{ij}, of the old (not counting the kth row and column of the old matrix) by the relation

$$\xi'_{ij} = \xi_{ij}-\xi_{ik}\cdot\xi_{jk}/\xi_{kk}.$$

The matrix of coefficients is then found by postmultiplying the reduced covariance matrix by the matrix of sums and products of the independent variables, excluding the row appropriate to the eliminated independent variable.

Thus with the equations for the refractive indices of anthophyllite, we have originally (omitting the terms below the principal diagonal of the covariance matrix, which may be inserted by symmetry):

$$\begin{Vmatrix} s_\Gamma & s_B & s_A \\ t_\Gamma & t_B & t_A \\ f_\Gamma & f_B & f_A \\ m_\Gamma & m_B & m_A \\ c_\Gamma & c_B & c_A \end{Vmatrix} = \begin{Vmatrix} 6{\cdot}891 & -18{\cdot}836 & -12{\cdot}072 & -\textit{11·434} & -4{\cdot}325 \\ & 82{\cdot}349 & 36{\cdot}260 & \textit{35·962} & 18{\cdot}808 \\ & & 24{\cdot}058 & \textit{21·699} & 9{\cdot}634 \\ & & & \textit{20·348} & \textit{8·939} \\ & & & & 10{\cdot}342 \end{Vmatrix} \cdot \begin{Vmatrix} -8{\cdot}1507 & -8{\cdot}6495 & -8{\cdot}6471 \\ 1{\cdot}2389 & 1{\cdot}3102 & 1{\cdot}4811 \\ 5{\cdot}3684 & 5{\cdot}2956 & 5{\cdot}9468 \\ \textit{-12·5707} & \textit{-12·9211} & \textit{-13·8756} \\ 0{\cdot}1760 & 0{\cdot}2208 & 0{\cdot}1167 \end{Vmatrix} \times 10^{-2}.$$

To adjust the coefficients when we eliminate the terms in M we must eliminate the fourth row and fourth column of the covariance matrix (the italicized terms) by the above procedure, and delete the fourth row of

the other two matrices. For the reduced covariance matrix we may write:

$$\underset{\text{STFN}}{\|\xi\|} =$$

$$\left\|\begin{array}{llll} 6{\cdot}891-\dfrac{(11{\cdot}434)^2}{20{\cdot}348} & -18{\cdot}836+\dfrac{(11{\cdot}434\times 35{\cdot}962)}{20{\cdot}348} & -12{\cdot}072+\dfrac{(11{\cdot}434\times 21{\cdot}699)}{20{\cdot}348} & -4{\cdot}325+\dfrac{(11{\cdot}434\times 8{\cdot}939)}{20{\cdot}348} \\ & 82{\cdot}349-\dfrac{(35{\cdot}962)^2}{20{\cdot}348} & 36{\cdot}260-\dfrac{(35{\cdot}962\times 21{\cdot}699)}{20{\cdot}348} & 18{\cdot}808-\dfrac{(35{\cdot}962\times 8{\cdot}939)}{20{\cdot}348} \\ & & 24{\cdot}058-\dfrac{(21{\cdot}699)^2}{20{\cdot}348} & 9{\cdot}364-\dfrac{(21{\cdot}699\times 8{\cdot}939)}{20{\cdot}348} \\ & & & 10{\cdot}342-\dfrac{(8{\cdot}939)^2}{20{\cdot}348} \end{array}\right\|.$$

The pattern in this expression should be readily apparent. After evaluating it and making a second reduction by the same procedure to eliminate the terms in c, which also prove to be non-significant, we arrive at an equation for the remaining coefficients:

$$\left\|\begin{array}{lll} s_\Gamma & s_B & s_A \\ t_\Gamma & t_B & t_A \\ f_\Gamma & f_B & f_A \end{array}\right\| = \left\|\begin{array}{lll} 0{\cdot}3901 & 1{\cdot}0443 & 0{\cdot}1103 \\ & 17{\cdot}3789 & -2{\cdot}1372 \\ & & 0{\cdot}9166 \end{array}\right\| \times 10^{-2} \left\|\begin{array}{lll} -8{\cdot}1507 & -8{\cdot}6495 & -8{\cdot}6471 \\ 1{\cdot}2389 & 1{\cdot}3102 & 1{\cdot}4811 \\ 5{\cdot}3684 & 5{\cdot}2956 & 5{\cdot}9468 \end{array}\right\|$$

$$= \left\|\begin{array}{lll} -0{\cdot}0129 & -0{\cdot}0142 & -0{\cdot}0117 \\ 0{\cdot}0180 & 0{\cdot}0242 & 0{\cdot}0400 \\ 0{\cdot}0143 & 0{\cdot}0110 & 0{\cdot}0133 \end{array}\right\|.$$

When the new regression equations with these coefficients are tested, it is found that while all the coefficients are significantly different from zero, t_Γ and f_Γ (0·0180 and 0·0143) are not significantly different from one another. We can therefore compound these terms if we wish, and we now consider how to do so.

Procedure for compounding two terms of a regression equation whose coefficients do not differ significantly from one another.

If it has been established that two of the coefficients of a regression equation, say the kth and the qth of the series a, b, c,..., do not differ significantly from one another, they can be adjusted to equality, and the remaining coefficients adjusted, by recalculating a new covariance matrix of reduced order $(n-1)$. The new, adjusted column and row, replacing the kth and the qth of the old matrix, can retain the place of either, the other being deleted; if we retain it in the kth place, the general term, ξ'_{ij}, of the new matrix will be derived from the corresponding term, ξ_{ij}, of the old (not counting the qth row and qth column of the old matrix) by the relation:

$$\xi'_{ij} = \xi_{ij} - (\xi_{ik} - \xi_{iq})(\xi_{jk} - \xi_{jq})/(\xi_{kk} + \xi_{qq} - 2\xi_{kq}),$$

except when i or $j = k$ or q; the terms (except ξ_{kk}) of the new kth row and kth column are given by:

$$\xi'_{ik} = (\xi_{ik}+\xi_{iq})-\{\xi_{ik}\xi_{kk}+\xi_{iq}\xi_{qq}-(\xi_{ik}+\xi_{iq})\,\xi_{kq}\}/(\xi_{kk}+\xi_{qq}-2\xi_{kq}),$$

while the new term $\xi'_{kk} = (\xi_{kk}\xi_{qq}-\xi^2_{kq})/(\xi_{kk}+\xi_{qq}-2\xi_{kq})$. The sum of products matrix $\|\sum X_i A_i\|$ will also be modified, the term $\sum X_i K_i$ being replaced by $\sum X_i K_i+\sum X_i Q_i$, and $\sum X_i Q_i$ omitted.

If we apply this procedure[1] to the coefficients of the regression equation for Γ, two of which (t_Γ and f_Γ) have been found not to differ significantly, we shall start from the equation:

$$\begin{Vmatrix} s_\Gamma \\ t_\Gamma \\ f_\Gamma \end{Vmatrix} = \begin{Vmatrix} 0{\cdot}3901 & 1{\cdot}0443 & 0{\cdot}1103 \\ & 17{\cdot}3789 & -2{\cdot}1372 \\ & & 0{\cdot}9166 \end{Vmatrix} \bullet \begin{Vmatrix} \times 10^{-2} \\ -8{\cdot}1507 \\ 1{\cdot}2389 \\ 5{\cdot}3684 \end{Vmatrix}.$$

The second and third rows (and columns) of the covariance matrix are to be compounded. Only one term of the new matrix, ξ'_{ss}, is derived by the first of the above formulae:

$$\xi'_{ss} = 0{\cdot}3901-(1{\cdot}0443-0{\cdot}1103)^2/(17{\cdot}3789+0{\cdot}9166+2\times 2{\cdot}1372)$$
$$= 0{\cdot}3525.$$

The symmetrically equal pair of terms, ξ'_{st} and ξ'_{ts}, are derived by the second formula:

$$\xi'_{st} = 1{\cdot}0443+0{\cdot}1103-$$
$$-\{1{\cdot}0443\times 17{\cdot}3789+0{\cdot}1103\times 0{\cdot}9166+2{\cdot}1372\,(1{\cdot}0443+0{\cdot}1103)\}/$$
$$(17{\cdot}3789+0{\cdot}9166+2\times 2{\cdot}1372)$$
$$= 0{\cdot}2365.$$

And the last term is derived by the third formula:

$$\xi'_{tt} = (17{\cdot}3789\times 0{\cdot}9166-2{\cdot}1372^2)/(17{\cdot}3789+0{\cdot}9166+2\times 1{\cdot}1372)$$
$$= 0{\cdot}5026.$$

We now have:

$$\begin{Vmatrix} s_\Gamma \\ (t+f)_\Gamma \end{Vmatrix} = \begin{Vmatrix} 0{\cdot}3525 & 0{\cdot}2365 \\ 0{\cdot}2365 & 0{\cdot}5026 \end{Vmatrix} \bullet \begin{Vmatrix} \times 10^{-2} \\ -8{\cdot}1507 \\ 1{\cdot}2389+5{\cdot}3684 \end{Vmatrix} \text{ giving } \begin{matrix} s = 0{\cdot}0130, \\ (t+f) = 0{\cdot}0142. \end{matrix}$$

[1] If the covariance matrix is of low order, as in the present case, it may be simpler to recompute the new matrix from the beginning rather than find it by this process. Referring back to the original matrix of sums of squares and products (the fifth-order square matrix on p. 80), the fourth and fifth columns and fourth and fifth rows, containing M and C, are simply suppressed; to write the second and third rows and columns, they are just added together, $\sum A_i(B_i+C_i) = \sum A_iB_i+\sum A_iC_i$, except for the four terms where the second and third columns cross the second and third rows; these four terms $\sum T^2$, $\sum F^2$, and $\sum TF$ (twice) are united by the relation $\sum (T+F)^2 = \sum T^2+\sum F^2+2\sum FT$.

The regression equations for anthophyllite.

In the foregoing discussion the selected example, anthophyllite, has been discussed at considerably greater length than will normally be necessary, in order to illustrate the general techniques. For instance, it may often be thought undesirable, or at least unnecessary, to unite two terms of a regression equation whose coefficients are not significantly different; it will often be quite obvious that certain coefficients are not significantly different from zero, without formal calculation of their standard deviations, Student's ratio, and the appropriate probabilities; and when terms are to be deleted or compounded and the coefficients adjusted, it will often be simpler to start anew from a reduced matrix of sums and products rather than to reduce the covariance matrix. But it was felt desirable to set out the procedure for all these operations, since they will sometimes be necessary.

So far as the selected example, anthophyllite, is concerned, we may set out our conclusions in four regression equations,[1] each with its standard deviation; and as a kind of appendix we may add the standard deviations of the several coefficients.[2]

$$\gamma = 1{\cdot}7249 - 0{\cdot}0130\ \mathrm{Si} + 0{\cdot}0140\ (\mathrm{Ti} + \mathrm{Fe'''} + \mathrm{Fe''} + \mathrm{Mn}) \pm 0{\cdot}0012,$$

$$\beta = 1{\cdot}7275 - 0{\cdot}0142\ \mathrm{Si} + 0{\cdot}024\ (\mathrm{Ti} + \mathrm{Fe'''}) + 0{\cdot}0110\ (\mathrm{Fe''} + \mathrm{Mn}) \pm 0{\cdot}0015,$$

$$\alpha = 1{\cdot}6951 - 0{\cdot}0117\ \mathrm{Si} + 0{\cdot}040\ (\mathrm{Ti} + \mathrm{Fe'''}) + 0{\cdot}0133\ (\mathrm{Fe''} + \mathrm{Mn}) \pm 0{\cdot}0025,$$

$$b(\text{Å.}) = 16{\cdot}44 + 0{\cdot}28\ \mathrm{Si} - 0{\cdot}13\ \mathrm{Mg} + 0{\cdot}40\ (\mathrm{Ca} + \mathrm{Na} + \mathrm{K}) \pm 0{\cdot}04,$$

$$\hat{\sigma}_{\gamma,s} = 0{\cdot}0007, \quad \hat{\sigma}_{\gamma,t} = 0{\cdot}0008, \quad \hat{\sigma}_{\beta,s} = 0{\cdot}0009, \quad \hat{\sigma}_{\beta,t} = 0{\cdot}004,$$

$$\hat{\sigma}_{\beta,f} = 0{\cdot}0015, \quad \hat{\sigma}_{\alpha,s} = 0{\cdot}0015, \quad \hat{\sigma}_{\alpha,t} = 0{\cdot}009, \quad \hat{\sigma}_{\alpha,f} = 0{\cdot}0025,$$

$$\hat{\sigma}_{b,s} = 0{\cdot}05, \quad \hat{\sigma}_{b,m} = 0{\cdot}03, \quad \hat{\sigma}_{b,c} = 0{\cdot}14.$$

It should, perhaps, be emphasized that the standard deviations of the dependent variables (γ, β, α, and b) cannot be taken as a true measure of the accuracy with which the given equations reproduce the true correlation of the physical data with the chemical composition of anthophyllite; they merely measure the accuracy with which the equations reproduce the selected data from which they are derived. It is therefore desirable to compare the whole of the available observations with the calculated values derived from the regression equations.

[1] The a and c cell-dimensions appear to be constant, within the experimental error of the available data.

[2] These serve as an indication of the amounts by which the several coefficients can be varied without gravely upsetting the agreement between observed and calculated values (after appropriate adjustment of the constant term).

Graphs have been prepared in which the residuals (the differences between observed and calculated values) are plotted against each of the five composition parameters, but these do not reveal any evidence of

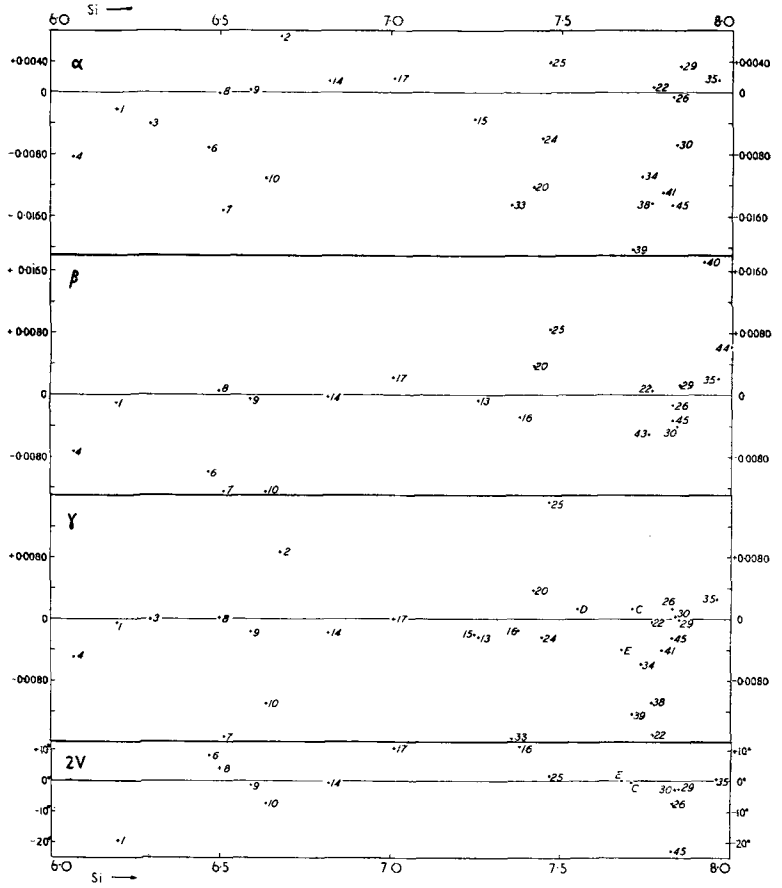

FIG. 1. Departures of the physical properties of anthophyllite from the values calculated from the regression equations, plotted against Si atoms per 24(O,OH,F). The numbers against the points are those used in J. C. Rabbitt's paper; C, D, and E refer to R. Pirani's data.

non-linear variation; one set of these graphs is shown in fig. 1. Graphs of the residuals plotted against the hydroxyl groups per 24(O,OH,F) and against fluorine have been prepared, and suggest that these variables

do not affect the optics, but in view of the limited number[1] of sets of observations and the doubtful accuracy of the chemical data on (OH)′ and F′, conclusions from these graphs must be viewed with great caution; on general grounds, we should expect F′ to lower the refractive indices notably.

On comparison of the observed and calculated refractive indices[2] it is noticeable that (except among the sets of superior data) there is a marked tendency for all three refractive indices to be high or low as a group, relative to the calculated values, rather than for one index to be high and another low (see analyses 6, 7, 10, 25, 33, 38, and 39). This result, which is emphasized by the fact that with only five exceptions (of which three are perhaps doubtful analyses) optic axial angles derived from the regression equations through the calculated refractive indices agree unexpectedly well with the observed values,[3] strongly supports the general accuracy of the regression equations, and suggests that several of the sets of observed refractive indices are subject to a systematic error. Two sets of data, for analyses 38 and 39, are probably low owing to the presence of 1·6 % and 2·4 % of adsorbed water respectively;[4] assuming the mixture law, this should lower their refractive indices by 0·005 and 0·007 respectively; actually, they give values of α 0·012 and 0·016 low, and of γ 0·009 and 0·011 low respectively.

There are two sets of data for which the above regression equations do not give satisfactory refractive indices: the Glen Urquhart gedrite and analysis no. 25. The latter is the 'picroamosite' of Serdyuchenko (1936),

[1] Where a large number of observations are available, the neglect of a variable will normally lead to fairly large residuals, which if plotted against the neglected variable will show a distinct trend. But if the neglected variable tends to follow one of those taken into account, its effect will be largely or wholly absorbed by the latter; and if there are only a few observations, false constants will probably be deduced, and the residuals will show only an irregular scatter. It is desirable that there should be at least ten times as many sets of observations as there are variables to be taken into account.

[2] As all the observed data have been published before, and all the calculated data may be derived from the regression equations, it seemed unnecessary to print a table of observed and calculated data, but such a table has been drawn up and deposited in the library of the Mineral Department of the British Museum (Natural History), where it may be consulted, together with a full set of graphs of the residuals plotted against the composition parameters, including OH′ and F′.

[3] The mean difference, excluding the above five analyses, is only 4°, against an expected 10° as calculated from the standard deviations of α, β, and γ. If the cited 2V(+) for R. Pirani's anthophyllite from Alpe de Brez is a misprint or error for 2V(−), there would be good agreement in this case also.

[4] G. H. Francis and M. H. Hey (in the press).

and has a remarkably high content of Fe_2O_3; the disagreement suggests that it may not be an ortho-amphibole, or alternatively that the chemical analysis is in error. For the Glen Urquhart gedrite, we know that the regression equations neglect the variation in total oxygen atoms per unit-cell, which in this material is definitely outside the limits of experimental error; if the observed data for this gedrite are taken, together with the empirical unit-cell contents, to establish approximate additional terms for the effect of extra oxygen (H_2O) in the regression equations, we find, approximately:

$$\alpha = \alpha_0 - 0{\cdot}023\,[\Sigma'\,(O,OH,F) - 24];$$

$$\beta = \beta_0 - 0{\cdot}020\,[\Sigma'\,(O,OH,F) - 24];$$

$$\gamma = \gamma_0 - 0{\cdot}016\,[\Sigma'\,(O,OH,F) - 24].$$

Here α_0, β_0, γ_0 are the values derived from the above incomplete regression equations (but with the composition in terms of one-quarter the empirical unit-cell contents), and Σ' (O,OH,F) is expressed for one-quarter the unit-cell contents. If these relations are even approximately true, they suggest that any anthophyllite having interstitial oxygen, as the Glen Urquhart gedrite has, will have strikingly low refractive indices in relation to its composition and should readily be detected by this property.

It may be of interest to note, in general terms, what the above regression equations imply regarding the variations in the cell-dimensions and optical properties of anthophyllite with composition (excluding possible variations in Σ (O,OH,F)). The *b*-axis of pure magnesio-anthophyllite should, from the regression equation, be 17·77 Å.; it is lowered by the substitution of Al_2 for MgSi, by 0·15 Å. for each MgSi per quarter unit-cell replaced, and raised by the substitution of Al_2 for Mg_3, by 0·13 Å. for each Mg per quarter unit-cell replaced; substitution of Ca, Na, or K for Mg appears to increase *b* markedly.

Turning to the optical properties, replacement of Mg by Fe″ leads to an increase in all three refractive indices, as usual, hardly any change in $\gamma-\alpha$, and a moderate increase in $\gamma-\beta$, and hence decrease in $2V_\gamma$; for one Fe″ replacing Mg per 24(O,OH,F) the increase is 0·014 for γ, 0·011 for β, and 0·013 for α. Replacement of MgSi by Al_2, in gedrites, leads to an unexpectedly large effect; Al_2 replacing MgSi increases α by 0·012, β by 0·014, γ by 0·013; $\gamma-\alpha$ and $2V_\gamma$ are increased, but $\gamma-\beta$ is decreased. Replacement of Al by Fe‴ or Ti (the data are insufficient to distinguish these) increases the refractive indices markedly. Replace-

ment of Mg by Ca or Na appears to have little effect. Replacement of 3Mg by 2Al also has no noticeable effect.

This work had just been completed when the author received a reprint of a paper by F. Hori,[1] who has applied similar methods to the correlation of the optical properties and chemical composition of the clinopyroxenes; in this group several substitutions that only occur to a minor extent in the anthophyllites go much farther, and it may be of interest to compare the general results of the two studies, though the structures are not so closely related that any close correspondence is to be expected. Hori's equations indicate that a replacement of Mg by Fe″ will increase the refractive indices markedly; for one atom Fe″ replacing Mg per 24 oxygen, the increases will be α 0·014, β and γ 0·016, a little greater than the effects in anthophyllite. Replacement of MgSi by Al_2, again on a comparable basis of 2Al per 24 oxygen, leads to increases in the refractive indices, namely α 0·017, β 0·025, γ 0·020, again greater effects than in anthophyllite. Replacement of Mg by Ca, Hori found, increases all three refractive indices in the clinopyroxenes by 0·003 to 0·005; it is possible that the effect is about the same in the anthophyllites, but the data are as yet inadequate to draw definite conclusions, and the same applies to most of the other possible replacements.

The use of weighting factors.

It may occasionally happen that there are a few sets of observations of superior or inferior accuracy, and it is desirable to weight such data appropriately. This is readily done by the use of a weighting factor, but in choosing this factor it should be remembered that, for example, physical data of special accuracy do not merit special weight unless the accompanying chemical analyses are of comparable quality to the rest of the analyses.

If it is decided to use a weighting factor for certain sets of data, those sets should be multiplied by their appropriate weighting factors, which may be fractional for inferior data, before adding them to derive the weighted means, which will be given by relations of the type $\bar{x} = \sum W_i x_i / \sum W_i$. The data are now tabulated in the form of differences from the weighted means, and a column of weighting factors is added; the squares and products are formed as usual, except that each square or product is multiplied by the appropriate weighting factor

[1] F. Hori, Sci. Papers Coll. General Education Univ. Tokyo, 1954, vol. 4, no. 1, p. 71.

TABLE III. The formation of the matrix equation for the regression coefficients when a weighting factor is applied to part of the data; exemplified by the data for the b cell-dimension of anthophyllite. Compare tables I and IIB.

No.	Si.	Mg.	Ca+Na+K.	b.	W.	WSi.	WMg.	W(Ca+Na+K).	Wb.	$S^*\times10^2$.	$M^*\times10^2$.	$C^*\times10^2$.
1	6·20	3·37	0·41	17·96	2	12·40	6·74	0·82	35·92	−88	−97	22
8	6·50	3·78	0·08	17·86	2	13·00	7·56	0·16	35·72	−58	−56	−11
9	6·60	3·93	0·12	17·84	2	13·20	7·86	0·24	35·68	−68	−41	−7
14	6·82	4·34	0·02	17·70	2	13·64	8·68	0·04	35·40	−26	0	−17
17	7·01	3·67	0·29	17·99	2	14·02	7·34	0·58	35·98	−7	−67	12
20	7·42	2·67	0·00	18·14	1	7·42	2·67	0·00	18·14	34	−67	−19
26	7·83	4·83	0·08	18·08	1	7·83	4·83	0·08	18·08	75	49	−11
29	7·85	5·95	0·37	17·94	2	15·70	11·90	0·74	35·88	77	61	18
30	7·84	5·51	0·16	18·02	2	15·68	11·02	0·32	36·04	76	17	−3
43	7·76	6·04	0·31	17·94	½	3·88	3·02	0·15	8·97	68	70	12
Sums	—	—	—	- -	16·5	116·77	71·62	3·13	295·81	—	—	—

No.	$X^*\times10^2$.	$WS^2\times10^4$.	$WSM\times10^4$.	$WSC\times10^4$.	$WM^2\times10^4$.	$WMC\times10^4$.	$WC^2\times10^4$.	$WX^2\times10^4$.	$WSX\times10^4$.	$WMX\times10^4$.	$WCX\times10^4$.
1	3	15488	17072	−3872	18818	−4268	968	18	−528	−582	132
8	−7	6728	6496	1276	6272	1232	242	98	812	784	154
9	−9	9248	5576	952	3362	574	98	162	1224	738	126
14	−23	1352	0	884	0	0	578	1058	1196	0	782
17	6	98	938	−168	8978	−1608	288	72	−84	−804	144
20	21	1156	−5678	−646	27889	3173	361	441	714	−3507	−399
26	15	5625	3675	−825	2401	−539	121	225	1125	735	−165
29	1	11858	24794	2772	51842	5796	648	2	154	322	36
30	9	11552	17784	−456	27378	−702	18	162	1368	2106	−54
43	1	2312	5780	408	14450	1020	72	1	34	85	6
Sums	—	65417	76437	325	161390	4678	3394	2239	6015	−123	762

$$\begin{Vmatrix} 6{\cdot}5417 & 7{\cdot}6437 & 0{\cdot}0325 \\ 7{\cdot}6437 & 16{\cdot}1390 & 3{\cdot}4678 \\ 0{\cdot}0325 & 3{\cdot}4678 & 0{\cdot}3394 \end{Vmatrix} \cdot \begin{Vmatrix} s_x \\ m_x \\ c_x \end{Vmatrix} = \begin{Vmatrix} 0{\cdot}6015 \\ -0{\cdot}0123 \\ 0{\cdot}0762 \end{Vmatrix}.$$

* Composition and b parameters as differences from the weighted means. Weighted means:

$\overline{\text{Si}} = 116{\cdot}77/16{\cdot}5 = 7{\cdot}08$, $\overline{\text{Mg}} = 71{\cdot}62/16{\cdot}5 = 4{\cdot}34$, $\overline{\text{Ca+Na+K}} = 3{\cdot}13/16{\cdot}5 = 0{\cdot}19$, $\bar{b} = 295{\cdot}81/16{\cdot}5 = 17{\cdot}93$.

before summation (table III), and the matrix equation is set up, the covariance matrix calculated, and the calculation of the coefficients completed as above. There is, however, one further complication; in the calculation of the standard deviations of the dependent variables we have $\acute{\sigma}_x^2 \sum W_i = \sum W_i X_i^2 - a_x \sum W_i X_i A_i - b_x \sum W_i X_i B_i - \ldots$ instead of $N\sigma_x^2 = \sum X_i^2 - a_x \sum X_i A_i - \ldots$. The formula for the standard deviations of the regression coefficients is not affected, and the number of degrees of freedom remains $N-n-1$.

The evaluation of the covariance matrix; a systematic procedure.

In the above discussion, the solution to the equation for the covariance matrix was shortly stated (p. 78) in the form: $\xi_{ij} = D_{ij}/D_0$, where ξ_{ij} is the element in the ith row and jth column of the covariance matrix, D_0 is the determinant of the matrix of sums of squares and products of the independent variables, and D_{ij} is the determinant derived by replacing the ith column of D_0 by the jth column of a unitary square matrix of order n, the number of independent variables. If $n = 2$ or 3, the formation and evaluation of these determinants presents no particular difficulties, but if there are many independent variables this process may become very tedious, and there are many opportunities for errors unless some systematic procedure is adopted. The procedure outlined below, based on successive pivotal reduction, has been found very effective in practice and includes adequate checks against arithmetical errors. It should be added that, since every step in the reduction of the determinants involves a subtraction of two numbers of the same order of magnitude, care must be taken to employ enough significant figures at each stage to ensure adequate accuracy in the final coefficients. With 2 or 3 significant figures in the observational data when expressed as differences from the means, 5 or 6 significant figures in the earlier stages of the reductions, and 4 or 5 in the last stages will not come amiss.

In the method of successive pivotal reduction for the evaluation of the determinants a determinant of order q is reduced to one of order $q-1$ by the following relation, which is then applied again to reduce the order to $q-2$, and so on:

$$\begin{vmatrix} a_1 & b_1 & c_1 & \ldots & q_1 \\ a_2 & b_2 & c_2 & \ldots & q_2 \\ a_3 & b_3 & c_3 & \ldots & q_3 \\ . & . & . & \ldots & . \\ . & . & . & \ldots & . \\ a_q & b_q & c_q & \ldots & q_q \end{vmatrix} = \begin{vmatrix} B_2 & C_2 & \ldots & Q_2 \\ B_3 & C_3 & \ldots & Q_3 \\ . & . & \ldots & . \\ . & . & \ldots & . \\ B_q & C_q & \ldots & Q_q \end{vmatrix} \times \frac{1}{a_1^{q-2}},$$

where $B_2 = a_1 b_2 - a_2 b_1$, $E_4 = a_1 e_4 - a_4 e_1$, and similarly throughout. This method has the great advantage that the quantities B_2, C_2, &c., of the 'denominator determinant' D_0 recur in all the derived determinants D_{ij} except the group D_{1j}. For if we replace the third column c_1, c_2,... of the left-hand determinant above by a new column s_1, s_2,..., then the determinant on the right will remain unchanged except for its second column C_2, C_3,..., which will be replaced by a new column S_2, S_3, where $S_2 = a_1 s_2 - a_2 s_1$, &c.; and if the kth column of the left-hand determinant is replaced by s_1, s_2,..., the $(k-1)$th column of the right-hand determinant will be replaced by S_2, S_3,.... The same, of course, applies to subsequent steps in the reductions. On this basis a systematic procedure for the reduction of the equation

for the covariance matrix is possible, in which D_0 and a proportion of the covariances are evaluated twice, so affording a check on the working.

Taking the covariance matrix for the refractive index data for anthophyllite as an example, we have:

$$\begin{Vmatrix} 4{\cdot}1476 & -0{\cdot}4356 & -1{\cdot}5162 & 4{\cdot}8148 & -0{\cdot}2226 \\ -0{\cdot}4356 & 0{\cdot}1265 & 0{\cdot}3478 & -0{\cdot}8315 & -0{\cdot}0175 \\ -1{\cdot}5162 & & 2{\cdot}0858 & -3{\cdot}6770 & -0{\cdot}0314 \\ 4{\cdot}8148 & & & 8{\cdot}2080 & -0{\cdot}1436 \\ -0{\cdot}2226 & & & & 0{\cdot}1888 \end{Vmatrix} \cdot \begin{Vmatrix} \xi_{ss} & \xi_{st} & \xi_{sf} & \xi_{sm} & \xi_{sc} \\ \xi_{ts} & \xi_{tt} & \xi_{tf} & \xi_{tm} & \xi_{tc} \\ \xi_{fs} & \xi_{ft} & \xi_{f} & \xi_{fm} & \xi_{fc} \\ \xi_{ms} & \xi_{mt} & \xi_{mf} & \xi_{mm} & \xi_{mc} \\ \xi_{cs} & \xi_{ct} & \xi_{cf} & \xi_{cm} & \xi_{cc} \end{Vmatrix} = \begin{Vmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{Vmatrix}.$$

The first pivotal reduction of D_0, the determinant of the left-hand matrix (the matrix of sums of squares and products of the independent variables), gives:

$$D_0 = \begin{vmatrix} 4{\cdot}1476\times 0{\cdot}1265 - \\ \quad -(0{\cdot}4356)^2 & & 4{\cdot}1476\times 0{\cdot}3478 - & & -4{\cdot}1476\times 0{\cdot}8315 + & & -4{\cdot}1476\times 0{\cdot}0175 - \\ & & \quad -0{\cdot}4356\times 1{\cdot}5162 & & \quad +0{\cdot}4356\times 4{\cdot}8148 & & \quad -0{\cdot}4356\times 0{\cdot}2226 \\ & & 4{\cdot}1476\times 2{\cdot}0858 - & & -4{\cdot}1476\times 3{\cdot}6770 + & & -4{\cdot}1476\times 0{\cdot}0314 - \\ & & \quad -(1{\cdot}5162)^2 & & \quad +1{\cdot}5162\times 4{\cdot}8148 & & \quad -1{\cdot}5162\times 0{\cdot}2226 \\ & & & & 4{\cdot}1476\times 8{\cdot}2080 - & & -4{\cdot}1476\times 0{\cdot}1436 + \\ & & & & \quad -(4{\cdot}8148)^2 & & \quad +4{\cdot}8148\times 0{\cdot}2226 \\ & & & & & & 4{\cdot}1476\times 0{\cdot}1888 - \\ & & & & & & \quad -(0{\cdot}2226)^2 \end{vmatrix} \times (4{\cdot}1476)^{-3}.$$

To find D_{ij}, we replace the ith column of the unreduced form of D_0 by the jth column of the right-hand unitary matrix; except when $i = 1$, the determinant so obtained will, after its first pivotal reduction, be identical with the first pivotal reduction of D_0, except in its $(i-1)$th column, which will be derived from the first column of D_0 and the jth column of the unitary matrix by the usual relation $B_2 = a_1b_2 - a_2b_1$.

We may therefore write the results of the first pivotal reduction thus:

$$D_0 = \begin{vmatrix} 0{\cdot}33492 & 0{\cdot}78208 & -1{\cdot}35140 & -0{\cdot}16955 \\ 0{\cdot}78208 & 6{\cdot}35220 & -7{\cdot}95052 & -0{\cdot}46774 \\ -1{\cdot}35140 & & 10{\cdot}86120 & 0{\cdot}47618 \\ -0{\cdot}16955 & & & 0{\cdot}73352 \end{vmatrix} \times (4{\cdot}1476)^{-3};$$

$$D_{(i+1),j} \propto \begin{Vmatrix} 0{\cdot}4356 & 4{\cdot}1476 & 0 & 0 & 0 \\ 1{\cdot}5162 & 0 & 4{\cdot}1476 & 0 & 0 \\ -4{\cdot}8148 & 0 & 0 & 4{\cdot}1476 & 0 \\ 0{\cdot}2226 & 0 & 0 & 0 & 4{\cdot}1476 \end{Vmatrix}$$

where the expression on the right is taken to indicate that $D_{(i+1),j}$ is derived by replacing the ith column of the left-hand expression for D_0 by the jth column of the right-hand matrix, each column of which is derived from the first column of D_0 and a column of the unitary matrix. Since i is necessarily between 1 and 5 inclusive and the left-hand expression for D_0 has only four columns, $i+1$ must lie between 2 and 5 inclusive, so that the expression on the right cannot define D_{ij} when $i = 1$.

We make further pivotal reductions in exactly the same manner, with the right-hand expression for $D_{(i+1),j}$ taking the place of the unitary matrix and combining with the first column of the reduced expression for D_0 to give the new expression for $D_{(i+2),j}$:

$$D_0 = \begin{vmatrix} 1{\cdot}51586 & -1{\cdot}60592 & -0{\cdot}02406 \\ -1{\cdot}60592 & 1{\cdot}81139 & -0{\cdot}06964 \\ -0{\cdot}02406 & & 0{\cdot}21693 \end{vmatrix} \times (4{\cdot}1476)^{-3} \times (0{\cdot}33492)^{-2};$$

$$D_{(i+2),j} \propto \begin{Vmatrix} 0{\cdot}16714 & -3{\cdot}24375 & 1{\cdot}38913 & 0 & 0 \\ -1{\cdot}02392 & 5{\cdot}60508 & 0 & 1{\cdot}38913 & 0 \\ 0{\cdot}14841 & 0{\cdot}70322 & 0 & 0 & 1{\cdot}38913 \end{Vmatrix}$$

$$D_0 = \begin{vmatrix} 0{\cdot}16684 & -0{\cdot}14420 \\ -0{\cdot}14420 & 0{\cdot}32826 \end{vmatrix} \times (4{\cdot}1476)^{-3} \times (0{\cdot}33492)^{2} \times (1{\cdot}51586)^{-1};$$

$$D_{(i+3),j} \propto \begin{Vmatrix} -1{\cdot}28371 & 3{\cdot}28732 & 2{\cdot}33083 & 2{\cdot}10573 & 0 \\ 0{\cdot}22899 & 0{\cdot}98793 & 0{\cdot}03342 & 0 & 2{\cdot}10573 \end{Vmatrix}.$$

Finally, we evaluate

$$D_0 = 0{\cdot}033970/(4{\cdot}1476) \times (0{\cdot}33492)^2 \times 1{\cdot}51586 = 2{\cdot}8000 \times 10^{-3},$$

while D_{4j} and D_{5j} ($j = 1, 2, 3, 4, 5$) are obtained by replacing the first and second columns respectively of the second-order determinant by the columns of the associated matrix of order 2×5.

$$\times 10^{-2}$$

$$D_{(i+3),j} = \begin{matrix} i+3 \\ 4 \\ 5 \end{matrix} \begin{Vmatrix} -3{\cdot}8837 & 12{\cdot}2155 & 7{\cdot}3711 & 6{\cdot}9122 & 3{\cdot}0365 \\ -1{\cdot}4691 & 6{\cdot}3886 & 3{\cdot}2727 & 3{\cdot}0365 & 3{\cdot}5131 \end{Vmatrix}$$

$$\qquad\qquad 1 \qquad 2 \qquad 3 \qquad 4 \qquad 5 \qquad = j.$$

And as $\xi_{ij} = D_{ij}/D_0$, we have:

$$\xi_{ij} = \begin{matrix} i \\ 4 \\ 5 \end{matrix} \begin{Vmatrix} 11{\cdot}434 & 35{\cdot}962 & 21{\cdot}699 & 20{\cdot}348 & 8{\cdot}939 \\ 4{\cdot}325 & 18{\cdot}808 & 9{\cdot}624 & 8{\cdot}939 & 10{\cdot}342 \end{Vmatrix}$$

$$\qquad\qquad 1 \qquad 2 \qquad 3 \qquad 4 \qquad 5 \qquad = j.$$

So far, we have only derived the last two rows of the covariance matrix. By inverting the initial matrix of sums of squares and products of the independent variables, bringing 0·1888 to the top left corner and 4·1476 to the bottom right, and reducing in the same way, we can derive two more rows (the first two). To derive the third row we must return to that stage of reduction at which D_0 is expressed as a third order determinant. If we replace the first column of this determinant by each of the five columns of the associated matrix of order 3×5, the five resulting third-order determinants will be the five D_{3j}; and since $5 - 2 = 1 + 2 = 3$, the same five determinants will result by treating the third-order stage of the reduction of the inverted matrix of sums of squares and products in the same way:

$$D_{31} = \begin{vmatrix} 0{\cdot}16714 & -1{\cdot}60592 & -0{\cdot}02406 \\ -1{\cdot}02392 & 1{\cdot}81139 & 0{\cdot}06964 \\ 0{\cdot}14841 & -0{\cdot}06964 & 0{\cdot}21693 \end{vmatrix} \times P^{-1}; \quad D_{32} = \begin{vmatrix} -3{\cdot}24375 & 1{\cdot}60592 & -0{\cdot}02406 \\ 5{\cdot}60508 & 1{\cdot}81139 & 0{\cdot}06964 \\ 0{\cdot}70322 & 0{\cdot}06964 & 0{\cdot}21693 \end{vmatrix} \times P^{-1};$$

$$D_{33} = 1{\cdot}38913 \times \begin{vmatrix} 1{\cdot}81139 & 0{\cdot}06964 \\ -0{\cdot}06964 & 0{\cdot}21693 \end{vmatrix} \times P^{-1}; \quad D_{34} = 1{\cdot}38913 \times \begin{vmatrix} 1{\cdot}60592 & 0{\cdot}02406 \\ -0{\cdot}06964 & 0{\cdot}21693 \end{vmatrix} \times P^{-1};$$

$$D_{35} = 1{\cdot}38913 \times \begin{vmatrix} 1{\cdot}60592 & 0{\cdot}02406 \\ 1{\cdot}81139 & 0{\cdot}06964 \end{vmatrix} \times P^{-1}; \text{ where } P = (4{\cdot}1476)^3 \times (0{\cdot}33492)^2.$$

These five determinants are then evaluated and divided by D_0 to give the five elements ξ_{3j}.

If our example had more independent variables, it would be necessary to go farther back to complete the covariance matrix: with seven variables, the first step, with direct and inverted matrices, would give rows 1, 2, 6, and 7; rows 3 and 5 could be derived from the third-order determinants, but row 4 must be derived from the fourth-order stage in the reduction of D_0 and its associated matrix.

In this procedure, every element of the covariance matrix is determined twice, except those in the principal diagonal, so affording an almost complete check. If the number of independent variables is high, it may be felt that so complete a check is superfluous, and it is quite a simple matter to omit some of the check calculations. Thus in our fifth-order example, after the first two and last two rows have been calculated, we may utilize the symmetry relation $\xi_{ij} = \xi_{ji}$ to write in the first two and last two columns, leaving only ξ_{33} to be determined.