Acta Cryst. (1959). **12**, 981

Regression Formulae and the Joint Distribution of Structure Factors

By Philip A. Vaughan

Rutgers, The State University of New Jersey, New Brunswick, New Jersey, U.S.A.

(*Received* 2 *October* 1957 *and in revised form* 10 *November* 1958)

The joint distribution (frequency) function of a set of structure factors can be obtained as an expansion in terms of a general set of orthogonal polynomials. The series given by Hauptman & Karle and also by Bertaut is a particular example of such an expansion. The question is considered from the standpoint of regression formulae and it is shown that the (terminated) sign-determining series of Hauptman & Karle does not represent a least-squares regression formula. A method of obtaining improved regression formulae is considered and illustrated in the case of space group $P\bar{1}$. A numerical example is presented for a synthetic structure.

Hauptman & Karle (1953) have proposed a method of determining the signs of structure factors which is based on the theory of probabilities. Specifically, they assume that the atomic coordinates are random variables which are uniformly distributed between 0 and 1 (for general positions). They then proceed to derive formulae for the joint distribution of a group of structure factors and show how this joint distribution forms the basis of a method of sign determination. Bertaut (1955*a*, *b*) has also derived the joint distribution by a method which is similar to that of Hauptman & Karle and which can be expected to give identical results provided the same terms are kept in the terminated series.

In this discussion a general formula for the joint distribution of normalized structure factors is given and it is pointed out that Hauptman & Karle's result is a special case. The question as to the best statistical means of phase determination will then be considered from the standpoint of regression formulae. It turns out that Hauptman & Karle's results for the joint distribution do not in general yield the best (least-squares) regression formulae. This is because of the lack of orthogonality of the individual terms in their expansion. A discussion of regression formulae which can be used to estimate $E_{\mathbf{H}}(h, k, l$ even) in space group $P\bar{1}$ is given.

One method of investigating this problem is by the method of orthogonal polynomials (see Cramer, 1946, for a discussion of one-dimensional orthogonal polynomials). A form of expansion of a function $g(x_1, x_2, \ldots, x_n)$ of n variables in terms of one-dimensional orthogonal polynomials is given by

$$g(x_1, \ldots, x_n) = \left[\prod_{j=1}^{n} f_j(x_j)\right] \sum_{\substack{k_1, \ldots, k_n \\ =0}}^{\infty} C^{(\nu)}_{k_1, \ldots, k_n} \prod_{j=1}^{n} p_{j,k_j}(x_j) , \tag{1}$$

in which the subscripts $k_1, \ldots, k_n$ give the orders of the polynomials $p_{j,k_j}(x_j)$, $\nu =$ total order $= k_1 + \ldots + k_n$, and

$$C^{(\nu)}_{k_1, \ldots, k_n} = \int_{-\infty}^{\infty} \cdots \int g(x_1, \ldots, x_n) \prod_{j=1}^{n} p_{j,k_j}(x_j) dx_j . \tag{2}$$

The last equation follows from the orthonormality of each set of polynomials $p_{j,k_j}(x_j)$ with respect to the corresponding weight functions $f_j(x_j)$; that is

$$\int_{-\infty}^{\infty} p_{j,k_j}(x_j) p_{j,l_j}(x_j) f_j(x_j) dx_j = \delta_{k_j l_j} \quad (j = 1, \ldots, n) . \tag{3}$$

The conditions on the $f_j(x_j)$ which make it possible to determine the set of polynomials $p_{j,k_j}(x_j)$ are discussed by Cramer (1946) and Szegö (1939). The interval of integration need not be infinite, as in (2) and (3). It is also clear that orthogonal polynomials may be found which are functions of all n variables. In this case, the n-dimensional weight function can not generally be factored, as in (1).

Equations (1) and (2) can be used to obtain any number of formal expansions of the joint distribution function. The function $g(x_1, \ldots, x_n)$ is replaced by the required joint distribution (frequency) function

$$P(A_{\mathbf{H}_1}, \ldots, A_{\mathbf{H}_n})$$

for the normalized structure factors $E_{\mathbf{H}_1}, \ldots, E_{\mathbf{H}_n}$. Then (2) can be written

$$C^{(\nu)}_{k_1, \ldots, k_n} = \overline{\prod_{j=1}^{n} p_{j,k_j}(E_{\mathbf{H}_j})} , \tag{4}$$

in which the bar indicates averaging over all atomic coordinates. Following Hauptman & Karle (1953), we assume that the atomic coordinates are uniformly distributed in the interval 0 to 1. The functions $f_j(A_{\mathbf{H}_j})$ can be any functions for which orthogonal polynomials can be found.

In particular, if

$$f_j(A_{\mathbf{H}_j}) = \frac{1}{(2\pi)^{\frac{1}{2}}} \exp\left[-\tfrac{1}{2} A^2_{\mathbf{H}_j}\right] ,$$

the orthogonal polynomials are the orthonormal Hermite polynomials,

$$\left(\frac{1}{k_j!}\right)^{\frac{1}{2}} H_{k_j}(A_{\mathbf{H}_j}) .$$

We then obtain

$$P(A_{\mathbf{H}_1}, \ldots, A_{\mathbf{H}_n}) = (2\pi)^{-\frac{1}{2}n} \prod_{j=1}^{n} \exp\left[-\tfrac{1}{2}A_{\mathbf{H}_j}^2\right] \sum_{\substack{k_1,\ldots,k_n \\ =0}}^{\infty} \overline{\left[\prod_{j=1}^{n} H_{k_j}(E_{\mathbf{H}_j})\right]} \prod_{j=1}^{n} \frac{1}{k!} H_{k_j}(A_{\mathbf{H}_j}) , \quad (5)$$

which is identical with Bertaut's (1955*b*) equation (III–1). Many equations similar to (5) can be obtained simply by choosing different functions $f_j(A_{\mathbf{H}_j})$. The question arises, which is best? One possible answer can be found by considering regression formulae.

Regression formulae

Regression curves and formulae are discussed in standard treatises on statistics (Cramer, 1946). If we have a group of random variables $y_1, \ldots, y_n$, then the regression surface for y_1 is the locus of the mean value of y_1 subject to the conditions $y_2 = x_2, \ldots, y_n = x_n$; that is, the equation

$$x_1 = m_1(x_2, \ldots, x_n) = \int_{-\infty}^{\infty} x_1' P(x_1', x_2, \ldots, x_n) dx_1' , \quad (6)$$

in which $P(x_1, \ldots, x_n)$ is the joint probability distribution function for $y_1, \ldots, y_n$, is the required regression formula. This gives the least-squares estimate of y_1, subject to the conditions $y_2 = x_2, \ldots, y_n = x_n$. Let us find the regression formula for the case in which $P(x_1, \ldots, x_n)$ is given by an expansion in orthogonal polynomials, as in equation (1):

$$m_1(x_2, \ldots, x_n) = \overline{x_1 p_{1,1}(x_1)} \prod_{j=2}^{n} f_j(x_j) \sum_{\substack{k_2,\ldots,k_n \\ =0}}^{\infty} C^{(\nu)}_{1,k_2,\ldots,k_n} \prod_{j=2}^{n} p_{j,k_j}(x_j) , \quad (7)$$

in which

$$C^{(\nu)}_{1,k_2,\ldots,k_n} = \overline{p_{1,1}(x_1) \prod_{j=2}^{n} p_{j,k_j}(x_j)} .$$

This equation follows from the fact that $\overline{x_1 p_{1,k_1}(x_1)} = 0$ for $k \neq 1$ if x_1 has a mean value of zero (which is assumed). In the crystallographic application, x_1 can be a normalized structure factor or an origin-invariant combination of structure factors reduced to a mean value of zero. The $x_j (j = 2, \ldots, n)$ can be either $E_{\mathbf{H}_j}$ or $(E^2_{\mathbf{H}_j} - 1)$. If (7) is used for sign determination, the sign of $m_1(x_2, \ldots, x_n)$ will depend only on the sum of polynomials if the functions $f_j(x_j)$ are symmetric. The Gaussian functions which appear in Hauptman & Karle's (1953) treatment are, of course, symmetric.

The difficulty of (7) is that, although it will provide the best (least squares) estimate of y_1 if $x_j (j = 2, \ldots, n)$ are given, the speed of convergence of the series of polynomials will be dependent on the particular weight functions $f_j(x_j)$ which are chosen. It has not been demonstrated that Gaussian functions are particularly advantageous.

Let us take another approach and assume a regression formula to estimate x, as a function of a set of observables $x_j (j = 1, \ldots, n)$, of the type

$$x \approx \sum_{\substack{k_1,\ldots,k_n \\ =0}}^{\nu=m} C^{(\nu)}_{k_1,\ldots,k_n} P^{(\nu)}_{k_1,\ldots,k_n}(x_1, \ldots, x_n) , \quad (8)$$

in which the $P^{(\nu)}_{k_1,\ldots,k_n}$ are polynomials of total order $\nu = k_1 + \ldots + k_n$, and $C^{(\nu)}_{k_1,\ldots,k_n}$ are constants to be determined. We assume that all possible, linearly independent, polynomials of total order equal to or less than m are included in the summation. Thus, there will be $(\nu+n-1)!/\nu!(n-1)!$ linearly independent polynomials of total order ν and a total of $(n+m)!/n!m!$ polynomials on the right side of (8).

We will assume best constants $C^{(\nu)}_{k_1,\ldots,k_n}$ are given by the least-squares principle; that is, we minimize the mean square difference between the left and right sides of (8). Straightforward minimization gives the set of equations

$$\sum_{\substack{k'_1,\ldots,k'_n \\ =0}}^{\nu'=m} C^{(\nu')}_{k'_1,\ldots,k'_n} \overline{P^{(\nu)}_{k_1,\ldots,k_n}(x_1, \ldots, x_n) P^{(\nu')}_{k'_1,\ldots,k'_n}(x_1, \ldots, x_n)} = \overline{x P^{(\nu)}_{k_1,\ldots,k_n}(x_1, \ldots, x_n)} , \quad (9)$$

in which the bar indicates mean value. Ordinarily, one would have to solve this set of simultaneous linear equations in order to determine the constants $C^{(\nu)}_{k_1,\ldots,k_n}$. However, if the polynomials are orthonormal with respect to a weight function which is the joint probability distribution function of the variables $x_j (j = 1, \ldots, n)$, then (9) becomes

$$C^{(\nu)}_{k_1,\ldots,k_n} = \overline{x P^{(\nu)}_{k_1,\ldots,k_n}(x_1, \ldots, x_n)} , \quad (10)$$

since

$$\overline{P^{(\nu)}_{k_1,\ldots,k_n}(x_1, \ldots, x_n) P^{(\nu')}_{k'_1,\ldots,k'_n}(x_1, \ldots, x_n)} = \delta_{k_1 k'_1} \ldots \delta_{k_n k'_n}$$

in this case. The regression formula (8) then becomes

$$x \approx \sum_{\substack{k_1,\ldots,k_n \\ =0}}^{\nu=m} \overline{x P^{(\nu)}_{k_1,\ldots,k_n}(x_1, \ldots, x_n)} P^{(\nu)}_{k_1,\ldots,k_n}(x_1, \ldots, x_n) . \quad (11)$$

The similarity between (11) and the summation on the right side of (7) is obvious since $p_{1,1}(x_1)$ is equal to a constant times x_1 (since the mean of x_1 is zero). Equation (11) is more general since the functions $P^{(\nu)}_{k_1,\ldots,k_n}(x_1, \ldots, x_n)$ cannot necessarily be factored, as in (7). The situation is now clear. If (7) is used *with a finite number of terms*, it will not generally prove to be a satisfactory regression formula since the polynomials $p_{j,k_j}(x_j)$ (or their products) are not orthogonal with respect to the joint probability distribution function. In particular, the Hermite polynomials and

their products are not orthogonal with respect to the joint probability distribution function of a set of normalized structure factors. Therefore, we can expect that it is possible to develop more powerful equations than those presented by Hauptman & Karle for the purpose of phase determination by starting with equation (8) and using the least-squares principle to determine the constants.

We must first discuss briefly what to use for the quantities x_j in crystallographic applications. At the beginning of a structure determination, no signs have been determined and we use the observed values of $|(E_{\mathbf{H}_j}|^2-1)$, which fulfill the requirement of having mean values of zero (we will assume, for convenience, that there are no atoms in special positions; appropriate modifications can be made in case this condition does not exist). In the event that some phases have been determined (by statistical means, by inequalities, or by structural requirements), the corresponding values of $E_{\mathbf{H}_j}$ could presumably be used with advantage. We shall not, however, consider this more complicated situation. We shall assume that the quantities $(|E_{\mathbf{H}_j}|^2-1)$ are to be used to estimate some origin-invariant product of normalized structure factors, reduced to mean value of zero, which we shall call G. For example, we might have

$$G=E_{\mathbf{H}_1}E_{\mathbf{H}_2}E_{-\mathbf{H}_1-\mathbf{H}_2}-\overline{E_{\mathbf{H}_1}E_{\mathbf{H}_2}E_{-\mathbf{H}_1-\mathbf{H}_2}}\,.$$

(For non-centrosymmetric structures, we may have to consider $G+G^*$, in which G^* refers to the inverted structure).

One possible method of procedure would be to construct a set of orthonormal polynomials from the set $(|E_{\mathbf{H}_j}|^2-1)$. This can always be done, for example, by the Schmidt method (Margenau & Murphy, 1943), since all of the moments of these quantities can be computed. Nevertheless, this is a difficult procedure and we will begin with the equivalent procedure of writing (8) in the form of a power series; that is put

$$P^{(\nu)}_{k_1,\ldots,k_n}(x_1,\ \ldots,\ x_n)=x_1^{k_1}\ldots x_n^{k_n}$$

and (8) becomes, on writing G for x and $(|E_{\mathbf{H}_j}|^2-1)$ for x_j:

$$G\approx\sum_{\substack{k_1,\ldots,k_n\\=0}}^{\nu=m}C^{(\nu)}_{k_1,\ldots,k_n}(|E_{\mathbf{H}_1}|^2-1)^{k_1}\ldots(|E_{\mathbf{H}_n}|^2-1)^{k_n}\,. \quad (12)$$

The constants $C^{(\nu)}_{k_1,\ldots,k_n}$ can be determined by solving equations (9). This also is a rather impractical procedure and in order to obtain useful results without excessive labor, we will make the following approximations:

(i) All terms will be omitted from (12) for which the covariances

$$\overline{G(|E_{\mathbf{H}_1}|^2-1)^{k_1}\ldots(|E_{\mathbf{H}_n}|^2-1)^{k_n}}$$

are zero.

(ii) All terms which have a common value for the above covariance have the same coefficient in (12).

These approximations result in a great reduction in the number of coefficients to be determined although there is some loss in accuracy of the final results.

Regression formulae for $E_{\mathbf{H}}$ in space group $P\bar{1}$

As an example, and to illustrate the suggested regression calculation, all terms ($\nu\leq 3$) which give non-zero covariance with $E_{\mathbf{H}}(h,k,l$ all even) in space group $P\bar{1}$ are listed in Table 1, along with their covariances (with $E_{\mathbf{H}}$) and variances for the case of N equal atoms per unit cell. The latter quantities (and similar quantities which will be used later) have also been obtained for the case of unequal point atoms; they are polynomials of the quantities

$$S_n=\sum_1^N g_j^n \qquad (E_{\mathbf{H}}=\sum_1^N g_j\exp\,[2\pi i\mathbf{H}\cdot\mathbf{r}_j])\,,$$

and in some cases are quite lengthy. The regression formula using these terms would be

$$E_{\mathbf{H}}\approx\sum_j C_jT^{(j)}_{\mathbf{H}}\,, \quad (13)$$

in which $T^{(j)}_{\mathbf{H}}$ are the quantities listed in the second column of Table 1. The coefficients C_j would be found by solving the equations

$$\sum_j C_j\overline{T^{(j)}_{\mathbf{H}}T^{(k)}_{\mathbf{H}}}=\overline{E_{\mathbf{H}}T^{(k)}_{\mathbf{H}}}. \quad (14)$$

This is still a rather formidable task but it is within the range of modern electronic computers.

Some approximations have been made in evaluating the variances and covariances listed in Table 1. It will be noticed that some of the terms listed contain others as special members. For example, term 3(i) is the member of 3(e) for which $\mathbf{K}=\frac{1}{2}\mathbf{H}$. For theoretical purposes, such members can be considered as omitted from the terms which are sums. Furthermore, the member for which $\mathbf{K}=0$ is not included in these sums. The variances of the sum terms

$$T_{\mathbf{H}}=\sum_{\mathbf{K}}t_{\mathbf{H},\mathbf{K}}$$

have been evaluated from the relation

$$\overline{T^2_{\mathbf{H}}}=\overline{(\sum_{\mathbf{K}}t_{\mathbf{H},\mathbf{K}})^2}=\sum_{\mathbf{K}}\overline{t^2_{\mathbf{H},\mathbf{K}}}+\sum'_{\mathbf{K}_1,\mathbf{K}_2}\overline{t_{\mathbf{H},\mathbf{K}_1}t_{\mathbf{H},\mathbf{K}_2}}$$
$$\approx n_K\overline{t^2_{\mathbf{H},\mathbf{K}}}+n_K(n_K-1)\overline{t_{\mathbf{H},\mathbf{K}_1}t_{\mathbf{H},\mathbf{K}_2}}\,. \quad (15)$$

The average $\overline{t_{\mathbf{H},\mathbf{K}_1}t_{\mathbf{H},\mathbf{K}_2}}$ has been evaluated for general values of $\mathbf{K}_1$ and $\mathbf{K}_2$. For some particular values of $\mathbf{K}_1$ and $\mathbf{K}_2$, for example if $\mathbf{K}_2=\frac{1}{2}\mathbf{K}_1$, this quantity is sometimes different from that given in Table 1. The resulting additional contributions to $T^2_{\mathbf{H}}$ have been neglected. It will be noted that the second term in (15) is, in general, not small compared to the first; this illustrates the non-orthogonality of the members of these sums.

Table 1. *Terms (order ≤ 3) in the regression formula for the estimation of $E_{\mathbf{H}}$ in space group $P\bar{1}$ (a)*

Designation	Term ($T_{\mathbf{H}}$)	Covariance $\overline{E_{\mathbf{H}} T_{\mathbf{H}}}^{(b)}$	Variance $\overline{T_{\mathbf{H}}^2}^{(b)}$
1(*a*)	$E^2_{\mathbf{H}/2}-1$	$1/N^{1/2}$	$2-(3/N)$
2(*a*)	$(E^2_{\mathbf{H}/2}-1)^2$	$[1/N^{1/2}][4-(8/N)]$	$60-(468/N)+(1275/N^2)-(1155/N^3)$
2(*b*)	$(E^2_{\mathbf{H}}-1)(E^2_{\mathbf{H}/2}-1)$	$[1/N^{1/2}][2-(3/N)]$	$4+(12/N)-(99/N^2)+(120/N^3)$
2(*c*)	$\sum_{\mathbf{K}}(E^2_{(\mathbf{H}/2)-\mathbf{K}}-1)(E^2_{\mathbf{K}}-1)$	$n_K/N^{3/2}$	$[n_K(n_K-1)/N^2][8-(15/N)]+n_K[4-(12/N)]$
3(*a*)	$\sum_{\mathbf{K}_1,\mathbf{K}_2}(E^2_{(\mathbf{H}/2)-\mathbf{K}_1-\mathbf{K}_2}-1)(E^2_{\mathbf{K}_1}-1)(E^2_{\mathbf{K}_2}-1)$	$n_K/N^{5/2}$	(*c*)
3(*b*)	$\sum_{\mathbf{K}}(E^2_{\mathbf{H}-\mathbf{K}}-1)(E^2_{\mathbf{K}}-1)(E^2_{\mathbf{H}/2}-1)$	$[n_K/N^{3/2}][4-(7/N)]$	$[n_K(n_K-1)/N^2][16+(42/N)-(435/N^2)+(576/N^3)]$ $+n_K[8-(36/N)+(246/N^2)-(1899/N^3)+(7776/N^4)-(9600/N^5)]$
3(*c*)	$\sum_{\mathbf{K}}(E^2_{(\mathbf{H}/2)-\mathbf{K}}-1)(E^2_{\mathbf{K}}-1)(E^2_{\mathbf{H}}-1)$	$[n_K/N^{3/2}][2-(3/N)]$	$[n_K(n_K-1)/N^2][16+(18/N)-(303/N^2)+(408/N^3)]$ $+n_K[8-(36/N)+(54/N^2)+(117/N^3)-(708/N^4)+(840/N^5)]$
3(*d*)	$\sum_{\mathbf{K}}(E^2_{(\mathbf{H}/2)-\mathbf{K}}-1)(E^2_{\mathbf{K}}-1)(E^2_{\mathbf{H}/2}-1)$	$[8n_K/N^{3/2}][1-(2/N)]$	$[n_K(n_K-1)/N^2][240-(2566/N)+(8731/N^2)-(9115/N^3)]$ $+n_K[8+(220/N)-(3722/N^2)+(21,189/N^3)-(51,524/N^4)+(44,297/N^5)]$
3(*e*)	$\sum_{\mathbf{K}}(E^2_{(\mathbf{H}/2)-\mathbf{K}}-1)(E^2_{\mathbf{K}}-1)(E^2_{2\mathbf{K}}-1)$	$[n_K/N^{3/2}][2-(3/N)]$	$[n_K(n_K-1)/N^3][4+(20/N)-(55/N^2)]$ $+n_K[8+(12/N)-(250/N^2)+(593/N^3)-(408/N^4)]$
3(*f*)	$\sum_{\mathbf{K}}(E^2_{(\mathbf{H}/2)-\mathbf{K}}-1)(E^2_{\mathbf{K}}-1)^2$	$[4n_K/N^{3/2}][1-(2/N)]$	$[n_K(n_K-1)/N^2][128-(1008/N)+(2624/N^2)-(2247/N^3)]$ $+n_K[120-(1116/N)+(3954/N^2)-(6135/N^3)+(3465/N^4)]$
3(*g*)	$\sum_{\mathbf{K}}(E^2_{(\mathbf{H}/2)-\mathbf{K}}-1)(E^2_{(\mathbf{H}/2)+\mathbf{K}}-1)(E^2_{\mathbf{K}}-1)$	$[4n_K/N^{3/2}][1-(2/N)]$	$[n_K(n_K-1)/N^3][144-(768/N)+(960/N^3)]$ $+n_K[8-(36/N)+(246/N^2)-(1547/N^3)+(3872/N^4)-(3200/N^5)]$
3(*h*)	$(E^2_{\mathbf{H}/2}-1)^3$	$[1/N^{1/2}][30-(141/N)+(165/N^2)]$	(*c*)
3(*i*)	$(E^2_{\mathbf{H}}-1)^2(E^2_{\mathbf{H}/2}-1)$	$[1/N^{1/2}][2-(3/N)]$	(*c*)
3(*j*)	$(E^2_{\mathbf{H}}-1)(E^2_{\mathbf{H}/2}-1)^2$	$[1/N^{1/2}][8-(44/N)+(56/N^2)]$	(*c*)

(*a*) See text for remarks concerning this table.

(*b*) n_K is the number of individual terms in $T_{\mathbf{H}}$.

(*c*) These quantities have not been evaluated; it is expected that they would be of small importance in a regression calculation.

It is possible to gain some idea of the reliability of (13) by assuming that

$$E_{\mathbf{H}} - \sum_j C_j T_{\mathbf{H}}^{(j)}$$

has a Gaussian distribution. We first note that

$$\sigma^2 = \overline{(E_{\mathbf{H}} - \sum_j C_j T_{\mathbf{H}}^{(j)})^2} = 1 - \sum_j C_j \overline{E_{\mathbf{H}} T_{\mathbf{H}}^{(j)}} . \quad (16)$$

Then it follows from the assumption of a Gaussian distribution that

$$P_+ = \tfrac{1}{2} + \tfrac{1}{2} \tanh (|E_{\mathbf{H}} \sum_j C_j T_{\mathbf{H}}^{(j)}| / \sigma^2) \quad (17)$$

is the probability that $E_{\mathbf{H}}$ and

$$\sum_j C_j T_{\mathbf{H}}^{(j)}$$

have the same sign.

It is obvious that regression formulae can be obtained from any single term in Table 1 or any group of terms. If a single term is used, we have

$$E_{\mathbf{H}} \approx (\overline{E_{\mathbf{H}} T_{\mathbf{H}}} / \overline{T_{\mathbf{H}}^2}) T_{\mathbf{H}} , \quad (18)$$

and

$$\sigma^2 = 1 - (\overline{E_{\mathbf{H}} T_{\mathbf{H}}})^2 / \overline{T_{\mathbf{H}}^2} . \quad (19)$$

Thus, the numerical values of $(\overline{E_{\mathbf{H}} T_{\mathbf{H}}})^2 / \overline{T_{\mathbf{H}}^2}$ provide an initial estimate of the relative importance of the various terms in Table 1 in determining the sign of $E_{\mathbf{H}}$.

Four of the terms listed in Table 1 have non-zero mean values. These are

$$\begin{aligned} \overline{2(a)} &= 2 - (3/N) , \\ \overline{3(d)} &= n_K[(4/N) - (7/N^2)] , \\ \overline{3(h)} &= 8 - (36/N) + (40/N^2) , \\ \overline{3(j)} &= (6/N) - (11/N^2) . \end{aligned} \quad (20)$$

If any of these terms are used in a regression calculation, the corresponding mean values should be subtracted from them in applying equations (13) and (14).

Tables similar to 1 can be obtained for other invariants of space group $P\bar{1}$ and for invariants of other space groups. In fact, a regression formula for determining the phase (or sign) of $E_{\mathbf{H}_1} E_{\mathbf{H}_2} E_{-\mathbf{H}_1-\mathbf{H}_2}$ for space groups $P1$ and $P\bar{1}$ has already been presented (Vaughan, 1958). Extension can also be made to any desired order. However, the evaluation of the covariances and variances, as well as the solution of the resulting set of linear equations, becomes increasingly more difficult. A means of finding terms which give non-zero covariance with any invariant of any space group will be discussed in a later communication. Bertaut (1955) has given a discussion which applies to this problem.

A numerical example

To show how a regression formula can be obtained from the terms listed in Table 1, and to demonstrate the possible utility of such an equation for sign determination, a numerical test has been made with a synthetic structure. A two-dimensional structure was considered with eight equal atoms at the positions

$$\pm(x, y) = (0{\cdot}162, 0{\cdot}389),\ (0{\cdot}307, 0{\cdot}233),\ (0{\cdot}568, 0{\cdot}214),\ (0{\cdot}794, 0{\cdot}418) .$$

In constructing regression formulae, only terms 1(a), 2(b), 2(c), 3(b), 3(e), and 3(g) in Table 1 were used. The principal reason for disregarding terms was, of course, to keep the calculations within reasonable limits. In selecting these particular terms, consideration was given to the expected magnitudes of $(\overline{E_{\mathbf{H}} T_{\mathbf{H}}})^2 / \overline{T_{\mathbf{H}}^2}$; also, terms were omitted which seemed to be only higher orders of terms which were included. It is admitted that the best choice of terms may not have been made. In carrying out the calculations, it was assumed that all values of $(|E_{\mathbf{H}}|^2 - 1)$ were known within the limitations $|h| \leq 10$, $|k| \leq 10$.

In order to construct regression formulae it was necessary to evaluate the covariances between the six terms which were used. These were computed from the following equations:

$$\begin{aligned}
\overline{1(a)2(b)} &= [1/N][6-(11/N)] \\
\overline{1(a)2(c)} &= [n_K/N][4-(7/N)] \\
\overline{1(a)3(b)} &= [n_K/N^2][12-(23/N)] \\
\overline{1(a)3(e)} &= [n_K/N^2][10-(19/N)] \\
\overline{1(a)3(g)} &= [2n_K/N^2][1-(2/N)] \\
\overline{2(b)2(c)} &= [n_K/N^2][10-(19/N)] \\
\overline{2(b)3(b)} &= [n_K/N][8+(22/N)-(259/N^2)+(368/N^3)] \\
\overline{2(b)3(e)} &= [n_K/N^2][72-(364/N)+(440/N^2)] \\
\overline{2(c)3(b)} &= [n_1 n_2/N^3][20-(39/\mathrm{N})] \\
&\quad +[n_{12}^+/N][18-(10/N)-(59/N^2)+(96/N^3)] \\
&\quad +[n_{12}^-/N^3][32-(63/N)] \\
\overline{2(c)3(e)} &= [n_1 n_2/N^3][18-(36/N)] \\
&\quad +[n_{12}^+/N][12-(40/N)+(33/N^2)] \\
&\quad +[n_{12}^-/N][8+(22/N)-(223/N^2)+(296/N^3)] \\
\overline{2(c)3(g)} &= [n_1 n_2/N^3][4-(8/N)] \\
&\quad +[n_{12}^{\pm}/N^2][48-(236/N)+(280/N^2)] \\
\overline{3(b)3(e)} &= [n_1 n_2/N^3][8+(12/N)-(55/N^2)] \\
&\quad +[n_{12}^+/N^2][24+(42/N)-(611/N^2)+(864/N^3)] \\
&\quad +[n_{12}^-/N^2][16+(90/N)-(711/N^2)+(936/N^3)] \\
\overline{3(b)3(g)} &= [n_1 n_2/N^3][144-(748/N)+(920/N^2)] \\
&\quad +[n_{12}^{\pm}/N^2][128-(480/N)-(360/N^2)+(1616/N^3)] \\
\overline{3(e)3(g)} &= [n_1 n_2/N^3][8-(28/N)+(24/N^2)] \\
&\quad +[n_{12}^{\pm}/N][32-(8/N)-(1148/N^2) \\
&\quad +(3864/N^3)-(3584/N^4)] .
\end{aligned} \quad (21)$$

In this tabulation, n_K, n_1, and n_2 stand for the number of members in a sum term, the latter two being used for the cases in which two sum terms are involved. The symbols n_{12}^{+}, n_{12}^{-}, and $n_{12}^{\pm}$ refer to the number of cases for which $\mathbf{K}_1=\mathbf{K}_2$, $\mathbf{K}_1=-\mathbf{K}_2$, and $\mathbf{K}_1=\pm\mathbf{K}_2$, in a double summation over $\mathbf{K}_1$ and $\mathbf{K}_2$; these were the only special terms which were considered. Again, approximations similar to those described in connection with Table 1 were used.

Since the number of terms which are used to compute the terms 2(*c*), 3(*b*), 3(*e*), and 3(*g*) enter into the determination of the coefficients in a regression formula, one would ideally have to compute different coefficients for almost every **H** for which one wished to determine a sign. However, for any given set of structure factors these numbers vary within rather restricted limits, and only two sets of coefficients were actually determined in this numerical example. The cases to which these coefficients apply are defined in the following way.

Case (1): n_K for $2(c)>163$, n_K for $3(b)>114$

Case (2): n_K for $2(c)<163$, n_K for $3(b)<114$.

It is to be noted that n_K for 3(*g*) is always one less than n_K for 3(*b*), and n_K for 3(*e*) is invariably 119. The numbers n_K which were actually used to compute the coefficients for these two cases were the average values taken over all **H** for which the corresponding inequalities are satisfied, and for which $|E_{\mathbf{H}}|>1\cdot00$. Thus, the following numbers were used.

Case (1): n_K for $2(c)=178$, n_K for $3(b)=139$, n_K for $3(e)=119$, n_K for $3(g)=138$

Case (2): n_K for $2(c)=152$, n_K for $3(b)=96$, n_K for $3(e)=119$, n_K for $3(g)=95$.

The values for n_{12}^{+}, n_{12}^{-}, and $n_{12}^{\pm}$ were estimated averages. Equations (14) were solved for each of these two cases.

Case (1): $\Sigma_{\mathbf{H}}=1\cdot048[1(a)]-0\cdot1337[2(b)]-0\cdot02147[2(c)]+0\cdot004136[3(b)]+0\cdot009297[3(e)]+0\cdot005094[3(g)]$

$\sigma^2=0\cdot622$.

Case (2): $\Sigma_{\mathbf{H}}=0\cdot9256[1(a)]-0\cdot06482[2(b)]-0\cdot02237[2(c)]+0\cdot005204[3(b)]+0\cdot01037[3(e)]+0\cdot003525[3(g)]$

$\sigma^2=0\cdot657$.

These equations give $\Sigma_{\mathbf{H}}$, a least-squares estimate (approximately) of $E_{\mathbf{H}}$ with variance σ^2. These equations were used to compute $\Sigma_{\mathbf{H}}$ for all **H** such that $|E_{\mathbf{H}}|\geq 0\cdot5$. The results for all cases for which $|E_{\mathbf{H}}|\geq 1\cdot0$ are given in Table 2. Column 10 in Table 2 gives the results of the application of equation (17), and gives the (approximate) probabilities (P_+) that $\Sigma_{\mathbf{H}}$ and $E_{\mathbf{H}}$ have the same sign. It will be noted that in eight cases P_+ is greater than 0·95, and in the only case of an incorrect sign determination (8,8) P_+ is only 0·61. There are, of course, a number of examples for which sign determinations would have to be considered unreliable because of low values of P_+. In only one case (0,10) was a sign reliably determined for which $0\cdot5\leq|E_{\mathbf{H}}|\leq 1\cdot0$; in all other examples in this category P_+ was less than 0·8, although the sign of $\Sigma_{\mathbf{H}}$ was the same as that of $E_{\mathbf{H}}$ in twelve out of seventeen cases. In none of the fourteen cases for which $P_+>0\cdot7$ was the sign of $E_{\mathbf{H}}$ incorrectly determined by $\Sigma_{\mathbf{H}}$. Although these results may well be partly fortuitous, they do provide some evidence that the six-term regression formula is useful and that the estimates of P_+ are reasonably conservative.

It is of interest to compare the results obtained with the six-term regression formula with other methods of sign determination. Karle, Hauptman, Karle & Wing (1958) have used $E_{\mathbf{H}/2}^2-1$ to determine the sign of $E_{\mathbf{H}}$. The regression formula obtained with this one term is

$$E_{\mathbf{H}}\approx N^{\frac{1}{2}}(2N-3)^{-1}(E_{\mathbf{H}/2}^2-1)\,, \qquad (22)$$

with

Table 2. *Sign determinations with six-term regression formula*

	Terms in regression formula								P_+ (product)		
h,k	1(*a*)	2(*b*)	2(*c*)	3(*b*)	3(*e*)	3(*g*)	$\Sigma_{\mathbf{H}}$	$E_{\mathbf{H}}$	$E_{\mathbf{H}}\Sigma_{\mathbf{H}}$	$E_{\mathbf{H}}\|(a)$	Case
4,10	−0·99	−4·66	−23·3	−80·1	−14·9	− 6·8	−0·69	−2·39	0·993	0·75	2
8,2	+0·66	+2·59	+ 9·1	+66·7	+16·4	+24·2	+0·71	+2·22	0·993	0·67	1
8,0	−0·99	−2·92	−22·4	−89·4	−19·2	+12·8	−0·65	−1·99	0·985	0·72	1
2,6	−0·99	−2·58	−55·4	−55·4	+30·0	−36·6	+0·36	+1·90	0·901	0·71	1
4,10	+0·55	+1·20	− 8·2	+28·4	+15·8	+13·6	+0·97	+1·79	0·995	0·61	2
0,6	−0·89	−1·83	− 7·9	−59·1	+15·5	+10·2	−0·57	−1·78	0·964	0·68	1
8,6	+0·58	+1·09	+20·3	+18·6	− 5·4	− 2·6	+0·05	+1·70	0·56	0·61	2
10,6	−1·00	−1·88	−12·4	−27·1	−13·2	+27·1	−0·71	−1·70	0·975	0·69	2
6,8	+0·25	+0·30	− 2·2	+ 7·6	+23·1	+16·3	+0·14	+1·49	0·65	0·54	2
4,4	−0·99	−1·18	−44·9	−37·4	+37·4	−15·5	+0·20	+1·48	0·72	0·67	1
4,8	−0·93	−1·06	−11·9	−29·2	−16·8	+ 2·0	−0·85	−1·46	0·977	0·66	2
10,4	+0·45	+0·36	+11·8	+ 7·9	+17·3	+12·2	+0·40	+1·35	0·88	0·57	2
8,4	−0·70	−0·50	− 3·8	−18·0	+15·6	+21·7	−0·40	−1·31	0·87	0·61	2
6,4	−1·00	−0·70	−27·5	− 1·9	+18·6	+13·3	−0·13	−1·30	0·63	0·65	1
8,8	+1·19	+0·39	+37·8	+15·3	− 2·3	−44·4	+0·13	−1·16	0·61	0·66	2
6,2	−0·91	−0·15	−30·0	+19·3	+21·9	+23·3	+0·12	+1·08	0·60	0·61	1
0,10	+4·12	−0·37	+113·1	−43·0	−39·4	−26·7	+1·26	+0·96	0·980	0·87	1

$$\sigma^2=(2N-4)/(2N-3)=0\cdot923 \quad \text{for} \quad N=8\,. \quad (23)$$

Equation (17) would then become

$$P_+=\tfrac{1}{2}+\tfrac{1}{2}\tanh\{|E_{\mathbf{H}}(E^2_{\mathbf{H}/2}-1)|/[N^{\frac{1}{2}}(2-(4/N))]\}\,, \quad (24)$$

which is the same as the formula of Cochran & Woolfson (1955) if the term $4/N$ is omitted. This term should, of course, be included since $P_+=1$ exactly for $N=2$. The results of applying (24) to the numerical example described above are given in column 11 of Table 2. The improvement obtained by using the six-term formulae is seen to be considerable. It should, of course, be remembered that the joint distribution of $E_{\mathbf{H}}$ and $E^2_{\mathbf{K}/2}-1$ is not very close to being Gaussian. Thus, the estimates obtained from (24) are probably rather crude. We note, for example, that the sign of $E_{0,10}$ is actually positive with probability 1 because of the Harker–Kasper inequality $U_{\mathbf{H}}\geq 2U^2_{\mathbf{H}/2}-1$. Within the limitations $|h|\leq 10$, $|k|\leq 10$, the signs of only $E_{0,10}$ and $E_{10,\bar{2}}$ can be determined to be positive with this inequality.

Another interesting formula is that due to Cochran (1954),

$$E_{\mathbf{H}}=N^{\frac{1}{2}}[2(E^2_{\mathbf{H}/2}-1)-N\overline{(E^2_{(\mathbf{H}/2)-\mathbf{K}}-1)(E^2_{\mathbf{K}}-1)}^{\mathbf{K}}]\,, \quad (25)$$

in which the bar indicates averaging over **K**. This equation is valid if all the **K** vectors are included in computing this average. An equation resembling (25) can be obtained by considering a two-term regression formula (see Table 1),

$$E_{\mathbf{H}}\approx C_1[1(a)]+C_2[2(c)]\,. \quad (26)$$

The values of C_1 and C_2 determined by least-squares minimization are

$$C_1 = (2n_KN^{\frac{1}{2}}+\tfrac{1}{2}N^{\frac{5}{2}}P_1)/(n_K+\tfrac{1}{2}N^3P_2)\,,$$

and

$$C_2 = -N^{\frac{3}{2}}/(n_K+\tfrac{1}{2}N^3P_2)\,,$$

in which

$$P_1 = [4N-12+(1/N)+(15/N^2)]/[N-2]\,,$$

$$P_2 = [8N-36+(38/N)+(27/N^2)-(45/N^3)]/[N-2]\,,$$

and n_K is the number of terms in $2(c)$. It is seen that (26) becomes identical with (25) in the limit $n_K\rightarrow\infty$ for $N\neq 2$. Furthermore,

$$\sigma^2=1-(C_1/N^{\frac{1}{2}})-(n_KC_2/N^{\frac{3}{2}})\,, \quad (27)$$

and $\lim_{n_K\rightarrow\infty}\sigma^2=0$, as was to be expected. As Cochran has noted, the fact that $\sigma^2=0$ for infinite n_K does not necessarily mean that (25) or (26) will succeed in practical examples. In the numerical example described above, we would have $\sigma^2=0\cdot82$ for case (1) and $\sigma^2=0\cdot83$ for case (2) if (26) were used for sign determination. Thus, it is clear that (26) is superior to (22), but is still considerably inferior to the six-term formulae which were considered.

Note

Klug (1958) has recently discussed joint distributions of structure factors in considerable detail. He has obtained an expansion in terms of Hermite polynomials in which the terms are ordered such that the coefficients are increasing powers of $1/N^{\frac{1}{2}}$ (in the case of equal atoms). A difficulty with this approach is that it becomes intractable when applied to the joint distribution of a large number of structure factors. Also, from the point of view of regression formulae, Klug's arguments concerning order (the power of $1/N^{\frac{1}{2}}$) lose their validity when the number of structure factors is large because the least-squares coefficients of summations such as $2(c)$, $3(e)$, etc. (see Table 1) depend on n_k as well as N. The important consideration is the relation between n_k and some power of N. This point has been emphasized in a discussion of other regression formulae (Vaughan, 1958) and has also been discussed by Cochran (1958).

References

Bertaut, E. F. (1955*a*). *Acta Cryst.* **8**, 537.

Bertaut, E. F. (1955*b*). *Acta Cryst.* **8**, 823.

Cochran, W. (1954). *Acta Cryst.* **7**, 581.

Cochran, W. (1958). *Acta Cryst.* **11**, 579.

Cochran, W. & Woolfson, M. (1955). *Acta Cryst.* **8**, 1.

Cramer, H. (1946). *Mathematical Methods of Statistics.* New Jersey: Princeton University Press.

Hauptman, H. & Karle, J. (1953). *Solution of the Phase Problem I. The Centrosymmetric Crystal.* A.C.A. Monograph No. 3, Brooklyn: Polycrystal Book Service.

Karle, I. L., Hauptman, H., Karle, J. & Wing, A. B. (1958). *Acta Cryst.* **11**, 257.

Klug, A. (1958). *Acta Cryst.* **11**, 515.

Margenau, H. & Murphy, G. M. (1943). *The Mathematics of Physics and Chemistry*, p. 298. New York: Van Nostrand.

Szegö, G. (1939). *Orthogonal Polynomials, American Mathematical Society Colloquium Publications*, Vol. XXIII. New York.

Vaughan, P. A. (1958). *Acta Cryst.* **11**, 111.