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

such procedures to find a correct solution unambiguously.

A general method of phase-problem solution which makes possible full utilization of *a priori* structure information hidden in the probability distributions of seminvariants has been reported in previous papers (Hašek, 1984*a*, *b*, *c*) denoted hereafter as papers I, II and III. The criteria of correctness of the phase set described therein are based on a comparison between the empirical probability distributions calculated for the trial set of phases and the corresponding theoretical probability distributions (paper II) or on a comparison of the respective low-order distribution moments (paper III). The third method for distribution fitting, based on the Kolmogorov test, is outlined in this paper.

2. Cumulative probability distributions of seminvariants

In a manner analogous to § 2 of paper II, let the parameter space of the probability distribution $P(\psi|R_1,\ldots,R_m)$ be divided into regions of magnitudes $R_1,\ldots,R_m$ and intervals of seminvariant values ψ. N_{ijk} denotes the number of seminvariants of the kth type belonging to the jth region and the ith interval; N_{jk} is the sum of N_{ijk} overall r intervals of ψ values, $N_{jk}=\sum_{i=1}^{r} N_{ijk}$. Then the values

$$\mathscr{P}_{ujk}^{\text{emp}}=\sum_{i=1}^{u} N_{ijk}/N_{jk} \tag{1}$$

converge for $N_{jk}\to\infty$ to the empirical cumulative probability distribution of seminvariants. For the correct phases and magnitudes, the empirical probability distribution is equal to the true cumulative probability distribution of seminvariants and, therefore, the $\mathscr{P}_{ujk}^{\text{emp}}$ values are unbiased and consistent estimates of the true cumulative probability distribution $\mathscr{P}^{\text{true}}$ (Bickel & Doksum, 1977). The trial cumulative distribution of seminvariants for any wrong set of phases cannot, in a statistical sense, fit the true cumulative distribution of seminvariants better than the empirical one. Thus, the fit between the trial cumulative distributions and the corresponding theoretical estimate of the true cumulative probability distribution may be used as a criterion of the correctness of the trial phase set.

According to the definition given above, $\mathscr{P}_{rjk}^{\text{emp}}$ for the last interval of seminvariant values is just unity. In the case of centrosymmetric structures the seminvariants may assume only two values and therefore the distribution may be estimated by computing

$$\mathscr{P}_{1jk}^{\text{emp}}=N_{1jk}/N_{jk}$$

values, converging for increasing number of seminvariants, and the correct set of phases to the true probability distribution of a positive sign of the respective product of the structure factors [compare with equation (7) in paper II].

Suppose that the theoretically derived distribution of seminvariants is identical with the true distribution. If the size of intervals and regions tends to zero and the number of seminvariants in them tends to infinity, the $\mathscr{P}_{ujk}^{\text{emp}}$ values would correspond exactly to those of the theoretical cumulative probability distribution. In practice, of course, the intervals and regions must be chosen large enough to contain sufficiently high numbers of seminvariants. Therefore, the $\mathscr{P}_{ujk}^{\text{theor}}$ values corresponding to the $\mathscr{P}_{ujk}^{\text{emp}}$ values are computed using the relations (*a*), (*b*) or (*c*).

(*a*)

$$\mathscr{P}_{ujk}^{\text{theor}}=V_{jk}^{-1}\int P_k(\psi|R_1,\ldots,R_m)\,\mathrm{d}R_1\ldots\mathrm{d}R_m\,\mathrm{d}\psi, \tag{2}$$

where the integration runs over all the first u intervals of ψ values, and over the jth region of magnitudes. The normalizing constant V_{jk} is

$$V_{jk}=\int P_k(\psi|R_1,\ldots,R_m)\,\mathrm{d}R_1\ldots\mathrm{d}R_m\,\mathrm{d}\psi,$$

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

(*b*)

$$\mathscr{P}_{ujk}^{\text{theor}}\doteq\mathscr{N}_{ujk}^{-1}\sum_{l} P_k(\psi|R_{1l},\ldots,R_{ml}), \tag{3}$$

where the summation runs over all $\mathscr{N}_{ujk}=\sum_{i=1}^{u} N_{ijk}$ seminvariants in the first u intervals contained in the jth region of magnitudes of the kth probability distribution.

(*c*)

$$\mathscr{P}_{ujk}^{\text{theor}}=\sum_{i=1}^{u} Q_{ijk}^{\text{theor}}, \tag{4}$$

where the relative frequencies Q_{ijk}^{theor} are given by equation (9) of paper II.

The $\mathscr{P}_{rjk}^{\text{theor}}$ value for the last interval of ψ values is equal to unity. Therefore, in the case of centrosymmetric structures, where only two intervals of ψ values are possible, the distribution in a single region of magnitudes is described by only one value $\mathscr{P}_{1jk}^{\text{theor}}$.

3. Kolmogorov test

By the Glivenko–Cantelli theorem,

$$D_N=\sup_{\psi_0}|\mathscr{P}_{\psi_0}^{\text{emp}}-\mathscr{P}_{\psi_0}^{\text{true}}| \tag{5}$$

converges for increasing sample size, $N\to\infty$, to zero in probability (Fisz, 1963). The Kolmogorov test based on this property enables us to reject the hypothesis that $\mathscr{P}^{\text{trial}}=\mathscr{P}^{\text{true}}$ in favour of the hypothesis $\mathscr{P}^{\text{trial}}\neq\mathscr{P}^{\text{true}}$ for large values of D_N.

Considering that the numbers of seminvariants in the individual regions of magnitudes N_{jk} remain the same for all the tested sets of phases, it is convenient in practice to replace the cumulative probabilities by

the cumulative numbers of seminvariants:

$$\mathcal{N}_{ijk} = N_{jk}\mathcal{P}_{ijk}^{\text{trial}} \quad (6)$$

$$\mathcal{N}_{ijk}^{\text{theor}} = N_{jk}\mathcal{P}_{ijk}^{\text{theor}}. \quad (7)$$

Suppose now that the theoretical distribution corresponds exactly to the true distribution of seminvariants, then a suitable random variable to test is

$$D_{jk} = \sup_i |\mathcal{N}_{ijk}^{\text{theor}} - \mathcal{N}_{ijk}|. \quad (8)$$

Maximal admissible deviations, $D_{jk}^{\text{crit}} = |\mathcal{N}_{ijk}^{\text{theor}} - \mathcal{N}_{ijk}|_{\max}$, at significance level 0·01, are dependent on the number of seminvariants N_{jk} in the tested region; these are given in Table 1.

The $|\mathcal{N}_{rjk}^{\text{theor}} - \mathcal{N}_{rjk}|$ values for the last interval of ψ values are identically zero because $\mathcal{N}_{rjk}^{\text{theor}} = \mathcal{N}_{rjk} = N_{jk}$, where N_{jk} is the number of seminvariants in the jth region of the kth distribution. Therefore, only $(r-1)$ first intervals of ψ values are tested for maximal differences between cumulative distributions. If in some region and interval the theoretical distribution does not correspond to the true distribution, then the critical values taken from Table 1 must be increased.*

In the case of special seminvariants there are only two possible seminvariant values and also only two intervals. Thus, the number of seminvariants in the second interval is uniquely determined by the number of seminvariants in the first interval and in this case only one difference $D_{jk} = |N_{1jk}^{\text{theor}} - N_{1jk}|$ is tested against Table 1 for one region of magnitudes. For centrosymmetric structures the efficiency of the Kolmgorov test based on testing

$$D_{jk} = |P_{+jk}^{\text{theor}} - P_{+jk}^{\text{trial}}|_{\max}$$

may be compared with that of the coefficient

$$K = \sum w_{ijk}(P_{+jk}^{\text{theor}} - P_{+jk}^{\text{trial}})^2 / \sum w_{ijk}$$

used in the x^2 test [equation (19) in paper II] under a restrictive condition that the only contribution to the summation is a maximal difference, *i.e.* $K' = (P_{+jk}^{\text{theor}} - P_{+jk}^{\text{trial}})_{\max}^2$. Thus, the Kolmogorov test is expected to have worse discriminative abilities than the x^2 test, but its convenience may be seen in its simplicity.

Table 1. *Critical values* $D_{jk}^{crit} = |\mathcal{N}_{ijk}^{theor} - \mathcal{N}_{ijk}|_{max}$ *of the Kolmogorov–Smirnov test at the* 0·01 *significance level*

N_{jk}	2	6	10	15	20	40	>40
D_{jk}^{crit}	2	4	5	6	7	10	$1{\cdot}63\sqrt{N_{jk}}$

The author would like to thank Professor H. Schenk for helpful discussions and support of this work.

References

BICKEL, P. J. & DOKSUM, K. A. (1977). *Mathematical Statistics: Basic Ideas and Selected Topics.* San Francisco: Holden-Day.
FISZ, M. (1963). *Probability Theory and Mathematical Statistics.* New York: Wiley.
HAŠEK, J. (1984*a*). *Acta Cryst.* A**40**, 338–340.
HAŠEK, J. (1984*b*). *Acta Cryst.* A**40**, 340–346.
HAŠEK, J. (1984*c*). *Acta Cryst.* A**40**, 346–350.

* The optimal values D_{jk}^{crit} have to be selected by experience for each type of theoretical distribution.

Acta Cryst. (1984). A**40**, 352–355

Diffraction Scattering at Angles Far From the Bragg Angle and the Structure of Thin Subsurface Layers

BY A. M. AFANAS'EV,* P. A. ALEKSANDROV,* R. M. IMAMOV, A. A. LOMOV AND A. A. ZAVYALOVA

Institute of Crystallography of the USSR Academy of Sciences, Leninsky prospect 59, *Moscow* 117333, *USSR*

(*Received* 3 *June* 1983; *accepted* 13 *January* 1984)

Abstract

It has been shown that by measuring the angular dependence of X-ray diffraction scattering far from the Bragg peak, information on the structure perfection of thin subsurface layers can be directly obtained. This is associated with the fact that the waves generated in the crystal bulk compensate one another, and the intensity of rocking-curve tails is due mainly to scattering in the subsurface layer. The typical thickness of a scattering layer is related to the deviation angle by a simple relationship: $\Delta z \simeq L_{ex}\omega_0/\alpha$, where α is the deviation angle of the specimen from the exact Bragg position, ω_0 the diffraction maximum width, and L_{ex} the extinction length. The method of three-crystal diffractometry permitted the observation for the first time with a conventional X-ray source of a distorted layer with a thickness of ~10 nm.

* Present address: Kurchatov Institute of Atomic Energy, Kurchatov Square 46, Moscow 123182, USSR

0108-7673/84/040352-04$01.50