MINERALOGICAL MAGAZINE, MARCH 1969, VOL. 37, NO. 285

The determination of multiple correlations between several variables, with especial reference to the correlation of physical properties and chemical composition

M. H. HEY

Department of Mineralogy British Museum (Natural History)

SUMMARY. If the true values of a number (n) of variables are exactly related by one or more (m) equations but all experimental values are liable to experimental error then, provided the errors of measurement can be assumed to be symmetrically distributed about a zero mean, provided the relations are linear, and provided the assessed probable errors of measurement of the several variables remain in constant ratio from one set of measurements to another, the best estimate of the relations derived from N sets of measurements will be given by the latent vectors corresponding to the m smallest latent roots of the n-square matrix whose terms are $\Sigma w_i \xi_{ij} \xi_{ik}$ (summed from $i = 1$ to $i = N$) for all j and k from 1 to n, where w_i is the weight of the ith set of measurements and ξ_{ij} is the standardized value of the ith measurement of the jth variable measured from its weighted mean.

It is also shown that the alleged inconsistency of the maximum-likelihood estimate of the variance in such a case is simply a confusion of the root-mean-square residual with the root-mean-square normal residual.

An Autocode computer programme has been written to carry out the necessary operations for correlating the physical properties and chemical properties of an isomorphous series of minerals using this procedure.

HARDLY any mineral is a pure chemical compound, and one of the principal tasks of mineralogy is to attempt to correlate the physical properties of minerals with their chemical composition. There may be a unique, one-to-one relation, or the physical properties of a mineral of a given structure and composition may vary in consequence of order–disorder phenomena, crystal imperfections, etc.; in what follows we shall ignore these complications and assume that for any given crystal structure a knowledge of the exact chemical composition will suffice to define the physical properties.

Since all the quantities involved are assumed to be linked by an exact relation, the 'true regressions' of Lindley (1947) are identical with his 'functional relation'; but the quantities are all liable to experimental error, and ordinary regression analysis is not applicable, because it assumes, in effect, that all error falls on the 'dependent' variables.

Two lines of approach to the problem are open, but each runs into certain difficulties. Considering first the maximum likelihood approach, and beginning with a simple two-dimensional example: suppose we have n paired observations $X_1, Y_1, X_2, Y_2, \ldots, X_n, Y_n$, and the distribution of error for each pair of observations

is measured by a pair of probability density functions $\phi_i(X_i-x)$, $\phi_i(Y_i-y)$, then the likelihood that the points all obey a relation $f(x, y) = 0$ is given by

$$\ln P = \text{const.} - \sum_{i=1}^{n} \ln \int \phi_i(X_i-x)\, dx + \ln \int \phi_i(Y_i-y)\, dy,$$

the integrals being taken over the whole range of values (x, y) satisfying $f(x, y) = 0$.

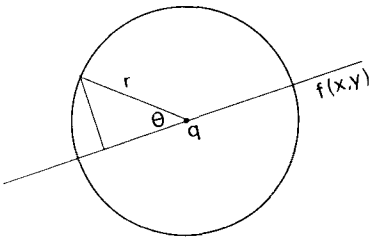

FIG 1.

If this relation is of known form but contains unknown coefficients, which we are seeking to adjust to the best possible fit to the observed data, it will be obvious that a knowledge of the form and parameters of the probability density functions is essential. If the functions approximate to normal Gaussian, with variances σ_{xi}^2, σ_{yi}^2, the likelihood P that any pair of true values x, y will give rise to observed values X, Y is given by $\ln P = K-(X-x)^2/\sigma_x^2-(Y-y)^2/\sigma_y^2$ where K is a constant. Let q be any point on the graph of the function $f(x, y) = 0$ taking σ_x, σ_y as units along the axes of x and y respectively (fig. 1); then all points on a circle with centre q are equally likely to be the representation of the observed values corresponding to the true point q, and all values of θ (fig. 1) are equally likely. Accordingly, if $f(x, y)$ is linear, the mean value of the square on the normal to $f(x, y)$ from any point at a distance r from q is

$$\bar{r}^2 = \frac{1}{\pi} \int_0^{\pi} r^2 \sin^2\theta\, d\theta = r^2/2;$$

and if $f(x, y)$ is non-linear, $r^2/2$ is a good approximation to $\bar{r}^2$ provided the radius of curvature of the graph in the neighbourhood of q is large compared with r.

Similarly in three dimensions

$$\bar{r}^2 = \int_0^{\pi/2} r^2 \sin^2\theta \cos\theta\, d\theta \Big/ \int_0^{\pi/2} \cos\theta\, d\theta = r^2/3,$$

and generally in n dimensions $\bar{r}^2 = r^2/n$.

If, then, we have N sets of observations of n related variables $(X_{ij}, i = 1, 2,..., N; j = 1, 2,..., n)$, the true values of which are known to be connected by a relation $f(x) \equiv f(a, b,..., x_{i1}, x_{i2}, x_{in}) = 0$, of known form but with unknown coefficients, we can apply the maximum likelihood method by minimizing $\Sigma R_i^2 = \Sigma\Sigma (X_{ij}-x'_{ij})^2/\sigma_{ij}^2$ (summed from $i = 1$ to N and $j = 1$ to n) provided the errors are distributed normally or quasi-normally, and provided $f(x)$ does not depart too far from linearity. But we must remember that in so doing we cannot locate the true values, x_{ij}, but only those values x'_{ij} for which the sum of the standardized normal squared residuals is minimized, and the expectation of this sum is only $1/n$ that of the sum of the true standardized squared residuals.[1] However, we are not limited to cases where the relation

[1] This point has commonly been overlooked, e.g. Lindley, 1947, p. 237; Kendall, 1956, p. 65; cf. Hey and Hey, 1960, p. 616.

between the variables is defined by a single equation. If it requires m equations (some of which may be known *a priori*) to define their relations completely, the relations can be represented by a surface of $n-m$ dimensions in n-space; and this can be defined by any m surfaces of $n-1$ dimensions (corresponding each to a single equation) that intersect in it. Therefore when we consider the expectation of the relation of the mean squared standardized normal residual to the mean squared standardized true residual (which will be 1) we may define the $(n-m)$-dimensional surface by m locally orthogonal primes[1] provided its local curvatures are not too great. As the mean squared normal residual to each of these primes will have an expected value of $1/n$, the total mean squared normal residual to the $(n-m)$-dimensional surface will be m/n.

The alternative approach is by a least-squares technique, which has the advantage of not requiring a normal or quasi-normal distribution of the errors of measurement, but merely a symmetrical one (Hamilton, 1964), but here we have a difficulty deciding what set of squared residuals to minimize. To take the simplest example, if x, y are linearly related, we could write the equation $y = ax+b$, or $x = cy+d$, or $px+qy = 1$, and the residuals $R = y-ax-b$, etc.; minimizing the sums of squared residuals leads to $\hat{a} = \Sigma XY/\Sigma X^2$, $1/\hat{c} = \Sigma Y^2/\Sigma XY$, and $\hat{p}/\hat{q} = \tan\theta$, where $\cot 2\theta = (\Sigma X^2-\Sigma Y^2)/2\Sigma XY$, respectively. As we suppose X and Y are both subject to errors of measurement, the first two solutions, which are in fact the two regressions, can be rejected because they offer unsymmetrical solutions to a symmetrical problem; the third is also unsatisfactory since it is affected by change of scale. To obtain a solution independent of scales we must eliminate all units and express our variables as pure numbers, and the only symmetrical and consistent solution appears to be that suggested by Kummell (1879); and this leads to the same result as the maximum likelihood technique.[2] Geometrically, Kummell's least squares solution can be expressed as: for each set of observations separately, let the data be standardized by taking their standard deviations of error[3] as units, and graphed in n-space; note the variation in the length R_i of the normal from the point to the surface $f(x) = 0$ with the coefficients a, b,..., and adjust these coefficients so that ΣR_i^2 is a minimum.

It will be obvious that although Kummell only considered a single equation connecting the n variables, corresponding to a surface of $n-1$ dimensions in n-space, his procedure is equally applicable where the relation can be represented by a surface of fewer, say $n-m$, dimensions, to define which m equations of relation will be required.

[1] In n dimensions, a surface defined by m linear equations is called an $(n-m)$-flat; if $n-m = 1$ we have a line, if $n-m = 2$ a plane, while the surface of $n-1$ dimensions is a prime (the term hyperplane is ambiguous when $n > 4$).

[2] Kummell's equation 4 (loc. cit., p. 98) is $\Sigma\Sigma(X_{ij}-x'_{ij})^2/\sigma_{ij}^2 =$ a minimum, in the symbols used here.

[3] We assume that the standard errors of all the measurements are known, or at least can be estimated, and are symmetrically distributed about a zero mean. It is sometimes stated (Madansky, 1959, p. 179) that if all the error variances are known we have an over-identified situation, but this is another consequence of the failure to distinguish between residuals and normal residuals; in Madansky's example of two variables whose error variances σ_y^2, σ_x^2 are known and do not vary from one pair of measurements to another, knowledge of the absolute values of the variances instead of merely their ratio simply enables us to assess the average distance between the true points and the feet of the standardized normals from the experimental points to the line.

In general, the minimum of ΣR_i^2 can only be found by successive approximations;[1] for a direct solution $f(x)$ must be linear and the ratios $\sigma_{i1}^2:\sigma_{i2}^2:\ldots:\sigma_{in}^2$ must not vary with i. Under these conditions the solution for the bivariate case was given by Kummell (1879);[2] Pearson (1901) derived a solution for the multivariate relation, on the implicit assumption that all the σ_{ij}^2 are equal, but as it involves finding the roots of a determinantal equation it has been little used; a solution on the same assumption for the case where there are several linear relations between the variables is implicit in the work of Brown and Fereday (1958).

The extension of Kummell's solution to multiple linear relations between several variables, subject to the condition[3] that the ratio of error variances is constant from one set of experimental data to another, is straightforward. We can write $\sigma_{ij}^2 = \sigma_j^2/w_i$, where w_i are the weights of the N sets of observations and σ_j^2 the error variances of the n variables of a set of unit weight. It will be convenient to adopt the σ_j as units along the several axes of n-space, and also to transfer the origin to the weighted mean of the observations, when we write $\xi_{ij} = (X_{ij}-\bar{X}_j)/\sigma_j$, $\xi'_{ij} = (x'_{ij}-\bar{X}_j)/\sigma_j$, where $\bar{X}_j = \Sigma w_i X_{ij}/\Sigma w_i$ (summations from $i = 1$ to $i = N$).

The set of linear relations between the variables, which we take to be m in number, define an $(n-m)$-flat in n-space, and since this is equally well defined by any m independent primes (i.e. $(n-1)$-flats) that intersect in it, we can with loss of generality postulate m mutually orthogonal equations with normalized coefficients, $\Sigma\alpha_{hj}\xi'_j = c_h$ (summation from $j = 1$ to N; $h = 1, 2,\ldots, m$), with the condition $\Sigma\alpha_{hj}\alpha_{kj} = 0$ if $h \neq k$, or 1 if $h = k$.

With these co-ordinates and postulates, it can readily be shown (cf. Appendix, p. 88) that for a minimum of ΣR^2 the desired $(n-m)$-flat passes through the weighted mean of the observations, so that all the c_h are zero, and is defined by the latent vectors corresponding to the m smallest latent roots of the n-square matrix whose terms are $\Sigma w_i \xi_{ij}\xi_{ik}$ (summed from $i = 1$ to N) for all j and k from 1 to n. The resulting equations will, of course, be in terms of ξ', and must be destandardized and the origin restored to its original position; they will usually be in all n variables, which is inconvenient, but can be remedied since we can properly combine them to eliminate up to $m-1$ variables and so derive equations for any variable in terms of any $n-m$ of the others (such recombination would, of course, be out of order with a set of regression equations).

[1] See, e.g. Deming, 1943, chap. 4.

[2] Madansky (1959, pp. 200–2) points out that this solution, also cited by Deming (1943, p. 184), by Lindley (1947, p. 236), and by Kendall (1956, p. 64), is not quite correctly stated, and in consequence only yields a correct result if dy/dx is positive. A correct solution is given by

$$dy/dx = \{P+\sqrt{(p^2+k^2Q^2)}\}/Q,$$

where $P = \Sigma(Y-\bar{Y})^2-k^2\Sigma(X-\bar{X})^2$, $Q = 2\Sigma(X-\bar{X})(Y-\bar{Y})$, and $k = \sigma_y/\sigma_x$.

[3] This condition is not as restrictive as it might at first appear. As Kummell (1879) and Hamilton (1964) point out, it will normally be preferable to assess the probable standard deviation of error from experience of the measuring technique rather than to rely on internal estimates derived from a small number of repetitions. And if the sets of data are of varying precision (as will usually be the case with data culled from the literature) the loss of information involved in weighting each set of observations according to an assessment of its least accurate member is not a serious penalty to pay for a direct method of computation.

While it is a very simple matter to set up the matrix $\Sigma w_i \xi_{ij} \xi_{ik}$, computation of its smallest latent roots and the corresponding latent vectors by a desk calculator is tedious as soon as the number of variables exceeds four, and prohibitively so as soon as there is more than one equation of relation. Fortunately, programmes are available to handle the task on all but the smallest computers.

Applications in mineralogy. The above discussion has been general, applicable to any collection of related variables; in most mineralogical problems the data are subject to certain known relations in addition to the unknown ones we are seeking, the most obvious being that the sum of the chemical data is 100% within the experimental error. It might appear desirable to utilize this knowledge to eliminate one or more variables at the start, but this is not so; in fact we shall obtain slightly different results according to which variable or variables we eliminate; completely consistent results only result from completely symmetrical treatment. Fortunately, most computer programmes for latent roots and vectors are competent to handle a singular matrix, so the known relation is reproduced automatically as a latent vector corresponding to a zero latent root.

A computer programme has been written in Autocode for use on the Atlas computer,[1] by which linear relations between the chemical composition and the physical properties of a mineral of variable composition can be derived; and if there are p independent chemical components, the relations between any $p+1$ of the variables, chemical and physical, can be displayed.

The programme also computes a number of other useful statistics: since marked correlation, positive or negative, between any pair of variables[2] results in very unstable coefficients in any equation in which both appear, the matrix of partial correlation coefficients is computed. Since the latent vectors are orthogonal, the sum of the latent roots corresponding to the m vectors computed is ΣR^2, and if the standard errors of measurement were correctly estimated and there are no unrecognized sources of variation, we should have the root-mean-square standardized residual

$$\bar{R} = \sqrt{(n\Sigma R^2/m\Sigma w)} \approx 1;$$

in practice $\bar{R}$ will usually be a good deal larger, because of such factors as non-linearity due to ordering, neglected chemical components, etc., but it does provide a useful check and is therefore computed. The simplest check on linearity is to plot the residuals of the several sets of data for the several equations against one or more of the variables, and the programme is arranged to provide these if desired.

It should be noted that if an equation is derived for one of the physical variables in terms of the chemical composition at least one chemical component will necessarily be eliminated, and the coefficients of the equation obtained must therefore be interpreted with caution (see, e.g. Louisnathan and Smith, 1968; Smith, Stephenson, Howie, and Hey, 1969). In as simple a system as an iron-free hornblende with Na, Ca, Mg,

[1] This programme in typescript is available from the author, with notes, but excluding the routine for latent roots and vectors. (This latter is London Computer Centre Library Programme −519.)

[2] Which may be intrinsic, as when all the refractive indices are similarly affected by an ionic substitution, or may be an accident of the selection of specimens for study.

Al^{vi}, Al^{iv}, and Si as chemical variables (calculated to 24(O, OH) and assuming OH = 2), two of the variables must be eliminated to meet the relations Σ(valencies)=46 and $Si+Al^{iv} = 8$, and it may be convenient to choose Al^{iv} and Al^{vi} for elimination; then since there is no other element of the same valency, only such substitutions as $NaAl \rightleftharpoons Si$, $Na_2 \rightleftharpoons Ca$, or $NaSi \rightleftharpoons MgAl$ are possible, and no direct significance can be attached to individual coefficients.

Finally, we may note that density data cannot properly be included in correlation calculations, because the error distribution in density determinations is notoriously skew; and that because the birefringences of a mineral are often known with higher precision than the difference of the reported refractive indices, it is usually preferable to use one refractive index and the birefringences in correlation studies. Optic axial angles, being correlated, though non-linearly, with the birefringences, can be used to adjust the observed birefringences, and it is hoped to write a computer programme for this purpose.

REFERENCES

BROWN (R. L.) and FEREDAY (F.), 1958. *Biometrika*, **45**, 1.
DEMING (W. E.), 1943. *Statistical Adjustment of Data*. New York (Wiley).
HAMILTON (W. C.), 1964. *Statistics in Physical Science*. New York (Ronald Press).
HEY (E. N.) and HEY (M. H.), 1960. *Biometrics*, **16**, 606.
KENDALL (M. G.), 1956. *A Course in Multivariate Analysis*. London (Griffin).
KUMMELL (C. H.), 1879. *The Analyst (Des Moines)*, **6**, 97.
LINDLEY (D. V.), 1947. *Journ. Roy. Statist. Soc.*, suppl. **9**, 218.
LOUISNATHAN (S. J.) and SMITH (J. V.), 1968. *Min. Mag.* **36**, 1123.
MADANSKY (A.), 1959. *Journ. Amer. Statist. Ass.* **54**, 173.
PEARSON (K.), 1901. *Phil. Mag.*, ser. 6, **2**, 559.
SMITH (J. V.), STEPHENSON (D. A), HOWIE (R. H.), and HEY (M. H.), 1969. *Min. Mag.* **37**, 90.

APPENDIX

We have seen above that in the system of co-ordinates we have adopted both maximum likelihood and least-squares techniques require us to minimize the sum of weighted squared normal residuals to the correlation $(n-m)$-flat to obtain the best linear fit. If the flat is defined by m linear, mutually orthogonal equations $f_h(\xi') = \sum_{j=1}^{n} \alpha_{hj}\, \xi'_j + c_h$ $(h = 1, 2, \ldots, m)$, this sum is

$$\Sigma R^2 = \sum_{h=1}^{m} \sum_{i=1}^{N} w_i \Big(\sum_{j=1}^{n} \alpha_{hj}\, \xi_{ij} + c_h \Big)^2 .$$

For a minimum we must have

$$0 = \partial \Sigma R^2 / \partial c_h = \sum_{i=1}^{N} \sum_{j=1}^{m} w_i\, \alpha_{hj}\, \xi_{ij} + c_h$$

for all h; as the observed values ξ_{ij} are measured from their weighted means, all $\sum_{i=1}^{N} w_i\, \xi_{ij}$ are zero, and hence all c_h are zero.

We must also have for all h and j, remembering the orthogonality conditions $\Sigma \alpha_{hj}\, \alpha_{kj} = 0$ or 1:

$$0 = \tfrac{1}{2} \partial \Sigma R^2 / \partial \alpha_{hj} - \mu_h\, \alpha_{hj} = \sum_{k=1}^{n} \sum_{i=1}^{N} w_i\, \xi_{ij}\, \xi_{ik}\, \alpha_{hj} - \mu_h\, \alpha_{hj}$$

These mn equations can be written as a matrix equation $\mathbf{W \cdot A} = \mathbf{A \cdot M}$, where $\mathbf{W}$ is the n-square matrix whose k, jth term is $\Sigma w_i\, \xi_{ij}\, \xi_{ik}$ (summed from $i = 1$ to N); $\mathbf{A}$ is the matrix

of n rows and m columns whose terms are α_{kj} and has been postulated to be orthogonal by columns and $\mathbf{M}$ is the m-square diagonal matrix whose non-zero terms are μ_h.

Now as $\mathbf{W}$ is symmetrical, we have $\mathbf{W} = \mathbf{V}\cdot\mathbf{\Lambda}\cdot\tilde{\mathbf{V}}$, where $\mathbf{\Lambda}$ is the diagonal matrix of latent roots of $\mathbf{W}$, and $\tilde{\mathbf{V}}$ its (orthogonal) matrix of row vectors; hence $\mathbf{V}\cdot\mathbf{\Lambda}\cdot\tilde{\mathbf{V}}\cdot\mathbf{A} = \mathbf{A}\cdot\mathbf{M}$, and premultiplying by $\tilde{\mathbf{V}}$, $\mathbf{\Lambda}\cdot\tilde{\mathbf{V}}\cdot\mathbf{A} = \tilde{\mathbf{V}}\cdot\mathbf{A}\cdot\mathbf{M}$. Let λ_p, μ_q be the pth and qth diagonal terms of $\mathbf{\Lambda}$ and $\mathbf{M}$ respectively and v_{pq} the term in the pth row and qth column of $\tilde{\mathbf{V}}\cdot\mathbf{A}$; then $\lambda_p v_{pq} = \mu_q v_{pq}$ for all $p = 1, 2,..., n$ and $q = 1, 2,..., m$; the only solutions to this are either that all terms of $\mathbf{\Lambda}$ are equal (which will not be true), and also all terms of $\mathbf{M}$ are equal, or that $\tilde{\mathbf{V}}\cdot\mathbf{A} = \mathbf{J}$, where $\mathbf{J}$ is one of the permutations of the columns of the $n \times m$ matrix $[\mathbf{D}\ \mathbf{0}]$, where $\mathbf{D}$ is diagonal. Further, since both $\mathbf{V}$ and $\mathbf{A}$ are orthogonal, the only solutions (giving stationary values of ΣR^2) are when $\mathbf{A}$ consists of any m columns of the latent vector matrix $\mathbf{V}$; and direct expansion shows that the stationary values of ΣR^2 will be equal to the sum of the corresponding latent roots. Hence the minimum minimorum will be given by the latent vectors of $\mathbf{W}$ corresponding to its m smallest latent roots.

[*Manuscript received 3 June 1968; revised 18 October 1968*]