and does not change its value during any shift of the origin of the unit cell.

Definition 2. N-phase seminvariant (or, shortly, seminvariant) is any quantity which depends on N symmetrically independent structure-factor phases and does not change its value after a translation between any two equivalent origins of the unit cell.

The following statements can be easily proved.

1. Any invariant is also seminvariant.
2. Any universal structure invariant is invariant.
3. Any structure seminvariant is also seminvariant.
4. Let $\Phi_1, \ldots, \Phi_n$ be seminvariants (invariants). Then any function $\psi(\Phi_1, \ldots, \Phi_n)$ of only these seminvariants (invariants) is also a seminvariant (invariant).

APPENDIX B

List of *a priori* structure information

I. *A priori structure information necessary for the solution of the phase problem*

1. Measured intensities.
2. Electron density concentrated around individual atoms has unimodal distribution.
3. Overlap of electron density of different atoms may be neglected.

II. *Further a priori structure information used in ab initio methods*

4. The electron density distribution around individual atoms is known.
5. Approximation of spherically symmetrical atoms is applicable.
6. Approximation of scattering factors by an 'overall shape factor' is applicable.
7. Approximation of temperature factors by an 'overall temperature factor' is applicable.
8. The number of 'heavy atoms' (atoms which determine the main features of the diffraction pattern) is known.
9. The number of individual types of atoms in the unit cell is known.
10. Electron density is non-negative everywhere in the unit cell.
11. The crystallographic symmetry is known.
12. The non-crystallographic symmetry is known.
13. Interatomic vectors in the asymmetric part of the unit cell fill the vector space uniformly.

III. *Partial knowledge of the structure*

14. Inner structure of atomic groups with unknown positions and orientations is known.
15. The inner structure and orientation of atomic groups with unknown positions are known.
16. Positions of some atoms are known.

IV. *Repeated intensity measurements under changed conditions*

17. Intensity measurements of isomorphous derivatives.
18. Intensity measurements using the wavelength for which a small number of atoms shows strong anomalous scattering.

References

BUERGER, M. J. (1959). *Vector space.* New York: Wiley.
GIACOVAZZO, C. (1977). *Acta Cryst.* **A33**, 933–944.
HAŠEK, J. (1984*a*). *Acta Cryst.* **A40**, 340–346.
HAŠFK, J. (1984*b*). *Acta Cryst.* **A40**, 346–350.
HAŠEK, J. (1984*c*). *Acta Cryst.* **A40**, 350–352.
HAUPTMAN, H. & KARLE, J. (1956). *Acta Cryst.* **9**, 45–55.
STANFORD, R. H. JR. (1971). *Acta Cryst.* **B27**, 2036.

Acta Cryst. (1984). **A40**, 340–346.

On the Solution of the Phase Problem. II.* Seminvariant Distributions Fitted by Comparing their Function Values

BY J. HAŠEK

Institute of Macromolecular Chemistry, Czechoslovak Academy of Sciences, 162 06 *Prague* 6, *Czechoslovakia*

(*Received* 1 *October* 1982; *accepted* 3 *January* 1984)

Abstract

An *a posteriori* method of the determination of a correct set of structure-factor phases based on a comparison between the trial and theoretical distribution functions of semivariants, using the χ^2 test, makes possible the full utilization of *a priori* structure information contained in the phase relationships. It is expected that the application of this method should raise the efficiency of existing direct methods.

1. Introduction

Direct methods (Giacovazzo, 1980; Ladd & Palmer, 1980; Main, Hull, Lessinger, Germain, Declercq &

* Part I: Hašek (1984*a*).

0108-7673/84/040340-07$01.50

Woolfson, 1978; Hauptman, 1972; Karle & Karle, 1966; Hauptman & Karle 1953) have become a widely used tool for the solution of the phase problem. The good efficiency of these methods requires the proper use of *a priori* structure information contained in the distribution functions of seminvariants (Hašek, 1984*a*). The existing direct methods usually consider only the best possible fit between the trial and the theoretical mean seminvariant values and do not make full use of information on the shape of the distributions. The method suggested here requires an agreement between the whole distributions, and hence makes better use of the *a priori* structure information contained in the phase relationships.

Let us define the following notions.

Empirical distribution of seminvariants, calculated for the correct set of phases, is a function $P^{\rm emp}(\Psi, R_1, \ldots, R_m)$ to which the frequency function of seminvariants for the structure under study converges if the size of regions and intervals, in which the relative frequencies of seminvariants are calculated, go to zero in the limit, while the numbers of seminvariants therein approach infinity.

Trial distribution of seminvariants is defined as an empirical distribution, but the correct set of phases is replaced by some trial set of phases proposed by *ab initio* methods.

True distribution of seminvariants denotes the distribution to which the empirical distribution for the structure under study converges, assuming the exact values of phases and magnitudes.

Theoretical distribution of seminvariants $P^{\rm theor}(\Psi, R_1, \ldots, R_m)$ denotes the probability distribution of seminvariants (usually given analytically) derived theoretically on the basis of *a priori* structure information or semiempirically as a generalization of empirical distributions for a number of different structures.

Following these definitions a correct solution of the phase problem can be found among a certain number of trial solutions according to the best fit between the trial and true distributions of seminvariants. However, since the true distribution of seminvariants remains unknown until the structure is known, it must be approximated in our procedures by a suitable theoretical distribution. Thus, the basic principle lies in finding such a set of phases, whose trial distribution of seminvariants gives the optimal fit with the corresponding theoretical distribution which is assumed to be a sufficiently exact approximation of the true distribution.

2. Empirical and trial distributions of seminvariants

The joint probability distribution of seminvariants $P^{\rm theor}(\Psi, R_1, \ldots, R_m)$ is generally a function of the seminvariant value Ψ and of m magnitudes $R_1, \ldots, R_m$. The m-dimensional space of magnitudes may be divided into regions so that for arbitrary $R_1, \ldots, R_m$ values in the same region, the function $P^{\rm theor}(\Psi|R_1, \ldots, R_m)$ assumes approximately the same values for a fixed Ψ value. Then, it can be assumed that seminvariants belonging to the same region of magnitudes correspond approximately to the same one-dimensional conditional distribution $P^{\rm theor}(\Psi|R_1, \ldots, R_m)$ and the relative frequencies of randomly selected semiinvariant values which occur in various intervals of Ψ values may be used as an estimate of the empirical conditional distribution $P^{\rm emp}(\Psi|R_1, \ldots, R_m)$.

Let the total number of seminvariants in a selected region be denoted by N and the probability that a randomly selected seminvariant from this region falls into the ith interval of Ψ by p_i. Then the numbers of seminvariants $x_1, \ldots, x_r$ falling into the individual intervals of Ψ are given by a multinomial distribution (*cf.* Appendix *A*).* This distribution may be approximated in the limit for $N \to \infty$ by the r-dimensional normal distribution (Bickel & Doksum, 1977)

$$P(\mathbf{x}) = (2\pi)^{-N/2} \det(\mathbf{M}^{-1})^{1/2} \times \exp[-1/2(\mathbf{x} - N\mathbf{p})\mathbf{M}^{-1}(\mathbf{x} - N\mathbf{p})], \quad (1)$$

where the vector $\mathbf{x} \equiv (x_1, \ldots, x_r)$, the vector $\mathbf{p} \equiv (p_1, \ldots, p_r)$ and $\mathbf{M}^{-1}$ is a matrix of rank $(r-1)$ inverse to the variance–covariance matrix $\mathbf{M}$, the diagonal elements of which are

$$\operatorname{var}(x_i) = Np_i(1 - p_i) \quad (2)$$

and off-diagonal elements

$$\operatorname{cov}(x_i, x_j) = -Np_ip_j. \quad (3)$$

Thus, by increasing the number of randomly selected seminvariants the relative frequencies $Q_i^{\rm emp} = x_i/N$ approach the theoretical probabilities p_i that the selected seminvariant lies within the ith interval with a variance $\sigma_i = [p_i(1-p_i)/N]^{1/2}$. Since for $N \to \infty$ we have $\sigma_i \to 0$, the relative frequency of seminvariants in the ith interval of Ψ,

$$Q_i^{\rm emp} = x_i/N, \quad (4)$$

is a consistent estimate of the true probability p_i (see Appendix *A*). The $Q_i^{\rm emp}$ values calculated using (4) approximate at several points the conditional probability distribution of seminvariants $P(\Psi|R_1, \ldots, R_m)$ for fixed values of the magnitudes.†

* Appendices *A*, *B* and *C* have been deposited with the British Library Lending Division as Supplementary Publication No. SUP 39114 (10 pp.). Copies may be obtained through The Executive Secretary, International Union of Crystallography, 5 Abbey Square, Chester CH1 2HU, England. †

† It is possible to estimate the joint probability distribution directly, but, to diminish problems arising from irregular distributions of seminvariants in different regions, it is better to substitute for the empirical joint probability distribution the renormalized frequency function composed of a number of one-dimensional conditional distributions for individual regions of magnitudes.

An example of the possible form of the distribution function of seminvariants is shown in Fig. 1. To draw the function $Q(\Psi, R_1, \ldots, R_m) = Q(\Psi, \mathbf{w})$ in three dimensions, the m-dimensional space of magnitudes is represented by a vector $\mathbf{w}=(R_1, \ldots, R_m)$. The m-dimensional space of magnitudes is considered to be divided into *regions* and every region is further divided into *intervals* according to the value of the seminvariant. Equation (4) allows an estimate of the distribution in one region of magnitude, *i.e.* an estimate of just one profile in Fig. 1. To obtain the description of the whole distribution, the calculation must be repeated for each region of magnitudes. The relative frequency of seminvariants of the kth type belonging to the jth region and to the ith interval is then

$$Q^{\rm emp}_{ijk} = N_{ijk}/N_{jk}, \tag{5}$$

where N_{ijk} is the number of the seminvariants in the jth region, the value of which lies in the ith interval, and N_{jk} is the total number of seminvariants in the jth region. The calculated $Q^{\rm emp}_{ijk}$ values are expected to better approximate the empirical distribution, the smaller the regions of magnitudes and the greater the numbers of seminvariants in the individual regions of magnitudes.

In the case of special seminvariants, the phases of which can assume only two values owing to the crystallographic symmetry, the condition (5) gives

$$Q^{\rm emp}_{1jk} = 1 - Q^{\rm emp}_{2jk} \tag{6}$$

for all the regions $j = 1, \ldots, n$ and all the distribution types $k = 1, \ldots, s$. Hence, to obtain a full description of an empirical distribution, it is sufficient to calculate only one $Q^{\rm emp}_{1jk} = N_{1jk}/N_{jk}$ value for each region. Evidently, $Q^{\rm emp}_{1jk}$ is an estimate of the probability $P^{\rm emp}_{1jk}$ that a seminvariant of the kth type belonging to the jth region assumes just the value belonging to the first interval. An estimate of the probability of a positive sign for centric structure seminvariants is

$$P^{\rm emp}_{+jk} = Q^{\rm emp}_{1jk} = N_{+jk}/N_{jk}, \tag{7}$$

where N_{+jk} is the number of structure seminvariants which assume the value $0 \bmod (2\pi)$ and N_{jk} is the total number of structure seminvariants in the jth region. An example of such a distribution is shown in Fig. 2.

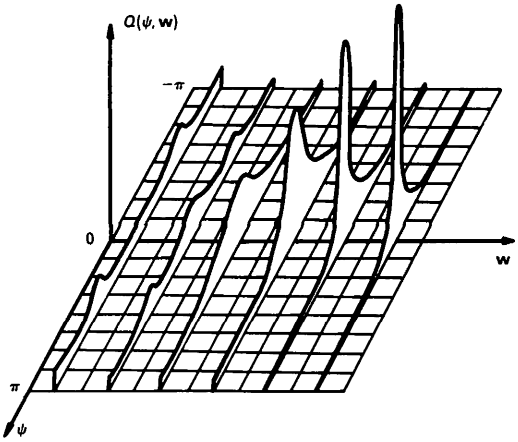

Fig. 1. An example of the distribution of acentric seminvariants. The $(m+1)$-dimensional space of parameters of the distribution $Q(\Psi, \mathbf{w})$ is divided into regions of $\mathbf{w} = f(R_1, \ldots, R_m)$ and into intervals of Ψ.

A trial distribution of seminvariants $Q^{\rm trial}_{ijk}$ ($P^{\rm trial}_{+jk}$) is calculated using the same formalism, only the correct set of phases is replaced by a trial one, proposed by *ab initio* methods.

3. Theoretical distribution function

The probability distribution of any seminvariant can be calculated from the assumed distribution of $\mathbf{r}_j$ vectors (for fixed $\mathbf{H}$ vectors) or from the assumed distribution of diffraction vectors $\mathbf{H}$ (for fixed $\mathbf{r}_j$ vectors.* In the first case, where the position vectors of atoms or of groups of atoms are taken as primitive random variables, it holds that with increasing number of randomly chosen structures the probability $\int P(\Psi|R_1, \ldots, R_m)\,{\rm d}\Psi\,{\rm d}R_1 \ldots {\rm d}R_m$ converges to the fraction of seminvariants of a given type, whose values Ψ and magnitudes $R_1, \ldots, R_m$ lie in the range of integration limits.

In the second case,† where the diffraction vectors $\mathbf{H}_1, \ldots, \mathbf{H}_r$ are taken as primitive random variables, whereas the structure is fixed, with increasing number of randomly chosen seminvariants Ψ of a given type,

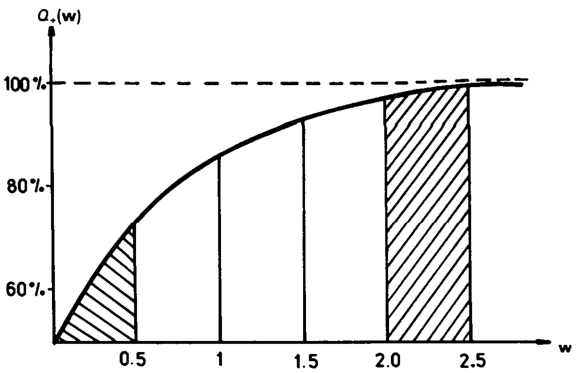

Fig. 2. An example of the centric seminvariant distribution. The m-dimensional space of parameters of the distribution $Q_+(\mathbf{w})$ is divided into regions of magnitudes $\mathbf{w} \equiv f(R_1, \ldots, R_m)$.

* In practice, the same probability of finding any atom or any group of atoms at any site in the unit cell or a random choice of diffraction vectors $\mathbf{H}$ in reciprocal space is usually assumed.

† It should be noted that the second case corresponds to the procedure employed in the calculation of the empirical distribution.

the probability $\int P'(\Psi|R_1,\ldots,R_m)\,d\Psi\,dR,\ldots,dR_m$ converges also to a fraction of those seminvariants Ψ, which assume values within the integration interval Ψ and whose magnitudes $R_1,\ldots,R_m$ lie in the range of integration limits.

Since the position vectors of atoms **r** and the diffraction vectors **H** appear in any seminvariant only through the product **Hr**, the same distribution of seminvariants $p(\Psi|R_1,\ldots,R_n)$ is obtained in both cases, assuming the same distribution of scalar product, **Hr**, although their interpretation is different (Hauptman & Karle, 1953). The equivalence of the two distributions allows us to use the theoretical distributions of seminvariants obtained by both procedures for comparison with the empirical distributions estimated using the relative frequencies of seminvariant values.

In the case of acentric seminvariants* (*i.e.* of quantities which can be expressed as a monotonic function of one or several structure seminvariants, at least one of which is acentric) the distribution $P(\Psi, R_1,\ldots,R_m)$ is generally a function of $(m+1)$ variables. If all the structure seminvariants Φ_i $(i=1,\ldots,s)$ forming the seminvariant Ψ are centric, the distribution $P(\Psi|R_1,\ldots,R_m)$ is discrete, and the seminvariant Ψ may assume 2^s values at most. Hence, the total joint probability distribution can be fully described using (2^s-1) m-dimensional functions. In particular, in the case where the seminvariant Ψ is regarded as a function of only one centric structure seminvariant, the distribution is fully described by a single function, $P_{\Psi_0}(R_1,\ldots,R_m)$. For centric structure seminvariants the description of the distribution $P(\Phi, R_1,\ldots,R_m)$ is reduced to the description of the probability that the structure seminvariant Φ assumes the values 0 or π, corresponding to a positive or a negative sign of the respective structure-factor product. Since the probability of the positive sign $P_+=1-P_-$, the distribution can be adequately described using merely the m-dimensional function P_+ which expresses the probability that the structure seminvariant assumes just the value 0.

Suppose that the theoretical distribution is identical with the true distribution of seminvariants. For the correct set of phases and fixed magnitudes $R_1,\ldots,R_m$, the relative frequency of occurrence of randomly chosen seminvariants in the interval (Ψ_i, Ψ_{i+1}) approaches in the limit for increasing number of seminvariants its theoretical value

$$Q^{\text{theor}} = \int_{\Psi_i}^{\Psi_{i+1}} P(\Psi|R_1,\ldots,R_m)\,d\Psi, \qquad (8)$$

corresponding to the probability that a randomly chosen seminvariant lies in the interval (Ψ_i, Ψ_{i+1}), supposing that the normalizing condition is fulfilled. However, comparatively large regions of magnitudes are necessary to achieve sufficiently high numbers of seminvariants for the calculation of Q_{ijk}^{emp} values. Then, assuming a random choice of seminvariants, the relative frequency Q_{ijk}^{emp} approaches in the limit for $N_{jk}\to\infty$ the theoretically derived value

$$Q_{ijk}^{\text{theor}} = V_{jk}^{-1}\int P_k(\Psi|R_1,\ldots,R_m)\,dR_1\ldots dR_m\,d\Psi, \qquad (9a)$$

where integration runs over the ith interval of Ψ values and over the jth region of magnitudes. The normalizing constant V_{jk} is

$$V_{jk} = \int P_k(\Psi|R_1,\ldots,R_m)\,dR_1\ldots dR_m\,d\Psi,$$

where integration proceeds over the whole jth region and all possible seminvariant values.

It very often happens in practice that there are small numbers of non-uniformly distributed seminvariants within the individual regions of magnitudes. Then, the estimate of the theoretical distribution by the mean value of contributions from the individual seminvariants in the corresponding region is expected to be a better approximation of the empirical probability distribution,

$$Q_{ijk}^{\text{theor}} \doteq N_{jk}^{-1}\sum_l \int P_k(\Psi|R_{1l},\ldots,R_{ml})\,d\Psi. \qquad (9b)$$

The summation over the index l runs over all N_j seminvariants contained in the jth region of magnitudes where the lth seminvariant has phasing magnitudes $R_{1l},\ldots,R_{ml}$.

If a linear approximation of the probability density is sufficient in the interval (Ψ_i, Ψ_{i+1}), the Q_{ijk}^{theor} value will be close to the theoretical probability density for the mean value $\langle\Psi\rangle_i$ in this interval:

$$Q_{ijk}^{\text{theor}} \simeq |\Psi_i - \Psi_{i+1}|P_k(\langle\Psi\rangle_i|\langle R_1\rangle_j,\ldots,\langle R_m\rangle_j), \qquad (9c)$$

where the normalization condition

$$\sum_{i=1}^{r} |\Psi_i - \Psi_{i+2}|P_k(\langle\Psi\rangle_i|\langle R_1\rangle_j,\ldots,\langle R_m\rangle_j) = 1$$

must be fulfilled for every region $(j=1,\ldots,n)$ and every distribution $(k=1,\ldots,q)$, r being the number of intervals. It is obvious from (9a), (9b), (9c) that $0\le Q_{ijk}^{\text{theor}}\le 1$ and

$$\sum_{i=1}^{r} Q_{ijk}^{\text{theor}} = 1 \quad \text{for every } j, k. \qquad (10a)$$

For seminvariants which, owing to the crystallographic symmetry, may assume only two values, the normalizing condition (10a) is reduced to

$$Q_{1jk}^{\text{theor}} = 1 - Q_{2jk}^{\text{theor}} \quad \text{for every } j, k. \qquad (10b)$$

For centric structure seminvariants, Q_{1jk}^{theor} is equivalent to the probability P_{+jk}^{theor} of a positive sign of the corresponding structure-factor product. It can

*For the difference between 'centric' and 'centrosymmetric' see Rogers (1965).

easily be proved, in this case, that the expressions (9) are reduced to

$$P_{+jk}^{\text{theor}} = V_{jk}^{-1} \int P_{+k}(R_1, \dots, R_m)\, \mathrm{d}R_1 \dots \mathrm{d}R_m \qquad (11a)$$

$$P_{+jk}^{\text{theor}} \doteq N_{jk}^{-1} \sum_l P_{+k}(R_{1l}, \dots, R_{ml}) \qquad (11b)$$

$$P_{+jk}^{\text{theor}} \simeq P_{+k}(\langle R_1 \rangle_j, \dots, \langle R_m \rangle_j). \qquad (11c)$$

4. Distribution-fitting method

The frequency function of seminvariants converges for the correct set of phases and for increasing numbers of randomly selected seminvariants to the true distribution. Thus, using exact magnitudes, the empirical distribution is a consistent estimate of the true distribution. Thus, no trial distribution, calculated for an incorrect set of phases, can approximate the true distribution better than the empirical one. Of course, in practice, the unknown true distribution has to be replaced by the corresponding theoretical distribution. Hence, the basic principle of the phase-problem solution by direct methods should consist of finding such a set of phases, for which the trial distribution of seminvariants fits best the corresponding theoretical distribution. The discriminative power of the method depends greatly on the uncertainty in the theoretical estimation of the true distribution. Thus, the choice of a good theoretical or semiempirical approximation of the true distribution and the analysis of expected differences have to be given special care. However, in this paper, only the general formalism of the distribution-fitting method will be considered. The analysis of the individual distribution types will be given later, in the discussion of practical results using these methods.

There are three basic measures of the fit between two distributions that can be used to find the trial distribution which fits best the corresponding theoretical one.

1. A comparison of the trial and theoretical probability distributions by their function values. Usually, the set of phases which gives the minimum sum of squared deviations is expected to be the correct one (χ^2 test).

2. A comparison of the characteristic functions corresponding to the theoretical and trial probability distributions, *i.e.* a comparison of the low-order distribution moments.

3. A test of maximal difference between the corresponding theoretical and trial cumulative probability distributions (Kolmogorov test).

Only some simple tests related to the first procedure have been described previously (*e.g.* Hašek, 1974, 1979). The second procedure may be related to a number of figures of merit frequently used in direct methods. However, most of them compare only characteristics related to the first distribution moments, thus losing significant amounts of *a priori* structure information. Reviews of these methods may be found in Schenk (1980) and Ladd & Palmer (1980). The utilization of information on second distribution moments has been described by Hašek (1975). The third method analysing the fit between the cumulative distributions of seminvariants has not yet been used. These three methods will be dealt with separately. The first, being of primary importance, is discussed below.

Minimization of the sum of weighted squared deviations

Comparison of the theoretical and empirical probability distributions of seminvariants by their function values shows that a suitable criterion of the best fit [see Appendix *B*, equations (*B*4), (*B*6)]* is the minimum of the sum of quadratic forms

$$K = \sum_j^n (\mathbf{Q}_j^{\text{trial}} - \mathbf{Q}_j^{\text{theor}})\mathcal{M}_j^{-1}(\mathbf{Q}_j^{\text{trial}} - \mathbf{Q}_j^{\text{theor}}), \qquad (12)$$

where $\mathcal{M}_j^{-1}$ is an inverse matrix to the variance–covariance matrix, vectors $\mathbf{Q}_j^{\text{trial}} \equiv (Q_{1j}^{\text{trial}}, \dots, Q_{rj}^{\text{trial}})$, $\mathbf{Q}_j^{\text{theor}} \equiv (Q_{1j}^{\text{theor}}, \dots, Q_{rj}^{\text{theor}})$ and the summation runs over all regions of magnitudes $R_1, \dots, R_m$.

If the theoretical distributions describe exactly the corresponding true probability distributions, the criterion (12) would give a minimum-variance unbiased estimate of structure-factor phases. However, owing to the differences between the theoretical and true distributions, the reliability of the estimator is restricted by the extent of *a priori* structure information and by approximations made in the calculation of the theoretical distributions used in the test.

If q various types of seminvariants are tested, the criterion of the fit can be written as a sum of quadratic forms corresponding to the individual distributions [Appendix *B*, equation (*B*9)].* In each of these quadratic forms the matrix $\mathcal{M}_j^{-1}$ can be replaced by a generalized inverse to the matrix $\mathcal{M}_j$ (see Appendix *C*).* Thus, the sum of the quadratic forms K_k is reduced to the weighted sum of squares of the differences $Q^{\text{trial}} - Q^{\text{theor}}$.

A general criterion for a determination of the most probable set of structure-factor phases is then

$$K = \sum_{k=1}^{q} \sum_{j=1}^{n_k} \sum_{i=1}^{r_{jk}} w_{ijk}(Q_{ijk}^{\text{trial}} - Q_{ijk}^{\text{theor}})^2. \qquad (13)$$

Only $(r_{jk} - 1)$ values of Q_{ijk}^{trial} and Q_{ijk}^{theor} are independent for fixed indexes j, k because of constraints $\sum_{i=1}^{r} Q_{ijk}^{\text{trial}} = 1$ and $\sum_{i=1}^{r} Q_{ijk}^{\text{theor}} = 1$. The index i runs over r_{jk} intervals of the seminvariant values,† the

* See deposition footnote.

† If anomalous scattering is neglected, probability distributions are symmetrical around zero and therefore it is convenient to deal with absolute values of structure seminvariants, so that the computation is kept in the interval $\langle 0, \pi \rangle$.

index j runs over n_k regions of magnitudes and the index k runs over q various probability distribution types. The most probable set of phases is denoted by the minimal value of the distribution-fitting coefficient (13).

The weights w_{ijk} generally depend on the type of the seminvariant, on the respective region and interval, on the quality of *a priori* structure information contained in the theoretical distributions and on the restrictions and approximations used in their derivation. The relative importance of seminvariants in intervals corresponding to small or high Ψ values can be stressed or lowered depending on the index i. Depending on the index j, the weights w_{ijk} express a measure of the reliability of the determination of the differences $Q_{ijk}^{\text{trial}} - Q_{ijk}^{\text{theor}}$ in the individual regions of magnitudes.

Let us suppose that the function values of the theoretical distribution Q_{ijk}^{theor} represent exactly the function values of the true distribution in the whole area tested. Then the differences between the empirical and theoretical distributions are described by a multinomial distribution [see Appendix *A* equation (*A*1)]*, converging for increasing number of randomly selected seminvariants to the normal distribution (1). The general inverse matrix V_{jk} to the variance–covariance matrix $\mathcal{M}_{jk}$ (the indices j, k denote the region of magnitudes and seminvariant type) has diagonal elements $v_{rr}^{jk} = N_{jk} Q_{rjk}^{\text{theor}} = N_{rjk}^{\text{theor}}$, and off-diagonal elements $v_{rs}^{jk}\,(r \neq s)$ are zero. Neglecting all other sources of errors and using relative frequencies instead of numbers of seminvariants [see Appendix *B*, equation (*B*4)],* one obtains the weights w_{ijk} in (13):

$$w_{ijk} = N_{jk}^2 / N_{ijk}^{\text{theor}} = N_{jk} / Q_{ijk}^{\text{theor}}. \qquad (14)$$

Under these conditions the distribution-fitting coefficient is

$$K = \sum_k^q \sum_j^{n_k} N_{jk} \sum_i^{r_{jk}} \frac{(Q_{ijk}^{\text{trial}} - Q_{ijk}^{\text{theor}})^2}{Q_{ijk}^{\text{theor}}} \qquad (15)$$

and, with the normalizing condition (10*a*),

$$K = \sum_k^q \sum_j^{n_k} N_{jk} \left[\sum_j^{r_{jk}} \frac{(Q_{ijk}^{\text{trial}})^2}{Q_{ijk}^{\text{theor}}} - 1 \right]. \qquad (16)$$

In the procedures usually employed for the calculation of trial sets of phases, seminvariants of a certain type (usually triplets) are preferred and moreover they are restricted to seminvariants belonging to a special region of magnitudes (*e.g.* only to triplets with high values of $|E_H E_K E_{-H-K}|$). All sets of phases thus derived already give a comparatively good fit in the regions of magnitude used in their calculation. Therefore, it is advantageous to consider especially those regions of magnitudes in which the fit has not yet been ensured by any means, and such types of seminvariants which have not been used in the calculation of the trial sets. For example, if triplet relations are used for the calculation of trial sets, then the quartets, especially in regions corresponding to the mean theoretical value of quartet $\Psi = \varphi_H + \varphi_K + \varphi_L + \varphi_{-H-K-L} \simeq \pi$ or $\pi/2 \bmod (2\pi)$, are of great importance (Schenk, 1980; Ladd & Palmer 1980). The optimal choice of the proper types of probability distributions, the division of space into regions and intervals, and the optimal choice of weights, so that *a priori* structure information can be used in the most effective way, requires further extensive study.

Special seminvariants

It has been shown in preceding sections that in the case of special seminvariants which, owing to the crystallographic symmetry, can assume only two values (*e.g.* centric structure seminvariants), the description of distribution is reduced to the description of the probability that the seminvariant assumes one of the two possible values. From (6) and (10*b*) the distribution-fitting coefficient (13) may be written as

$$K = \sum_k \sum_j (w_{1jk} + w_{2jk})(Q_{1jk}^{\text{trial}} - Q_{1jk}^{\text{theor}})^2. \qquad (17)$$

If the weights are chosen according to (14), then

$$K = \sum_k \sum_j \frac{N_{jk}}{Q_{1jk}^{\text{theor}}(1 - Q_{1jk}^{\text{theor}})} (Q_{1jk}^{\text{trial}} - Q_{1jk}^{\text{theor}})^2. \qquad (18)$$

In the case of centric structure seminvariants, the Q_{1jk}^{trial} and Q_{1jk}^{theor} values have the meaning of the probabilities of a positive sign for the corresponding structure semivariants P_{+jk}^{emp} and P_{+jk}^{theor}, respectively. Moreover, if only one type of probability distribution is used in the test, the summation over the index k can be omitted in (18) and the distribution-fitting coefficient for centric seminvariants is

$$\mathcal{K} = \sum_j w_j (P_{+j}^{\text{trial}} - P_{+j}^{\text{theor}})^2, \qquad (19)$$

where $w_j = N_j / P_{+j}^{\text{theor}}(1 - P_{+j}^{\text{theor}})$. The coefficient $\mathcal{K}'$ defined for triplet relationships (Hašek, 1974) is similar to (19) with the difference that the contributions of highly reliable triplets were stressed by weighting. The efficiency of this coefficient was tested using the chlorate of 4,4′-bis(dimethylamino)diphenylamine radical (space group $P\bar{1}$) and *cis*-1,3,5-trichlorocyclohexane (space group $C2/c$).

5. Conclusion

Success of the phase-determining procedure depends on the quality of the *a priori* structure information used. The first source of this information is a sufficient

* See deposition footnote.

number of experimentally derived magnitudes of the normalized structure factors describing the simplified structure (non-vibrating point atoms). The second source, also necessary for the solution of the phase problem, is concealed in the function form of the distributions of seminvariants. Unlike all the preceding methods, the distribution fit proposed here makes full use of structure information contained in the seminvariant probability distribution functions and so is expected to be more powerful and efficient. The procedure outlined in this paper has been treated only from a general point of view. The optimal choice of the theoretical distribution functions, the determination of the generalized coordinates and the selection of seminvariants for the test so as to ensure an economic and reliable determination of the correct set of phases is discussed in the following papers (Hašek, 1984*b*, *c*).

The author thanks Dr K. Huml for valuable comments on this work.

References

Bickel, P. J. & Doksum, K. A. (1977). *Mathematical Statistics: Basic Ideas and Selected Topics.* San Francisco: Holden-Day.

Giacovazzo, C. (1980). *Direct Methods in Crystallography.* New York: Academic Press.

Hašek, J. (1974). *Acta Cryst.* **A30**, 576–579.

Hašek, J. (1975). *Acta Cryst.* **A31**, 818–819.

Hašek, J. (1979). In *Proceedings of Symposium on Special Problems in X-ray Structure Analysis*, pp. 108–111. Berlin: Hohengruen, ZIPC.

Hašek, J. (1984*a*). *Acta Cryst.* **A40**, 338–340.

Hašek, J. (1984*b*). *Acta Cryst.* **A40**, 346–350.

Hašek, J. (1984*c*). *Acta Cryst.* **A40**, 350–352.

Hauptman, H. (1972). *Crystal Structure Determination.* New York: Plenum Press.

Hauptman, H. & Karle, J. (1953). *Solution of the Phase Problem.* I. *Am. Chem. Soc. Monogr.* No. 3.

Karle, J. & Karle, I. L. (1966). *Acta Cryst.* **21**, 849–859.

Ladd, M. F. C. & Palmer, R. A. (1980). *Theory and Practice of Direct Methods in Crystallography.* New York: Plenum Press.

Main, P., Hull, S. E., Lessinger, L., Germain, G., Declercq, J.-P. & Woolfson, M. M. (1978). *MULTAN*78. Department of Physics, Univ. of York, England.

Rogers, D. (1965). In *Computing Methods in Crystallography*, edited by J. S. Rollet. London: Pergamon Press.

Schenk, H. (1980). In *Computing in Crystallography.* Bangalore; Indian Academy of Sciences.

Acta Cryst. (1984). **A40**, 346–350

On the Solution of the Phase Problem. III.* Distributions Fitted by Comparing their Moments

By J. Hašek

Institute of Macromolecular Chemistry, Czechoslovak Academy of Sciences, 162 06 *Prague* 6, *Czechoslovakia*

(*Received* 1 *October* 1982; *accepted* 3 *January* 1984)

Abstract

The proposed method of determination of the correct set of phases of structure factors enables in principle full benefit to be taken of *a priori* structure information contained in the probability distributions of seminvariants. Unlike the direct comparison of the probability distributions discussed in the preceding paper, the method discussed here, by neglecting the moments of higher orders, allows concentration on the main characteristics of the distributions taken for the test. The basic principle of the method for determination of the correct set of phases using the fit between moments of the theoretical and trial distributions has been widely used in different modifications. However, most of these figures of merit compare only first distribution moments. In many cases this results in insufficient discriminating ability. The comparison of the second moments raises the effectiveness of these methods and may be useful in the last stage of the phase-problem solution. The utilization of moments of higher orders may be dangerous, especially using the global coefficient of moments fitting and in the case of a small number of seminvariants (unreliable determination of higher moments). The method of successive comparison of moments of different orders seems to be more reliable and economical. It permits the survey of a large number of potential solutions, thus increasing the likelihood that a correct solution is included. From the economic point of view, it is convenient to include only those regions of magnitudes and those distribution types which have not been used in the preceding step of the search of the trial solutions. It explains the excellent results obtained using figures of merit based on the special seminvariant types, *e.g.* NQEST, NQC [De Titta, Edmonds, Langs & Hauptman (1975). *Acta Cryst.* **A31**, 472–479; Schenk (1974). *Acta Cryst.* **A30**, 477–481].

* Part II: Hašek (1984*b*).

0108-7673/84/040346-05$01.50