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.* A**30**, 576–579.
HAŠEK, J. (1975). *Acta Cryst.* A**31**, 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.* A**40**, 338–340.
HAŠEK, J. (1984*b*). *Acta Cryst.* A**40**, 346–350.
HAŠEK, J. (1984*c*). *Acta Cryst.* A**40**, 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). *MULTAN78.* 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). A**40**, 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.* A**31**, 472–479; Schenk (1974). *Acta Cryst.* A**30**, 477–481].

* Part II: Hašek (1984*b*).

0108-7673/84/040346-05$01.50

1. Introduction

Hašek (1984*b*) presents a general method for the determination of a correct set of structure-factor phases based on a direct comparison of function values of theoretical and trial probability distributions of seminvariants. However, both probability distributions* can be uniquely described by means of their characteristic functions. Thus also a comparison of the corresponding moments of the theoretical and trial distributions can serve as a criterion of correctness of the calculated set of phases. The thoroughness of the description of both distributions depends on the order of moments used, on the reliability of their determination and on the fineness of partitioning of the m-dimensional space into regions of magnitude. It seems that, compared with the method of the direct distribution fitting (paper II), the method of moments may have some advantages in the case of an insufficient number of seminvariants for statistical treatment. It enables us to concentrate attention only on the main distribution characteristics contained in the two first moments. This idea leads to very simple criteria for finding the correct set of phases, still keeping the discrimination of the test in a satisfactory range. As is shown in § 4.2, both methods coincide in the case of seminvariants, which owing to the crystallographic symmetry may assume only two values (*e.g.* centric structure seminvariants).

The theoretical and empirical distributions may be compared either by using moments with respect to zero, or by means of central moments and cumulants. The distribution moments with respect to zero are primary, but, if problems arise, then the variance and standardized cumulants, giving a better view of the differences between the distributions, are to be preferred.

2. Moments of the empirical distribution of seminvariants

For reasons explained in § 2 of paper II, it is convenient to replace in our calculations the overall $(m+1)$-dimensional empirical probability distribution $P^{\text{emp}}(\Psi, R_1, \ldots, R_m)$ by a function composed of a number of one-dimensional conditional probability distributions $P^{\text{emp}}(\Psi|R_1, \ldots, R_m)$ defined for the individual regions of magnitudes.

The distribution moments (with respect to zero) are calculated using the relation†

$$\mu_{rj}^{\text{emp}} = N_j^{-1} \sum_l \Psi_l^r, \tag{1}$$

where index r denotes the order of moment and index j the region of magnitudes. Summation runs over all seminvariant values in the jth region.

Starting from the second moments, the central moments possess a simpler geometrical meaning. The second central moment of the conditional distribution $P^{\text{emp}}(\Psi|\mathbf{R}_j)$ is denoted by $m_{2j}^{\text{emp}} = \text{var}\,(\Psi)_j^{\text{emp}}$ and calculated using the seminvariant values Ψ of the tested set of phases

$$m_{2j}^{\text{emp}} = \text{var}\,(\Psi)_j^{\text{emp}} = \mu_{2j}^{\text{emp}} - (\mu_{1j}^{\text{emp}})^2. \tag{2}$$

The third central moment of the distribution $P(\Psi|R_j)$ is calculated by

$$m_{3j}^{\text{emp}} = \mu_{3j}^{\text{emp}} - 3\mu_{2j}^{\text{emp}}\mu_{1j}^{\text{emp}} + 2(\mu_{1j}^{\text{emp}})^3. \tag{3}$$

The higher central moments may be computed similarly. Distributions of centric seminvariants are fully described by the first moments only, *i.e.* by

$$\mu_{1j}^{\text{emp}} = \frac{N_{1j}\Psi_{1j} + N_{2j}\Psi_{2j}}{N_{1j} + N_{2j}}$$

values, where N_{1j}, N_{2j} are the numbers of seminvariants assuming the values Ψ_{1j} or Ψ_{2j}, respectively. If the tested seminvariant type is a linear combination of phases, then $\Psi_{2j} = 0$ and

$$\mu_{1j}^{\text{emp}} = N_{1j}/(N_{1j} + N_{2j})\Psi_{1j} = Q_{1j}^{\text{emp}}\Psi_{1j}$$

and for cosine seminvariants simply

$$\mu_{1j}^{\text{emp}} = Q_{1j}^{\text{emp}},$$

where Q_{1j}^{emp} is the relative frequency of seminvariants in the jth region (see § 2 of paper II). All seminvariant values in a single interval may be approximated by the mid-point of the interval Ψ_0. The moment calculation (1) is thus simplified to

$$\mu_j^{\text{emp}} \doteq N_j^{-1} \sum_{i=1}^{s} N_{ij}\Psi_{0j}^r, \tag{4}$$

where r is the order of the moment, N_{ij} is the number of seminvariants in the ith interval and jth region of magnitudes and summation runs over all s intervals.

There are several semi-empirical rules for formation of the seminvariant intervals. Without considering the actual shape of the theoretical distribution, the intervals are recommended to be equidistant and their number is given by some of the following rules:

$$\begin{aligned} r &\leq 5 \log N_j \\ r &\doteq \sqrt{N_j} \\ r &\doteq 1 + 3{\cdot}3 \log N_j, \end{aligned} \tag{5}$$

where n is the sample size.

The use of the mid-points of intervals instead of the actual empirical seminvariant values simplifies the calculations at the expense of some accuracy. Supposing equidistant intervals, unimodal symmetrical distribution and a large sample size $N_j > 500$,

---

For definitions of the true, trial, empirical and theoretical probability distributions of seminvariants, *a priori* structure information and the seminvariant see preceding papers (Hašek, 1984*a*, *b*), referred to henceforth as papers I and II.

†Special care should be taken to keep the averaging correct modulo 2π.

the bias of the central moments of the even order can be corrected using

$$m_2' = m_2 - h^2/12$$
$$m_4' = m_4 - h^2 m_2/2 + 7h^4/240, \quad (6)$$

where h is the length of equidistant intervals. The mean is assumed to be unbiased with maximum possible error $h/2$.

3. Moments of the theoretical distribution of seminvariants

The moments of the conditional distribution $P(\Psi|\mathbf{R}_j)$ are given by

$$\mu_{rj}^{\text{theor}} = \int \Psi^r P(\Psi|\mathbf{R}_j)\, \mathrm{d}\Psi. \quad (7)$$

However, since in practice the regions of magnitudes are usually rather wide, one of the following three approximations has to be used to estimate the theoretical quantities approximating the empirical distribution moments μ_{ij} for a correct set of phases.

Supposing a uniform distribution of seminvariants in the jth region of magnitudes, the first moment may be estimated using the relation

$$\mu_{1j}^{\text{theor}} = \int\int \Psi P(\Psi|\mathbf{R})\, \mathrm{d}\Psi\, \mathrm{d}\mathbf{R}, \quad (8a)$$

where the integration proceeds over all seminvariant values and over the whole jth region. If it is impractical to ensure a uniform choice of seminvariants in the whole region, then it is better to estimate the theoretical moment in the jth region as the average value of moments for all seminvariant values in the jth region

$$\mu_{1j}^{\text{theor}} \doteq \frac{1}{N_j} \sum_l \int \Psi P(\Psi|\mathbf{R}_l)\, \mathrm{d}\Psi, \quad (8b)$$

where the summation runs over all the N_j seminvariants in the jth region of magnitudes and integration over all possible seminvariant values. In many cases it is sufficient to estimate the first moment using the simple relation

$$\mu_{1j}^{\text{theor}} \simeq \int \Psi P(\Psi|\langle\mathbf{R}\rangle_j)\, \mathrm{d}\Psi, \quad (8c)$$

where the conditional probability distribution is calculated for the average magnitude $\langle\mathbf{R}\rangle_j$ in the jth region.

The higher moments may be calculated similarly using the following relations:

$$\mu_{rj}^{\text{theor}} = \int\int \Psi^r P(\Psi, \mathbf{R})\, \mathrm{d}\Psi\, \mathrm{d}\mathbf{R} \quad (9a)$$

$$\mu_{rj}^{\text{theor}} \doteq \frac{1}{N_j} \sum_l \int \Psi^r P(\Psi|\mathbf{R}_l)\, \mathrm{d}\Psi \quad (9b)$$

$$\mu_{rj}^{\text{theor}} \simeq \int \Psi^r P(\Psi|\langle\mathbf{R}\rangle_j)\, \mathrm{d}\Psi, \quad (9c)$$

where all symbols used have the same meaning as in (8). The central moments m_{ij}^{theor}, cumulants k_{ij}^{theor} or standardized cumulants $\lambda_{ij}^{\text{theor}}$ may be calculated using simple relations.

In the case of centric seminvariants, (9c) may be rewritten for $r = 1$ as

$$\mu_{1j}^{\text{theor}} \simeq \Psi_+ P(\Psi_+|\langle R\rangle_j) + \Psi_-[1 - P(\Psi_+|\langle R\rangle_j)],$$

where $\Psi_+(\Psi_-)$ denote the seminvariant values corresponding to the positive (negative) signs of the respective product of the structure factors. If cosine seminvariants are used, then $\Psi_- = 0$ and $\Psi_+ = 1$ and

$$\mu_{1j}^{\text{theor}} \simeq P(\Psi_+|\langle R\rangle_j).$$

By analogy, relations (9a) and (9b) may also be simplified. The distribution moments of seminvariants of order $r > 1$ give no additional information for centrosymmetric structures because they may be simply derived from the first moment μ_{1j}^{theor}.

4. Distribution fitting using the distribution moments

Using the same arguments as in paper II, the empirical distribution of seminvariants for the correct set of phases of structure factors and for an increasing number of randomly selected seminvariants converges to the true distribution of seminvariants, which is assumed to be well approximated by the theoretical distribution. Thus, the correct set of phases may be looked for according to the fit between the trial and the theoretical distributions. The distribution function is unambiguously determined by its characteristic function and thus also by an infinite set of distribution moments. Furthermore, the main characteristics of unimodal distributions may be satisfactorily described by several first moments. Thus, the correct set of phases is expected to have the empirical distribution moments of low orders very close to the theoretically derived moments (Hašek, 1980).

Instead of moments, a number of simpler distribution characteristics of location, dispersion, asymmetry and excess may be used. Unlike the distribution moments, these characteristics usually do not reflect the whole profile of the distribution. Therefore, such criteria should be chosen carefully according to the expected distribution profile. Generally, they are simpler but less sensitive to the distribution profile than the moments.

4.1. *Method of successive comparison of the individual distribution characteristics*

One of the possible procedures in looking for the best fit between the trial and theoretical distributions is the successive testing of differences between the individual distribution characteristics. If the difference between the trial and theoretical values of any characteristic is larger than a defined limit, the tested set of phases is rejected. Confidence limits for the individual seminvariant types, regions of magnitudes and distribution characteristics are usually taken empirically. However, in the case where the two first

moments are taken as the only distribution characteristics, the following well known statistical procedures may be used for a rough estimate of these limits.

Suppose for simplicity that the theoretical distribution $P^{\text{theor}}(\Psi, R)$ fits exactly the true distribution of seminvariants $P^{\text{true}}(\Psi, R)$. Then a hypothesis that the true distribution and the trial distribution for the tested set of phases have the same mean if the true variance is known, and a hypothesis that they have the same variance assuming that the true mean is known may be tested (Hamilton, 1964).

(*a*) *First distribution moment.* The theoretical estimate of the true variance m_2^{theor} is known and we wish to test the hypothesis H_0: $\mu_1^{\text{trial}} = \mu_1^{\text{theor}}$ on the basis of a sample of N randomly chosen seminvariants with a sample mean μ_1^{sample} against the alternative H_1: $\mu_1^{\text{trial}} \neq \mu_1^{\text{theor}}$. The random variable

$$w = (\mu_1^{\text{sample}} - \mu_1^{\text{theor}})\sqrt{N}/m_2^{\text{theor}}$$

has a normal distribution $N(0, 1)$ and therefore the hypothesis H_0 is rejected at the $100\alpha\%$ significance level if

$$|w| > w_{\alpha/2},$$

where the 100α percentage points $w_{\alpha/2}$ of the normal distribution are given by the following table (Bickel & Doksum, 1977):

$w_{\alpha/2}$	1·96	2·58	2·81	3·29	3·89
100α	5%	1%	0·5%	0·1%	0·01%.

Thus, if the sample mean μ_1^{sample} does not lie in the interval

$$\mu_1^{\text{theor}} - m_2^{\text{theor}} w_{\alpha/2}/\sqrt{N} < \mu_1^{\text{sample}} < \mu_1^{\text{theor}} + m_2^{\text{theor}} w_{\alpha/2}/\sqrt{N},$$

the tested phase set is rejected.

(*b*) *Second central distribution moment.* Similarly, the $100\alpha\%$ confidence interval for variance may be established. Random variable

$$x^2 = (\sqrt{N-1}\, m_2^{\text{sample}}/m_2^{\text{theor}})^2$$

is distributed as x^2 with $N-1$ degrees of freedom. The hypothesis H_0: $m_2^{\text{trial}} = m_2^{\text{theor}}$ is then rejected at the $100\alpha\%$ significance level in favour of the hypothesis H_1: $m_2^{\text{trial}} \neq m_2^{\text{theor}}$ if

$$x^2 < x^2_{N-1,\alpha/2}$$

or

$$x^2 > x^2_{N-1,1-\alpha/2},$$

where percentage points $x^2_{N,\alpha}$ of the x^2 distribution defined by the relation

$$\int_0^{x^2_{N,\alpha}} \Phi(x^2)\, \mathrm{d}x^2 = 1 - \alpha$$

are extensively tabulated (*e.g.* Hamilton, 1964; Bickel & Doksum, 1977). This means that the phase set is not rejected only if

$$m_2^{\text{theor}} x^2_{N-1,\alpha/2}/\sqrt{N-1} < m_2^{\text{sample}} < m_2^{\text{theor}} x^2_{N-1,1-\alpha/2}/\sqrt{N-1}.$$

In practice, of course, the theoretical estimate of the probability distribution, especially in some regions of magnitude, does not correspond to the true distribution. Therefore, in order not to reject occasionally the correct set of phases, it is necessary to increase the range of confidence limits according to the distribution type and region. It can therefore be expected that at the end of the outlined procedure several sets of phases will not be rejected. The most probable set of phases can then be determined using some of the following criteria:

(1) according to the minimal weighted sum of squared differences between the individual characteristics of empirical and theoretical distributions (§ 4.2);

(2) using the x^2 test (see paper II, § 4);

(3) by eliminating those trial sets that have maximal values of

$$R = \min\,[(\mu_{ijk}^{\text{trial}} - \mu_{ijk}^{\text{theor}})/\mu_{ijk}^{\text{theor}}] \qquad (10)$$

taken for all regions of magnitudes (index j), all seminvariant types (index k) and all the tested moment orders i.

4.2. *The global coefficient of the distribution fitting using moments*

In the last paragraph, some criteria were described which enable us to delete the sets of phases, the characteristics of which significantly deviate from the theoretical ones. However, a unique criterion which evaluates the fit of all distributions using the moments of all seminvariant types by only one coefficient is more useful. A number of different norms may be proposed to describe differences between the two characteristic functions. One of them is the coefficient (Hašek, 1975)

$$M = \left(\sum_i \sum_j \sum_k w_{ijk}\right)^{-1} \sum_i \sum_j \sum_k w_{ijk} (\mu_{ijk}^{\text{trial}} - \mu_{ijk}^{\text{theor}})^2, \qquad (11)$$

where k denotes different types of seminvariants, j the region of magnitudes and i the order of the respective moment. The weights w_{ijk} generally depend on the importance and reliability of moments of different orders, on the number of seminvariants in the corresponding regions and on the type of distribution of seminvariants used. For weights known only on a relative scale, the coefficient M is normalized by dividing by the sum of all the weights used. The minimal value of the coefficient M denotes the set of phases which is expected to be the correct one with

the highest probability. Principal information on both distributions is contained in the low-order moments. Furthermore, the uncertainty in the determination of moments increases with their order and decreases with the number of seminvariants used in the calculation. Therefore, the weights should strongly reduce the influence of the moments of higher orders depending on the index i. The weights should be smaller, the smaller the number of seminvariants used for the calculation of μ^{emp} and the more restrictive the approximations used in the calculation of the corresponding theoretical distributions. The decrease in the weights* with the order of moments might be approximately expressed by the coefficient $(n!)^{-1}$.

Special seminvariants

In the case of special seminvariants, which owing to the crystallographic symmetry may assume only two values, the distributions are fully described only by their first moments. Hence, the summation over index i in (11) is omitted and the distribution-fitting coefficient is

$$M = \sum_j \sum_k w_{jk}(\mu_{1jk}^{trial} - \mu_{1jk}^{theor})^2. \quad (12)$$

If centric cosine seminvariants are used, then $\mu_{1jk}^{emp} = P_{1jk}^{emp}$ [compare with equation (7) of paper II] and the distribution-fitting coefficient M for one type of seminvariant may be written in the form equivalent to equation (19) in paper II:

$$N = \sum_j w_j(P_{+j}^{trial} - P_{+j}^{theor})^2. \quad (13)$$

* The weights should be properly modified when cumulants, standardized cumulants or other types of distribution characteristics are used.

References

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

DE TITTA, G. T., EDMONDS, J. W., LANGS, D. A. & HAUPTMAN, H. (1975). *Acta Cryst.* A**31**, 472–479.

HAMILTON, W. C. (1964). *Statistics in Physical Science.* New York: Ronald Press.

HAŠEK, J. (1975). *Acta Cryst.* A**31**, 818–819.

HAŠEK, J. (1980). In *Proceedings of the Symposium on Special Topics of X-ray Crystal Structure Analysis*, pp. 108–111. Zentralinstitut für Physikalische Chemie AW, German Democratic Republic.

HAŠEK, J. (1984*a*). *Acta Cryst.* A**40**, 338–340.

HAŠEK, J. (1984*b*). *Acta Cryst.* A**40**, 340–346.

SCHENK, H. (1974). *Acta Cryst.* A**30**, 477–481.

Acta Cryst. (1984). A**40**, 350–352

On the Solution of the Phase Problem. IV.* Distributions Fitted using the Kolmogorov Test

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 a correct set of phases is based on a comparison between the trial and theoretical distributions of seminvariants using the Kolmogorov test. If the Kolmogorov test is restricted to a single region of magnitudes where only a small variance around the mean seminvariant value is expected, then the test is reduced to a simple rule. *The smaller the number of seminvariants differing significantly from the expected mean value, the more probable the set of phases.* In this simple form the Kolmogorov test has been used since the very beginnings of direct methods. In spite of the fact that the method seems to be less efficient than the distribution fitting using the x^2 test [Hašek (1984). *Acta Cryst.* A**40**, 340–346], its simplicity and low claim on computing time enables one to survey a large number of trial sets and so to increase the power of the method based on a combination of the Kolmogorov test with the x^2 test, or with the low-order distribution moment test.

1. Introduction

In direct methods, *a priori* information on the structure necessary for the phase-problem solution is usually represented by 'probability relations' between the structure factors, *i.e.* by the function form of the probability distributions of seminvariants. Of course, some methods extract only information on the most probable seminvariant values and do not account for the fact that the probability distribution defines also seminvariants which *must* greatly differ from their 'ideal' value. This results in occasional failures of

* Part III: Hašek (1984*c*).

0108-7673/84/040350-03$01.50