Acta Cryst. (1979). A**35**, 351–359

A New Statistical Model for Describing Errors in Isomorphous Replacement Data: The Case of One Derivative

BY EDWARD A. GREEN

Medical Foundation of Buffalo, Inc., 73 High Street, Buffalo, New York 14203, *USA*

(*Received* 5 *May* 1978; *accepted* 10 *October* 1978)

Abstract

A probability method for phasing macromolecular isomorphous replacement data from the viewpoint of errors arising in the isomorphous replacement process is considered. The assumption of imperfect isomorphism between the atomic positions in the native crystal and the atomic positions in the native component of the derivative crystal leads to estimates of phase dependent on the value of $\sin\theta/\lambda$. The mathematical techniques used are similar to those employed in deriving probability distributions of structure seminvariants. Some of the formulas, which apply only to the case of one derivative, are compared with earlier results.

1. Introduction

In deriving the classical probability methods for macromolecular phase determination (Blow & Crick, 1959; Rossmann & Blow, 1961; Hendrickson & Lattman, 1970), based on the central limit theorem, it has been assumed that the combined errors from different sources have a Gaussian behavior. Other models describing the nature of possible errors are, however, at least as plausible, *a priori*, as this. Employing techniques similar to those used in deriving probability formulas of structure seminvariants (Karle & Hauptman, 1958), it is possible to identify the main sources of error in the isomorphous replacement method and to predict the consequences of the assumed models. Presumably, the more rigorous account of the effects of these errors, as described here, will yield more accurate phase estimates *via* the isomorphous-replacement technique.

This paper is concerned with errors arising from single isomorphous replacement, SIR, only, and with the probability distributions associated with SIR random variables. It examines some of the different types of errors prevalent in this method and shows how these errors manifest themselves in both joint and conditional probability distributions of the native structure factor or phase.

It is assumed that the native crystal and the native component of the single-isomorphous-derivative crystal contain N atoms, not necessarily identical, in the unit-cell. The heavy-atom component of the derivative crystal contains M dissimilar atoms, which may be viewed as one or more different kinds of atoms occupying several partial sites and having in general M different thermal parameters. Structure factors obey either the acentric distribution (when complex) or the centric distribution (when real or pure imaginary).

Denote by $D_{\mathbf{h}}$, $F_{\mathbf{h}}$, $H_{\mathbf{h}}$ the respective derivative, native, and heavy-atom structure factors defined by

$$D_{\mathbf{h}} = |D_{\mathbf{h}}| \exp(i\psi_{\mathbf{h}}) = \sum_{j=1}^{N} f_j(\mathbf{h}) \exp(2\pi i \mathbf{h}.\mathbf{r}_j) + \sum_{k=1}^{M} g_k(\mathbf{h}) \exp(2\pi i \mathbf{h}.\mathbf{r}'_k), \quad (1.1)$$

$$F_{\mathbf{h}} = |F_{\mathbf{h}}| \exp(i\varphi_{\mathbf{h}}) = \sum_{j=1}^{N} f_j(\mathbf{h}) \exp(2\pi i \mathbf{h}.\mathbf{r}_j), \quad (1.2)$$

$$H_{\mathbf{h}} = |H_{\mathbf{h}}| \exp(i\omega_{\mathbf{h}}) = \sum_{k=1}^{M} g_k(\mathbf{h}) \exp(2\pi i \mathbf{h}.\mathbf{r}'_k), \quad (1.3)$$

where $|D_{\mathbf{h}}|$, $|F_{\mathbf{h}}|$, $|H_{\mathbf{h}}|$, $\psi_{\mathbf{h}}$, $\varphi_{\mathbf{h}}$, $\omega_{\mathbf{h}}$ are the magnitudes and phases for each reciprocal-lattice vector $\mathbf{h}$. The atomic scattering factor and position vector of the atom labeled j in the native or native component of the derivative crystal are denoted by $f_j(\mathbf{h})$ and $\mathbf{r}_j$ respectively; and by $g_k(\mathbf{h})$, $\mathbf{r}'_k$ in the heavy-atom and derivative structure factors. The corresponding definitions for the structure factors in $P\bar{1}$ are obvious. In view of (1.1)–(1.3) it is further assumed that the three structure factors $D_{\mathbf{h}}$, $F_{\mathbf{h}}$, $H_{\mathbf{h}}$ are on an absolute scale.

Since (1.1)–(1.3) are applicable to a perfect SIR experiment, an unlikely situation, a more realistic set of structure factor equations is needed before random variables can be attached to these quantities, as shown next.

0567-7394/79/030351-09$01.00

2. Errors in the SIR experiment

2.1 *Imperfect isomorphism between the native structure and the native component of the derivative*

In this subsection, it is assumed that the native and calculated heavy-atom structure factors are defined by (1.2) and (1.3). A less than perfect isomorphism between the atomic positions in the native and the native part of the derivative crystal is expressed as a shift $\boldsymbol{\delta}_j$ of each position vector, $\mathbf{r}_j$, in the native component of the derivative, *i.e.*

$$\mathbf{r}_j\,(\text{derivative}) = \mathbf{r}_j\,(\text{native}) + \boldsymbol{\delta}_j. \tag{2.1}$$

Since the *a priori* value of each $\boldsymbol{\delta}_j$ is unknown, it is plausible to suppose that $\boldsymbol{\delta}_j$ is a random vector which follows a normal distribution; more precisely, that $\boldsymbol{\delta}_j$ is normally distributed with mean vector $\mathbf{O}$ and isotropic variance σ_{Dj}^2, *i.e.*

$$\boldsymbol{\delta}_j \simeq N(\mathbf{O},\sigma_{Dj}^2). \tag{2.2}$$

This assumption corresponds to the displacement of one or more atoms in the native crystal due to the infusion of the heavy-atom component. In a SIR experiment containing errors due only to this type of imperfect isomorphism, the set of structure factors, $D_\mathbf{h}$, $F_\mathbf{h}$, $H_\mathbf{h}$, for the non-centrosymmetric case are defined by

$$|D_\mathbf{h}|\exp(i\psi_\mathbf{h}) = \sum_{j=1}^{N} f_j(\mathbf{h})\exp[2\pi i\mathbf{h}.(\mathbf{r}_j+\boldsymbol{\delta}_j)] + \sum_{k=1}^{M} g_k(\mathbf{h})\exp(2\pi i\mathbf{h}.\mathbf{r}'_k), \tag{2.3}$$

$$|F_\mathbf{h}|\exp(i\varphi_\mathbf{h}) = \sum_{j=1}^{N} f_j(\mathbf{h})\exp(2\pi i\mathbf{h}.\mathbf{r}_j), \tag{2.4}$$

$$|H_\mathbf{h}|\exp(i\omega_\mathbf{h}) = \sum_{k=1}^{M} g_k(\mathbf{h})\exp(2\pi i\mathbf{h}.\mathbf{r}'_k). \tag{2.5}$$

2.2 *Imperfect isomorphism plus errors in the calculated heavy-atom component positions*

The assumption in the preceding subsection that the positions of the heavy atoms in the derivative crystal are known is here replaced by the weaker assumption that an estimate for each heavy-atom position and a corresponding variance σ_{Hk}^2, are known. Thus, as before, M random vectors $\boldsymbol{\varepsilon}_k$ are introduced and defined by

$$\boldsymbol{\varepsilon}_k = \mathbf{r}'_k\,(\text{derivative}) - \mathbf{r}'_k\,(\text{estimated}), \tag{2.6}$$

where

$$\boldsymbol{\varepsilon}_k \simeq N(\mathbf{O},\sigma_{Hk}^2). \tag{2.7}$$

The structure factor equations describing this SIR model are now

$$|D_\mathbf{h}|\exp(i\psi_\mathbf{h}) = \sum_{j=1}^{N} f_j(\mathbf{h})\exp[2\pi i\mathbf{h}.(\mathbf{r}_j+\boldsymbol{\delta}_j)] + \sum_{k=1}^{M} g_k(\mathbf{h})\exp(2\pi i\mathbf{h}.\mathbf{r}'_k), \tag{2.8}$$

$$|F_\mathbf{h}|\exp(i\varphi_\mathbf{h}) = \sum_{j=1}^{N} f_j(\mathbf{h})\exp(2\pi i\mathbf{h}.\mathbf{r}_j), \tag{2.9}$$

$$|H_\mathbf{h}|\exp(i\omega_\mathbf{h}) = \sum_{k=1}^{M} g_k(\mathbf{h})\exp 2\pi i\mathbf{h}.(\mathbf{r}'_k+\boldsymbol{\varepsilon}'_k). \tag{2.10}$$

The term $\boldsymbol{\varepsilon}_k$ in (2.10) corresponds to the discrepancy between the true position, $\mathbf{r}'_k$, of the atom labeled k in the derivative crystal and its estimated position, $\mathbf{r}'_k + \boldsymbol{\varepsilon}_k$, in the calculated heavy-atom crystal.

2.3 *Errors arising from non-isomorphous sources*

The two sources of error considered thus far are probably the major ones associated with imperfect isomorphism. This does not imply they are the only major errors in the SIR process. The most important source of error unrelated to isomorphous replacement is experimental inaccuracy in the estimation of the magnitudes $|D_\mathbf{h}|$ and $|F_\mathbf{h}|$. The effects of experimental errors can be incorporated into the non-centrosymmetric structure factor model by adding the complex error functions, χ_D, χ_F, where

$$\chi_D = |\chi_D|\exp(i\xi_D), \tag{2.11}$$

$$\chi_F = |\chi_F|\exp(i\xi_F), \tag{2.12}$$

to (2.8), (2.9) respectively. Note that each has magnitude and phase components. It is assumed that each pair $(|\chi_\alpha|,\xi_\alpha)$ has the joint probability distribution defined by

$$P(|\chi_\alpha|,\xi_\alpha) = \frac{|\chi|}{\pi\mu_\alpha}\exp\left(-\frac{|\chi_\alpha|^2}{\mu_\alpha}\right), \tag{2.13}$$

i.e. that all values of ξ_α are equally probable, where α is either D or F. The quantity μ_α in (2.13) is the variance in those errors unrelated to isomorphous replacement. Hence, the complete non-centrosymmetric structure factor model for single isomorphous replacement is given by the three structure factors $D_\mathbf{h}$, $F_\mathbf{h}$, $H_\mathbf{h}$ defined by

$$|D_\mathbf{h}|\exp(i\psi_\mathbf{h}) = \sum_{j=1}^{N} f_j(\mathbf{h})\exp[2\pi i\mathbf{h}.(\mathbf{r}_j+\boldsymbol{\delta}_j)] + \sum_{k=1}^{M} g_k(\mathbf{h})\exp(2\pi i\mathbf{h}.\mathbf{r}'_k) + |\chi_D|\exp(i\xi_D), \tag{2.14}$$

$$|F_{\mathbf{h}}| \exp(i\varphi_{\mathbf{h}}) = \sum_{j=1}^{N} f_j(\mathbf{h}) \exp(2\pi i \mathbf{h}.\mathbf{r}_j) + |\chi_F| \exp(i\xi_F), \quad (2.15)$$

$$|H_{\mathbf{h}}| \exp(i\omega_{\mathbf{h}} = \sum_{k=1}^{M} g_k(\mathbf{h}) \exp[2\pi i \mathbf{h}.(\mathbf{r}'_k + \boldsymbol{\varepsilon}_k)], \quad (2.16)$$

where $\boldsymbol{\delta}_j$, $\boldsymbol{\varepsilon}_k$ are specified by (2.2), (2.7) respectively. For structure factors obeying the centric distribution, the error variables χ'_D, χ'_F are real and normally distributed with variance E^2, *i.e.*

$$\chi'_\alpha \simeq N(\mathbf{O}, E^2_\alpha), \quad (2.17)$$

where α is either D or F. Hence, for the centrosymmetric zones

$$D_{\mathbf{h}} = 2 \sum_{j=1}^{N/2} f_j(\mathbf{h}) \cos[2\pi \mathbf{h}.(\mathbf{r}_j + \boldsymbol{\delta}_j)] + 2 \sum_{k=1}^{M/2} g_k(\mathbf{h}) \cos(2\pi \mathbf{h}.\mathbf{r}'_k) + \chi'_D, \quad (2.18)$$

$$F_{\mathbf{h}} = 2 \sum_{j=1}^{N/2} f_j(\mathbf{h}) \cos(2\pi \mathbf{h}.\mathbf{r}_j) + \chi'_F, \quad (2.19)$$

$$H_{\mathbf{h}} = 2 \sum_{k=1}^{M/2} g_k(\mathbf{h}) \cos[2\pi \mathbf{h}.(\mathbf{r}'_k + \boldsymbol{\varepsilon}_k)]. \quad (2.20)$$

2.4 *A structure factor model analogous to the formulation by Blow & Crick*

If, in the SIR experiment, errors from all sources are combined and described by the single complex random variable χ_D then,

$$|D_{\mathbf{h}}| \exp(i\psi_{\mathbf{h}}) = \sum_{j=1}^{N} f_j(\mathbf{h}) \exp(2\pi i \mathbf{h}.\mathbf{r}_j) + \sum_{k=1}^{M} g_k(\mathbf{h}) \exp(2\pi i \mathbf{h}.\mathbf{r}'_k) + |\chi_D| \exp(i\xi_D), \quad (2.21)$$

$$|F_{\mathbf{h}}| \exp(i\varphi_{\mathbf{h}}) = \sum_{j=1}^{N} f_j(\mathbf{h}) \exp(2\pi i \mathbf{h}.\mathbf{r}_j), \quad (2.22)$$

$$|H_{\mathbf{h}}| \exp(i\omega_{\mathbf{h}}) = \sum_{k=1}^{M} g_k(\mathbf{h}) \exp(2\pi i \mathbf{h}.\mathbf{r}'_k). \quad (2.23)$$

This set of structure factors defines the 'ensemble error' model. The primary difference between it and that of Blow & Crick is that, in the latter system, all errors, presumed Gaussian, are assumed to reside in only the magnitude of the derivative structure factor alone, even in $P1$; hence the error is always real. In (2.21) on the other hand, the error term is complex and all values of the phase ξ_D are equally probable.

Equations (2.14)–(2.16), and (2.18)–(2.20) form the basis from which is derived the conditional probability distribution of the phase $\varphi_{\mathbf{h}}$, given the three magnitudes $|D_{\mathbf{h}}|$, $|F_{\mathbf{h}}|$, $|H_{\mathbf{h}}|$ and phase $\omega_{\mathbf{h}}$, and the joint conditional probability distribution of the pair $(|F_{\mathbf{h}}|, \varphi_{\mathbf{h}})$, given the magnitudes $|D_{\mathbf{h}}|$, $|H_{\mathbf{h}}|$ and phase $\omega_{\mathbf{h}}$, as shown in the sequel.

3. The conditional probability distribution of the native phase $\varphi_{\mathbf{h}}$ given the three magnitudes $|D_{\mathbf{h}}|$, $|F_{\mathbf{h}}|$, $|H_{\mathbf{h}}|$, and phase $\omega_{\mathbf{h}}$

Assume the values of the three magnitudes $|D_{\mathbf{h}}|$, $|F_{\mathbf{h}}|$, $|H_{\mathbf{h}}|$ and phase $\omega_{\mathbf{h}}$ are known, and the value of the singe phase $\varphi_{\mathbf{h}}$ is unknown. Denote by

$$P_{1|3,1} = P(\Phi | D, F, H, \omega) \quad (3.1)$$

the conditional probability distribution of $\varphi_{\mathbf{h}}$, given the three magnitudes $|D_{\mathbf{h}}|$, $|F_{\mathbf{h}}|$, $|H_{\mathbf{h}}|$ and phase $\omega_{\mathbf{h}}$. Then, in view of the joint probability distribution, P, of the three complex structure factors, D, F, H, Appendix I, (I.2),

$$P_{1|3,1} = \int_0^{2\pi} P \, d\Psi \Big/ \int_0^{2\pi} P \, d\Psi \, d\Phi \quad (3.2)$$

$$= 1/K \exp \times I_0 \left\{ \frac{2D}{\Sigma} [s_1^2 \Gamma^2 F^2 + (s_2 + \mu_F)^2 \Upsilon^2 H^2 + 2s_1(s_2 + \mu_F) \Upsilon \Gamma F H \cos(\Phi - \omega)]^{1/2} \right\}, \quad (3.3)$$

where,

$$\Sigma = [(s_2 + s_1 + \mu_D)(s_2 + \mu_F) s_1 - s_1 \Gamma^2 - (s_2 + \mu_F) \Upsilon^2], \quad (3.4)$$

$$s_1 = \sum_{k=1}^{M} g_k^2(\mathbf{h}), \quad (3.5)$$

$$s_2 = \sum_{j=1}^{N} f_j^2(\mathbf{h}), \quad (3.6)$$

$$\Upsilon = \sum_{k=1}^{M} g_k^2(\mathbf{h}) \exp(-8\pi^2 \sigma_{Hk}^2 \sin^2 \theta/\lambda^2), \quad (3.7)$$

$$\Gamma = \sum_{j=1}^{N} f_j^2(\mathbf{h}) \exp(-8\pi^2 \sigma_{Dj}^2 \sin^2 \theta/\lambda^2), \quad (3.8)$$

$$\mu_F = \epsilon(\chi_F^2), \quad (3.9)$$

$$\mu_D = \epsilon(\chi_D^2), \quad (3.10)$$

and where ϵ denotes the expected value; K is the normalization factor, and I_0 is the zero-order modified Bessel-function. For the centrosymmetric zones, the appropriate conditional probability distribution

$P^{+}_{1|3,1}(P^{-}_{1|3,1})$ = the conditional probability that

$$\Phi = 0(\pi), \text{ given } D, F, H, \omega, \tag{3.11}$$

is obtained from equation (I.3) by similar methods,

$$P^{+}_{1|3,1} = \frac{1}{K}\exp\left\{\frac{\pm Y\Gamma FH\cos\omega}{\bar{\Sigma}}\right\} \times \cosh\left\{\frac{D}{\bar{\Sigma}}[s_1^2\Gamma^2F^2 + (s_2+E_F^2)^2Y^2H^2 \pm 2s_1(s_2+E_F^2)Y\Gamma FH\cos\omega]^{1/2}\right\}, \tag{3.12}$$

and the known phase ω is 0 or π. The parameter $\bar{\Sigma}$ is defined by

$$\bar{\Sigma} = [(s_2+s_1+E_D^2)(s_2+E_F^2)s_1 - s_1\Gamma^2 - (s_2+E_F^2)Y^2], \tag{3.13}$$

the parameters s_1, s_2, Y, Γ are defined by (3.5)–(3.8) respectively, and

$$E_D^2 = \epsilon(\chi_D'^2) \tag{3.14}$$

$$E_F^2 = \epsilon(\chi_F'^2). \tag{3.15}$$

It is informative to consider two special cases for the distribution given in (3.3).

3.1 *Imperfect isomorphism only*

Assume the value of σ_{Hk}^2 is zero, *i.e.* the heavy-atom positions are known with perfect accuracy, then (3.5) and (3.7) imply

$$Y = s_1. \tag{3.16}$$

Assume also that the values of μ_F and μ_D [(3.9), (3.10)] are zero, *i.e.* that the magnitudes $|F_{\mathbf{h}}|$ and $|D_{\mathbf{h}}|$ are known with perfect accuracy. Then (3.3) becomes

$$P_{1|3,1} = \frac{1}{K}\exp\left\{\frac{-2\Gamma FH\cos(\Phi-\omega)}{(s_2^2-\Gamma^2)}\right\} \times I_0\left\{\frac{2D}{(s_2^2-\Gamma^2)}[\Gamma^2F^2 + s_2^2H^2 + 2s_2\Gamma FH\cos(\Phi-\omega)]^{1/2}\right\}, \tag{3.17}$$

where K is the normalization factor. This distribution belongs to the model in which errors arise from imperfect isomorphism alone (§ 2.1). The explicit definition of Γ, (3.8), clearly shows the $\sin\theta/\lambda$ dependence in the reliability of the phase estimated from $P_{1|3,1}$. As the value of $\sin\theta/\lambda$ increases, the value of Γ decreases faster than s_2, which in turn yields increasingly larger values for

$$s_2^2 - \Gamma^2. \tag{3.18}$$

In short the variance of (3.17) increases with increasing $\sin\theta/\lambda$ so that the distribution $P_{1|3,1}$ yields less reliable estimates of phase for larger values of $\sin\theta/\lambda$.

3.2 *Ensemble error model*

For this special case, the values for σ_{Hk}^2 and σ_{Dj}^2 are assumed to be zero which, in view of (3.6) and (3.8), implies

$$\Gamma = s_2, \tag{3.19}$$

as well as (3.16). Assume also that the value of μ_F is zero. Then (3.3) becomes

$$P_{1|3,1} = \frac{1}{K'}\exp\left\{\frac{-2FH\cos(\Phi-\omega)}{\mu_D}\right\} \times I_0\left\{\frac{2D}{\mu_D}[F^2+H^2+2FH\cos(\Phi-\omega)]^{1/2}\right\}, \tag{3.20}$$

where K' is the normalization factor. The distribution belongs to the ensemble error model of §2.4. From the definition of μ_D, (2.13), it is clear that any estimate of $\varphi_{\mathbf{h}}$ from (3.20) is independent of the value of $\sin\theta/\lambda$. Thus, in this case, (3.20) implies that there is no relationship between the value of $\sin\theta/\lambda$ and the reliability of the estimate $\varphi_{\mathbf{h}}$. Depending on the values of the parameters involved, estimates of phase could be more reliable at high values of $\sin\theta/\lambda$ than at lower ones. Also, (3.20) shows that the reliability of the estimate of $\varphi_{\mathbf{h}}$ decreases with increasing μ_D.

3.3 *The conditional expected value of* $\exp(i\varphi_{\mathbf{h}})$ *given* $|D_{\mathbf{h}}|, |F_{\mathbf{h}}|, |H_{\mathbf{h}}|, \omega_{\mathbf{h}}$

Denote by $\epsilon[\exp(i\Phi)|D,F,H,\omega]$ the conditional expected value of $\exp(i\varphi_{\mathbf{h}})$, given the three magnitudes $|D_{\mathbf{h}}|, |F_{\mathbf{h}}|, |H_{\mathbf{h}}|$ and phase $\omega_{\mathbf{h}}$. Then,

$$\epsilon[\exp(i\Phi)|D,F,H,\omega] = \int_0^{2\pi}\exp(i\Phi)\,P_{1|3,1}\,d\Phi, \tag{3.21}$$

$$= X\exp(i\omega), \tag{3.22}$$

where X is the figure of merit and ω the best phase. Although the integral is best evaluated by numerical methods, it has an analytical representation given by (3.22) where

$$X = -\frac{\sum_{m=-\infty}^{\infty}(-1)^m I_m(A)I_m(B)I_{m-1}(C)}{\sum_{m=-\infty}^{\infty}(-1)^m I_m(A)I_m(B)I_m(C)}, \tag{3.23}$$

$$A = \frac{2s_1\Gamma DF}{\Sigma}, \tag{3.24}$$

$$B = \frac{2(s_2 + \mu_F)\Upsilon DH}{\Sigma}, \tag{3.25}$$

$$C = \frac{2\Gamma\Upsilon FH}{\Sigma}. \tag{3.26}$$

The modified Bessel functions are of order m. The parameters Σ, s_1, s_2, Υ, Γ, μ_F defined by (3.4)–(3.9) are assumed known. The expected value of $\exp(i\varphi_{\mathbf{h}})$ corresponding to the two special cases of imperfect isomorphism and ensemble error follow from the substitutions discussed in § 3.1 and § 3.2 respectively.

4. The conditional probability distribution of the native structure factor $F_{\mathbf{h}}$ given the two magnitudes $|D_{\mathbf{h}}|$, $|H_{\mathbf{h}}|$ and phase $\omega_{\mathbf{h}}$

It has been shown by Blow & Crick (1959) that the minimum mean square error in a computed electron density synthesis is achieved when the Fourier coefficients $\bar{F}_{\mathbf{h}}$ are defined by

$$\bar{F}_{\mathbf{h}} = \int_C F_{\mathbf{h}} P(F_{\mathbf{h}})\, \mathrm{d}F_{\mathbf{h}}, \tag{4.1}$$

where $P(F_{\mathbf{h}})$ is the probability distribution of the structure factor $F_{\mathbf{h}}$. In order to obtain this centroid or expected value of $F_{\mathbf{h}}$, one must first obtain $P(F_{\mathbf{h}})$ itself. Hence, in the case of single isomorphous replacement, the joint conditional probability distribution of the pair of random variables (F,Φ), given the two magnitudes $|D_{\mathbf{h}}|$, $|H_{\mathbf{h}}|$ and phase $\omega_{\mathbf{h}}$ is needed. If we denote by

$$P_{1,1|2,1} = P(F,\Phi|D,H,\omega) \tag{4.2}$$

the joint conditional probability distribution of the magnitude $|F_{\mathbf{h}}|$ and phase $\varphi_{\mathbf{h}}$ of the native structure factor $F_{\mathbf{h}}$, given the two magnitudes $|D_{\mathbf{h}}|$, $|H_{\mathbf{h}}|$ and phase $\omega_{\mathbf{h}}$, then, in view of the joint probability density function of the three complex structure factors, P, (I.2), $P_{1,1|2,1}$ is obtained by fixing the variables D, H, ω, integrating over the ψ variable from 0 to 2π, and renormalizing. Hence,

$$\begin{aligned} P_{1,1|2,1} = {} & \frac{GF}{\pi\Sigma I_0(2\Upsilon DH/G)} \\ & \times \exp\left\{-\frac{1}{\Sigma G}[s_1^2\Gamma^2D^2 + \Gamma^2\Upsilon^2H^2 + G^2F^2]\right\} \\ & \times \exp\left\{\frac{-2\Upsilon\Gamma FH\cos(\Phi-\omega)}{\Sigma}\right\} \\ & \times I_0\left\{\frac{2D}{\Sigma}[s_1^2\Gamma^2F^2 + (s_2+\mu_F)^2\Upsilon^2H^2 \right. \\ & + 2s_1(s_2+\mu_F)\Gamma\Upsilon FH \\ & \left. \times \cos(\Phi-\omega)]^{1/2}\right\}, \end{aligned} \tag{4.3}$$

where

$$G = [(s_2 + s_1 + \mu_D)s_1 - \Upsilon^2]. \tag{4.4}$$

The parameters Σ, s_1, s_2, Υ, Γ, μ_F, μ_D are defined by (3.4)–(3.10), and are presumed known. The probability distributions associated with the imperfect isomorphism and ensemble error special cases follow from the substitutions discussed in § 3.1 and § 3.2 respectively and will not be considered further.

4.1 *The expected value of* $|F_{\mathbf{h}}|\exp(i\varphi_{\mathbf{h}})$

In view of the joint conditional probability distribution, (4.3), of the pair of random variables (F, Φ), given the two magnitudes $|D_{\mathbf{h}}|$, $|H_{\mathbf{h}}|$, and phase $\omega_{\mathbf{h}}$, the expected value of $|F_{\mathbf{h}}|\exp(i\varphi_{\mathbf{h}})$ is easily obtained. Denote by

$$\epsilon = \epsilon\,[F\exp(i\Phi)|D,H,\omega] \tag{4.5}$$

the conditional expected value of $|F_{\mathbf{h}}|\exp(i\varphi_{\mathbf{h}})$, given the magnitudes $|D_{\mathbf{h}}|$, $|H_{\mathbf{h}}|$ and phase $\omega_{\mathbf{h}}$. Then,

$$\epsilon = \int_0^\infty \int_0^{2\pi} F\exp(i\Phi)\, P_{1,1\,2,1}\, \mathrm{d}F\, \mathrm{d}\Phi. \tag{4.6}$$

The integration is readily performed leading to

$$\epsilon = \frac{\Gamma}{G}\left[\frac{I_1(2\Gamma DH/G)}{I_0(2\Gamma DH/G)}\, s_1 D - \Upsilon H\right]\exp(i\omega), \tag{4.7}$$

where G is defined by (4.4).

4.2 *Imperfect isomorphism only*

Employing the assumptions of § 3.1, the expected value of $F\exp(i\Phi)$ for the special case of imperfect isomorphism, ϵ_I, follows directly and is given by

$$\epsilon_I = \frac{\Gamma}{s_2}\left[\frac{I_1(2DH/s_2)}{I_0(2DH/s_2)}D - H\right]\exp(i\omega). \tag{4.8}$$

4.3 *Ensemble error model*

In view of § 3.2, the expected value of $F\exp(i\Phi)$ for the special case of the ensemble error model ϵ_E, is given by

$$\epsilon_E = \frac{s_2}{(s_2+\mu_D)}\left\{\frac{I_1[2DH/(s_2+\mu_D)]}{I_0[2DH/(s_2+\mu_D)]}D - H\right\}\exp(i\omega). \tag{4.9}$$

It is apparent that errors associated with the two models manifest themselves in different ways; in particular, the Bessel functions of (4.8) are independent of the error source, Γ. Conversely, the values of the Bessel functions in (4.9) directly depend on the value of

μ_D, the source of errors for this model. If both models were free of errors then $\Gamma = s_2$, $\mu_D = 0$ and

$$\epsilon_I = \epsilon_E = \left[\frac{I_1(2DH/s_2)}{I_0(2DH/s_2)} D - H\right] \exp(i\omega), \qquad (4.10)$$

which is the weight function for a Sim-weighted difference Fourier synthesis (Sim, 1960; Nixon & North, 1976).

The various probability distributions and expectations assume as known, or at least well estimated, several parameters, μ_F, μ_D (3.9), (3.10) for the non-centrosymmetric functions, E_D^2, D_F^2, (3.14), (3.15), for the centrosymmetric ones, and s_1, s_2, Υ, Γ, (3.5)–(3.8), for both. Procedures for obtaining estimates for these parameters are considered next.

5. Estimates of unknown parameters

Excluding the three magnitudes $|D_\mathbf{h}|$, $|F_\mathbf{h}|$, $|H_\mathbf{h}|$ and phase $\omega_\mathbf{h}$, there are two classes of parameters associated with the probability density functions given in (3.3), (3.12), (4.3); those which are independent of, and those which are dependent on, the value of $\sin\theta/\lambda$.

The parameters independent of $\sin\theta/\lambda$ are μ_F, μ_D, E_F^2, E_D^2. They arise from experimental errors associated with the native and derivative magnitudes in the non-centrosymmetric and centrosymmetric cases. Consequently, *a priori* estimates for these variances may be made (Blow & Crick, 1959) based upon multiple estimates of equivalent reflections from different crystals or symmetry-related reflections in the same crystal.

The second class of parameters, dependent upon the value of $\sin\theta/\lambda$, consist of s_1, s_2, Υ, Γ. Estimates of the values of σ_{HK}^2 are obtainable from a refinement of the heavy-atom parameters. These estimates may then be used to compute the values of s_1, (3.5), and Υ, (3.7), for each reflection. If the contents of the unit cell of the native crystal are known, then s_2 may be directly evaluated for each reflection from its definition, (3.6). When the contents of the unit-cell are imprecisely known, then s_2 may be estimated in shells of $h = \sin\theta/\lambda$, by

$$s_2 = \langle |F_\mathbf{h}|^2 \rangle_h. \qquad (5.1)$$

The only parameter that remains to be considered is Γ. Estimation techniques for this parameter depend on the existence of centrosymmetric zones. These estimates may then be used in the initial phasing. A procedure for obtaining the value of Γ from non-centrosymmetric data is more complicated and beyond the scope of this paper.

Two procedures are considered for estimating the value of Γ; the first is by the method of moments, the second by the method of maximum likelihood (Kendall & Stuart, 1973). In view of (I.3) it is easily shown that

$$\epsilon[(D - F - H)^2] = 2(s_2 - \Gamma) + 2(s_1 - \Upsilon) + E_D^2 + E_F^2. \qquad (5.2)$$

From assumed values of s_2, s_1, Υ, E_D^2, E_F^2, estimates of Γ may be obtained by substituting the expected value, ϵ, for local averages in shells of $\sin\theta/\lambda$, *i.e.*

$$\epsilon[(D - F - H)^2] \simeq \langle (D - F - H)^2 \rangle_h, \qquad (5.3)$$

and solving for Γ. The individual terms contributing to the average in (5.3) may be computed by minimizing the discrepancy of each $(D - F - H)$ term from the average. A potential disadvantage with this estimate of Γ is that it is not a minimum variance one. For this reason, estimation by the method of maximum likelihood may be preferred, and is obtained from the solution of the cubic equation

$$\Gamma^3 - \Gamma^2 \left\langle F\left(D - \frac{\Upsilon}{s_1} H\right)\right\rangle_h + \Gamma\left[\left\langle \left(D - \frac{\Upsilon}{s_1} H\right)^2 + \frac{K}{s_1(s_2 + E_F^2)} F^2 \right\rangle_h - \frac{K}{s_1}\right](s_2 + E_F^2) - \frac{K(s_2 + E_F^2)}{s_1} \left\langle F\left(D - \frac{\Upsilon}{s_1} H\right)\right\rangle_h = 0, \qquad (5.4)$$

where

$$K = [(s_2 + s_1 + E_D^2)s_1 - \Upsilon^2], \qquad (5.5)$$

and where the signed values of $D_\mathbf{h}$, $F_\mathbf{h}$, $H_\mathbf{h}$ in (5.4) are obtained by minimizing the residual of (5.3). If more than one real root should happen to exist, that root which maximizes the likelihood equation (II.1) is selected. Also, since E_F^2, E_D^2 are the variances of the magnitudes from the centrosymmetric zones, the variances μ_F, μ_D, for the magnitudes from the non-centrosymmetric zones may be used by substituting

$$\mu_D = 2E_D^2 \qquad (5.6)$$

and

$$\mu_F = 2E_F^2 \qquad (5.7)$$

into (5.4). This result follows from the probability density defined by (2.13), (2.17).

6. A comparison with other SIR formulations

Clearly, the major distinction between other error models and the ones considered here is the notion of the imperfect-isomorphism model considered in § 2.1. In order to make a simple comparison between the latter and other statistical-error models, some numerical examples are presented. In particular, some graphs are given of conditional probability distributions of $\varphi_\mathbf{h}$,

given representative values of the three magnitudes $|D_{\mathbf{h}}|$, $|F_{\mathbf{h}}|$, $|H_{\mathbf{h}}|$ and heavy-atom phase $\omega_{\mathbf{h}}$. The effects of the error model on these probability distributions are illustrated in the light of two assumptions: there is perfect agreement between the true and calculated heavy-atom structure factors, and there are no experimental errors associated with the derivative and native structure-factor magnitudes. Under these conditions, the following probability distributions are compared: (3.17) (in the accompanying figures denoted by *II*) corresponding to the imperfect-isomorphism model; (3.21) (denoted by *EE*) associated with the ensemble error model; Blow & Crick (1959), equation (23), denoted by *BC*; and Einstein (1977), equation (8), denoted by *RE*.

Error estimates for the Blow–Crick and Einstein formulas are obtained from

$$E^2 = \langle (|D_{\mathbf{h}} - F_{\mathbf{h}}| - |H_{\mathbf{h}}|)^2 \rangle_h, \qquad (6.1)$$

and the estimate of errors from the ensemble error model is found from

$$\mu_D = 2E^2. \qquad (6.2)$$

For the imperfect-isomorphism model,

$$\langle (|D_{\mathbf{h}} - F_{\mathbf{h}}| - |H_{\mathbf{h}}|)^2 \rangle_h = 2(s_2 - \Gamma). \qquad (6.3)$$

Hence for this model, once s_2 is specified, Γ is determined. The figures acccompanying this section are plots of the four normalized probability distributions *versus* Φ in the interval $0° \leq \Phi \leq 360°$. In Figs. 1(*a*) and 1(*b*) the value of s_2 is arbitrarily $(200)^2$, the value of F^2 in Fig. 1(*a*). Hence, in view of the value of E^2 listed in the figures, the value of Γ is 38 750, which differs only slightly from the maximum-likelihood estimate of 38 745. In Fig. 1(*c*), the value of s_2 is $(400)^2$, or four times the value of F^2 in Fig. 1(*a*), and in Fig. 1(*d*), s_2 is one fourth the value of F^2. Table 1 lists the mean

Table 1. *Mean figure of merit*

	Distribution			
	II	*EE*	*BC*	*RE*
Fig. 1(*a*)	0·31	0·15	0·21	0·15
Fig. 1(*b*)	0·43	0·25	0·35	0·29
Fig. 1(*c*)	0·25	0·15	0·21	0·15
Fig. 1(*d*)	0·51	0·15	0·21	0·15

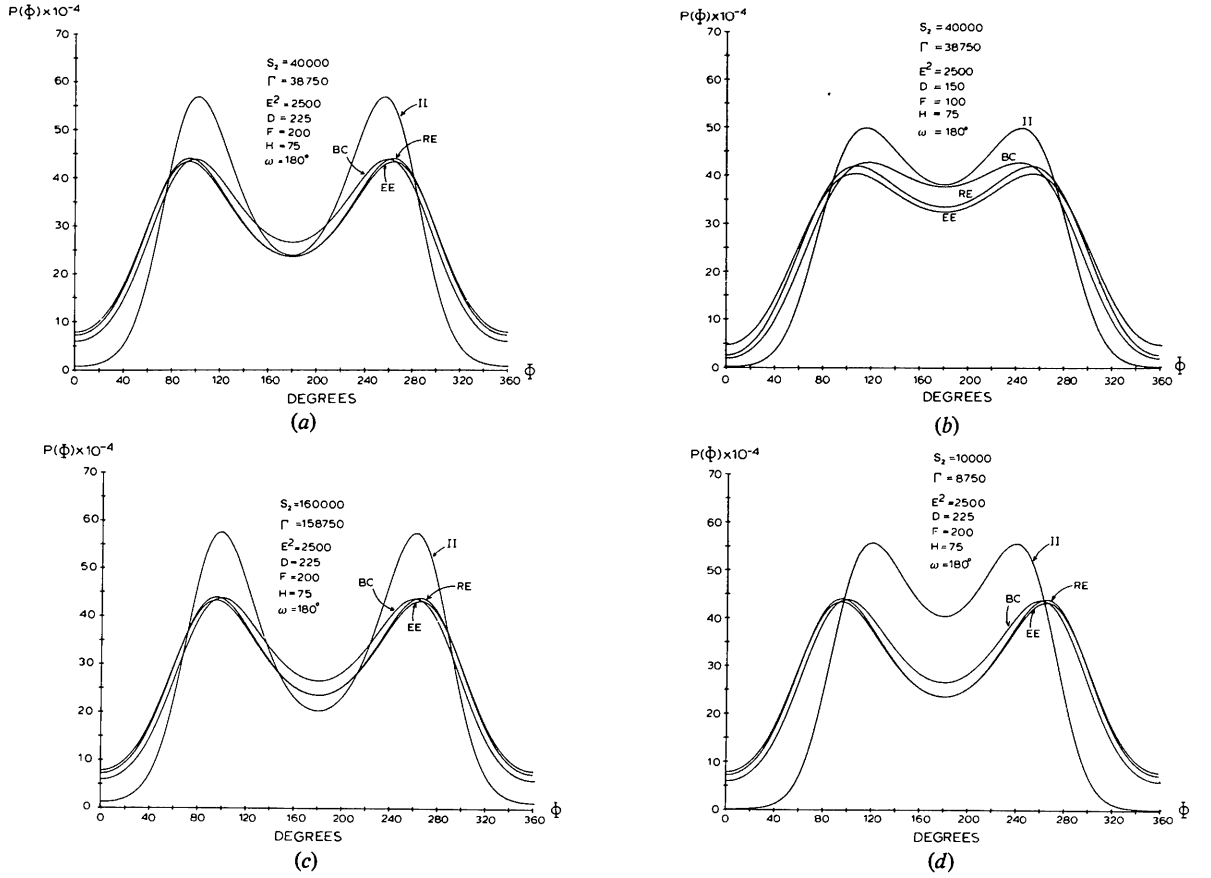

Fig. 1. The conditional probability distributions *II*, *EE*, *BC*, *RE* of the native phase $\varphi_{\mathbf{h}}$ given the three magnitudes $|D_{\mathbf{h}}|$, $|F_{\mathbf{h}}|$, $|H_{\mathbf{h}}|$ and phase $\omega_{\mathbf{h}}$ for the values listed; see § 6.

figures of merit for each distribution in the four figures.

The plot of the *II* distribution in Fig. 1(*d*) is quite different, not only in comparison with the other three distributions in this figure, but also with the *II* distributions in both Figs. 1(*a*) and 1(*c*). In particular, the modes of the distribution are at 120° and 240°, compared with 100° and 260° for the others. The effect of changing s_2 in the *II* distribution, (3.17), is to alter the normalization of the three magnitudes $|D_{\mathbf{h}}|$, $|F_{\mathbf{h}}|$, $|H_{\mathbf{h}}|$ which is not possible in the other three distributions. By defining the three normalized structure-factor magnitudes

$$|D'_{\mathbf{h}}| = |D_{\mathbf{h}}|/s_2^{1/2}, \tag{6.4}$$

$$|F'_{\mathbf{h}}| = |F_{\mathbf{h}}|/s_2^{1/2}, \tag{6.5}$$

$$|H'_{\mathbf{h}}| = |H_{\mathbf{h}}|/s_2^{1/2}, \tag{6.6}$$

and by defining

$$\rho = \Gamma/s_2, \tag{6.7}$$

(3.17) may be expressed in terms of the normalized variables F', D', H' and ρ. Thus,

$$P_{1|3,1} = \frac{1}{K}\exp\left\{\frac{-2\rho F'H'\cos(\Phi-\omega)}{(1-\rho^2)}\right\} \times I_0\left\{\frac{2D'}{(1-\rho^2)}\,[H'^2+\rho^2F'^2+2\rho F'H' \times \cos(\Phi-\omega)]^{1/2}\right\}. \tag{6.8}$$

Although not presented here, the probability distribution associated with equation (4), Rossmann & Blow, and equations (6), (17), Hendrickson & Lattman have been compared. The results from these two formulas are not particularly different from the *EE*, *BC*, and *RE* distributions.

7. Concluding remarks

The assumption of imperfect isomorphism between the atomic positions in the native crystal and the atomic positions in the native component of the derivative crystal leads to estimates of phase whose reliability is directly affected by the value of $\sin\theta/\lambda$. Furthermore, the direct incorporation of this type of error into the phase determining formalism yields distributions having a smaller variance with potentially different estimates of phase than that derived from distributions based on the assumption of Gaussian errors alone.

I wish to thank Dr Herbert Hauptman for his continued encouragement and support of this work, and for his reading and critique of this paper. This research was supported by Grant No. CHE76-17582 from the National Science Foundation.

APPENDIX I
The joint probability distribution of the three structure factors $D_{\mathbf{h}}$, $F_{\mathbf{h}}$, $H_{\mathbf{h}}$

It is assumed that a crystal containing N, not necessarily identical, atoms per unit cell in the space group $P1$ is given. The observed structure factor, $F_{\mathbf{h}}$, for this crystal is defined by (2.9). Another crystal containing the same N atoms, in positions approximately identical to those of the first crystal, as well as additional (heavy) atoms per unit cell, not necessarily identical, in $P1$ is also given. The observed structure factor, $D_{\mathbf{h}}$, for the latter crystal is defined by (2.8). Estimates of the atomic parameters of the M additional atoms may not correspond precisely to the parameters specifying the atoms within the second crystal. A calculated structure factor $H_{\mathbf{h}}$, relative to the true heavy-atom-component structure factor in (2.8), is defined by (2.10). The reciprocal-lattice vector $\mathbf{h}$ is fixed. The N and M atomic position vectors $\mathbf{r}_j$ and $\mathbf{r}'_k$, respectively, are assumed to be uniformly and independently distributed in the unit cell. The N (M) random-shift vectors $\boldsymbol{\delta}_j$ ($\boldsymbol{\varepsilon}_k$) are normally distributed, with density functions given by (2.2), (2.7). The two ensemble random variables χ_D, ξ_F, having components $(|\chi_D|,\xi_D)$ and $(|\chi_F|,\xi_F)$, have joint probability densities given by (2.13). Then the three structure factors $D_{\mathbf{h}}$, $F_{\mathbf{h}}$, $H_{\mathbf{h}}$, as functions of these enumerated primitive random variables, are themselves random variables. Denoting by

$$P = P(D,F,H,\psi,\Phi,\omega) \tag{I.1}$$

the joint probability density function associated with the three magnitudes $|D_{\mathbf{h}}|$, $|F_{\mathbf{h}}|$, $|H_{\mathbf{h}}|$ and phases $\psi_{\mathbf{h}}$, $\varphi_{\mathbf{h}}$, $\omega_{\mathbf{h}}$, then, complete to terms of first order,

$$P = \frac{DFH}{\pi^3\Sigma}\exp\left(\frac{1}{\Sigma}\{-s_1(s_2+\mu_F)D^2 - [(s_2+s_1+\mu_D)s_1-\Upsilon^2]F^2 - [(s_2+s_1+\mu_D)(s_2+\mu_F)-\Gamma^2]H^2 + 2s_1\Gamma DF\cos(\Psi-\Phi) + 2(s_2+\mu_F)\Upsilon DH\cos(\Psi-\omega) - 2\Upsilon\Gamma FH\cos(\Phi-\omega)\}\right). \tag{I.2}$$

The parameters Σ, s_1, s_2, Υ, Γ, μ_F, μ_D are respectively defined by (3.4)–(3.10). The analogous joint probability density function of the three centrosymmetric structure factors, defined by (2.12)–(2.14), is

$$\bar{P} = \bar{P}(D,F,H) = (2\pi)^{-3/2}\bar{\Sigma}^{-1/2}\exp\left(-\frac{1}{2\bar{\Sigma}}\{s_1(s_2+E_F^2)D^2 + [(s_2+s_1+E_D^2)s_1-\Upsilon^2]F^2 + [(s_2+s_1+E_D^2)(s_2+E_F^2)-\Gamma^2]H^2 - 2s_1\Gamma DF - 2(s_2+E_F^2)\Upsilon DH + 2\Upsilon\Gamma FH\}\right). \tag{I.3}$$

The parameters $\bar{\Sigma}$, s_1, s_2, Υ, Γ, E_D^2, E_F^2 are defined by (3.13), (3.5), (3.6), (3.7), (3.8), (3.14), (3.15). The mathematical techniques employed in deriving (I.2) and (I.3) are similar to those discussed by Hauptman (1975*a*,*b*). The reader is referred to these papers for a discussion of similar mathematical analyses.

APPENDIX II
The maximum-likelihood estimate of Γ

In view of equation (18.1), Kendall & Stuart, page 37, the likelihood function, L, of (I.3) based on n independent observations of the three centrosymmetric structure factors $D_{\mathbf{h}}$, $F_{\mathbf{h}}$, $H_{\mathbf{h}}$ is

$$L = (2\pi)^{-3n/2}\,\bar{\Sigma}^{-n/2}\exp\left\{\frac{-s_1(s_2+E_F^2)}{2\bar{\Sigma}}\right.$$
$$\times\sum_{j=1}^{n} D_j^2 - \frac{[(s_2+s_1+E_D^2)s_1-\Upsilon^2]}{2\bar{\Sigma}}$$
$$\times\sum_{j=1}^{n} F_j^2 - \frac{[(s_2+s_1+E_D^2)(s_2+E_F^2)-\Gamma^2]}{2\Sigma}\sum_{j=1}^{n} H_j^2$$
$$+\frac{s_1\Gamma}{\bar{\Sigma}}\sum_{j=1}^{n} D_jF_j + \frac{(s_2+E_F^2)\Upsilon}{\bar{\Sigma}}\sum_{j=1}^{n} D_jH_j$$
$$\left.-\frac{\Upsilon\Gamma}{\bar{\Sigma}}\sum_{j=1}^{n} F_jH_j\right\}, \qquad \text{(II.1)}$$

where

$$\bar{\Sigma} = \{(s_2+E_F^2)[(s_2+s_1+E_F^2)s_1-\Upsilon^2] - s_1\Gamma^2\}. \qquad \text{(II.2)}$$

The maximum-likelihood estimate of Γ is obtained from the solution of

$$\frac{dL}{d\Gamma} = 0. \qquad \text{(II.3)}$$

This equation assumes estimates for the parameters s_1, s_2, Υ, E_D^2, E_F^2, and estimates for the three structure factors $D_{\mathbf{h}}$, $F_{\mathbf{h}}$, $H_{\mathbf{h}}$. The differentiation of L is straightforward and leads to the cubic equation in Γ

$$\Gamma^3 - \Gamma_2^2\left\langle F\left(D-\frac{\Upsilon}{s_1}H\right)\right\rangle$$
$$+\Gamma\left[\left\langle\left(D-\frac{\Upsilon}{s_1}H\right)^2+\frac{KF^2}{s_1(s_2+E_F^2)}\right\rangle-\frac{K}{s_1}\right](s_2+E_F^2)$$
$$+\frac{K(s_2+E_F^2)}{s_1}\left\langle F\left(D-\frac{\Upsilon}{s_1}H\right)\right\rangle = 0, \qquad \text{(II.4)}$$

where

$$K = [(s_2+s_1+E_D^2)s_1-\Upsilon^2]. \qquad \text{(II.5)}$$

The averages are taken over the n independent observations per shell of $\sin\theta/\lambda$.

References

BLOW, D. M. & CRICK, F. H. C. (1959). *Acta Cryst.* **12**, 794–802.

EINSTEIN, R. J. (1977). *Acta Cryst.* A**33**, 75–85.

HAUPTMAN, H. (1975*a*). *Acta Cryst.* A**31**, 671–679.

HAUPTMAN, H. (1975*b*). *Acta Cryst.* A**31**, 680–687.

HENDRICKSON, W. A. & LATTMAN, E. E. (1970). *Acta Cryst.* B**26**, 136–143.

KARLE, J. & HAUPTMAN, H. (1958). *Acta Cryst.* **11**, 264–269.

KENDALL, M. G. & STUART, A. (1973). *The Advanced Theory of Statistics*, Vol. 2, p. 37. New York: Hafner.

NIXON, P. E. & NORTH, A. C. T. (1976). *Acta Cryst.* A**32**, 325–333.

ROSSMANN, M. G. & BLOW, D. M. (1961). *Acta Cryst.* **14**, 641–647.

SIM, G. A. (1960). *Acta Cryst.* **13**, 511–512.