Acta Crystallographica Section D
Biological Crystallography
ISSN 0907-4449

Application of the complex multivariate normal distribution to crystallographic methods with insights into multiple isomorphous replacement phasing

Navraj S. Pannu,[a,b,c]* Airlie J. McCoy[a] and Randy J. Read[a]*

[a]Department of Haematology, Cambridge Institute for Medical Research, Wellcome Trust/MRC Building, Hills Road, Cambridge CB2 2XY, England, [b]Biophysical Structural Chemistry, Leiden Institute of Chemistry, Gorlaeus Laboratories, Leiden University, PO Box 9502, 2300 RA Leiden, The Netherlands, and [c]Trinity College, Cambridge CB2 1TQ, England

Correspondence e-mail:
raj@chem.leidenuniv.nl, rjr27@cam.ac.uk

Received 14 April 2003
Accepted 2 September 2003

Probabilistic methods involving maximum-likelihood parameter estimation have become a powerful tool in computational crystallography. At the centre of these methods are the relevant probability distributions. Here, equations are developed based on the complex multivariate normal distribution that generalize the distributions currently used in maximum-likelihood model and heavy-atom refinement. In this treatment, the effects of various sources of error in the experiment are considered separately and allowance is made for correlations among sources of error. The multivariate distributions presented are closely related to the distributions previously derived in *ab initio* phasing and can be applied to many different aspects of a crystallographic structure-determination process including model refinement, density modification, heavy-atom phasing and refinement or combinations of them. The underlying probability distributions for multiple isomorphous replacement are re-examined using these techiques. The re-analysis requires the underlying assumptions to be made explicitly and results in a variance term that, unlike those previously used for maximum-likelihood multiple isomorphous replacement phasing, is expressed explicitly in terms of structure-factor covariances. Test cases presented show that the newly derived multiple isomorphous replacement likelihood functions perform satisfactorily compared with currently used programs.

1. Introduction

Multivariate probabilistic methods have a long history in computational crystallography and have played an important role in the advancement of the field. Multivariate distributions emerged in crystallography with the classic work of Hauptman & Karle (1953). They recognized that the relationship between structure-factor amplitudes alone can be used to obtain phase estimates *ab initio* and derived an approximation to a multivariate joint structure-factor probability distribution to obtain initial phase estimates. Their approach has been developed by many others, including Bertaut (1955), Klug (1958) and Bricogne (1984), in efforts to form more accurate and efficient approximations to the underlying distributions required in *ab initio* phasing (for a review, see Giacovazzo, 1998). Recently, Lunin *et al.* (1998) have elaborated on the above work and applied it to low-resolution *ab initio* phasing. Hauptman (1982) and Giacovazzo & Siliqi (1983) have applied multivariate joint distributions to the case of single-wavelength anomalous diffraction and this treatment has recently been further developed and generalized for substructure detection (Burla *et al.*, 2002) and phasing (Giacovazzo & Siliqi, 2001*a*,*b*). Finally, distributions necessary for maximum-likelihood structure refinement can be obtained from multivariate distributions (Bricogne, 1997; Pannu, 1997) and multivariate distributions have been applied to a maximum-likelihood analysis of molecular replacement (Read, 2001).

An appropriate multivariate distribution of structure factors provides a general basis for a variety of crystallographic experiments (*e.g.* experimental phasing, molecular replacement and model refinement), can consider different errors (*e.g.* lack of isomorphism, errors in atomic parameters and measurement errors) arising in these experiments separately and can account for the effect of correlated errors. These experiments (and their errors) can all be represented by a distribution of N structure-factor observations given M models,

$$P(|F_1|, \ldots, |F_N|; |F_{c1}|, \alpha_{c1}, \ldots, |F_{cM}|, \alpha_{cM}). \tag{1}$$

For example, in the case of a two-wavelength MAD phasing experiment, $N = 4$, $M = 4$ and the observed structure-factor amplitudes collected at a given wavelength λ_i and Friedel pair are represented as $|F_1| = |F^+_{\lambda_1}|$, $|F_2| = |F^-_{\lambda_1}|$, $|F_3| = |F^+_{\lambda_2}|$ and $|F_4| = |F^-_{\lambda_2}|$ with $|F_{c1}|$, α_{c1}, $\ldots$, $|F_{c4}|$, α_{c4}, corresponding to the calculated heavy-atom structure factors for a given wavelength λ_i and Friedel pair (*i.e.* $|F_{c1}| = |H^+_{\lambda_1}|$). As another example, in the case of molecular replacement with one data set and multiple choices of model, $N = 1$ and $M > 1$ (Read, 2001).

To derive the above distribution, the starting point is the multivariate distribution of a collection of $N + M$ complex structure factors,

$$P(F_1, \ldots, F_N, F_{c1}, \ldots, F_{cM}). \tag{2}$$

The central limit theorem will be applied to obtain a multivariate Gaussian distribution. In the analysis that follows, the atomic positions will be considered as the random variables in deriving a multivariate normal distribution of acentric structure factors. Acentric structure factors can be treated as real and imaginary terms or complex numbers (as shown below): centric structure factors obey real multivariate distributions and can be derived similarly with complex distributions. To apply the central limit theorem, it is assumed that there is a sufficiently large number of random vectors that are independent and identically distributed. This assumption is not necessarily valid: Shmueli & Weiss (1995) have shown that significant deviations from a Gaussian distribution can occur in the case of a crystal containing an outstandingly heavy atom with 14 C atoms in its asymmetric unit. Higher-level approximations to the joint probability distribution of structure factors, as developed by many others in the field of direct methods, would no doubt lead to a more accurate model of the joint distribution, but in the case of macromolecular phasing and refinement, where the errors are typically smaller than in direct methods and the number of atoms larger, the approximations will in general be better.

2. Complex multivariate Gaussian distribution

In Appendix A, the covariance terms for the real and imaginary parts of the structure factors are calculated and it is shown that $\langle A_i A_j \rangle \simeq \langle B_i B_j \rangle$ and $\langle A_i B_j \rangle \simeq -\langle A_j B_i \rangle$. Wooding (1956) showed that if the real and imaginary components of a set of complex numbers have a multivariate Gaussian distribution with the conditions

$$\langle A_i A_j \rangle = \langle B_i B_j \rangle,$$
$$\langle A_i B_j \rangle = -\langle A_j B_i \rangle,$$

then the distribution of $Z = (Z_1, \ldots, Z_N) = (A_1 + iB_1, \ldots, A_N + iB_N)$ is then a complex multivariate Gaussian,

$$P(Z) = \frac{1}{\pi^N \det(\Sigma)} \exp[-(Z^*)^T \Sigma^{-1} Z]. \tag{3}$$

In the above equation, Σ is the covariance matrix of the multivariate distribution with the i, jth entry of the covariance matrix (σ_{ij}) defined as $\sigma_{ij} = \langle Z_i Z_j^* \rangle$. In the case of crystallographic phasing and refinement, $Z_i = F_i$, the ith structure factor. In Appendix B, the terms of the covariance matrix are calculated analytically considering both coordinate and measurement errors along with anomalously scattering atoms.

3. Application to MIR phasing assuming uncorrelated errors

The complex multivariate normal distribution of structure factors can be used in a straightforward derivation of equations appropriate for multiple isomorphous replacement phasing when it is assumed that the errors affecting the different derivatives are uncorrelated. The equations developed here are similar to those discussed previously (Bricogne, 1991; Read, 1991) and implemented in *SHARP* (de La Fortelle & Bricogne, 1997), but express the variance term explicitly in terms of structure-factor covariances. Furthermore, the derivation based on multivariate statistics makes explicit the assumptions that are made.

To carry out heavy-atom refinement and phasing by maximum likelihood, the joint distribution of all the observations given the parameters of the model is needed. To obtain this, the joint probability distribution of the true structure-factor amplitude ($|F|$), phase (α) and structure-factor amplitude for derivative j for all $n + 1$ derivatives ($\{|F_{j0}|\}_{j=0\ldots n}$) conditional on the calculated heavy-atom structure factor for all $n + 1$ derivatives ($\{H_{jc}\}_{j=0\ldots n}$) is calculated (the native structure factor is denoted by the zeroth derivative, which will generally have a null heavy-atom model). The native and derivative structure factors are all highly correlated; to eliminate this correlation, the true structure factor F is introduced as a dummy variable in the joint probability distribution. As will be seen below, by making the distributions conditional on an assumed value of F the correlation is removed, but at the expense of integrating over all possible values of the dummy variable F.

For the acentric case, the distribution $P(F_0 \ldots F_n, F, H_0 \ldots H_n)$ will be approximated as a complex multivariate

Gaussian of mean zero and a covariance matrix that can be partitioned into submatrices as follows:

$$\begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}, \tag{4}$$

where Σ_{11} is an $n \times n$ matrix and Σ_{21} is the Hermitian transpose of Σ_{12}.

$$\Sigma_{11} = \begin{pmatrix} \langle F_0F_0^*\rangle & \langle F_0F_1^*\rangle & \dots & \langle F_0F_n^*\rangle \\ \langle F_0F_1^*\rangle^* & \langle F_1F_1^*\rangle & \dots & \langle F_1F_n^*\rangle \\ \vdots & \vdots & \ddots & \vdots \\ \langle F_0F_n^*\rangle^* & \langle F_1F_n^*\rangle^* & \dots & \langle F_nF_n^*\rangle \end{pmatrix}, \tag{5}$$

$$\Sigma_{12} = \begin{pmatrix} \langle F_0F^*\rangle & \langle F_0H_0^*\rangle & \dots & \langle F_0H_n^*\rangle \\ \langle F_1F^*\rangle & \langle F_1H_0^*\rangle & \dots & \langle F_1H_n^*\rangle \\ \vdots & \vdots & \ddots & \vdots \\ \langle F_nF^*\rangle & \langle F_nH_0^*\rangle & \dots & \langle F_nH_n^*\rangle \end{pmatrix}, \tag{6}$$

$$\Sigma_{22} = \begin{pmatrix} \langle FF^*\rangle & \langle FH_0^*\rangle & \dots & \langle FH_n^*\rangle \\ \langle FH_0^*\rangle^* & \langle H_0H_0^*\rangle & \dots & \langle H_0H_n^*\rangle \\ \vdots & \vdots & \ddots & \vdots \\ \langle FH_n^*\rangle^* & \langle H_0H_n^*\rangle^* & \dots & \langle H_nH_n^*\rangle \end{pmatrix}. \tag{7}$$

In the case of no anomalous diffraction, the covariance for $\langle F_iF_j^*\rangle$ can be simplified from Appendix *B*,

$$\langle F_iF_j^*\rangle \simeq \begin{cases} \sum\limits_{k=1}^{N_i} \varepsilon d_{kij} f_{ki} f_{kj} & \text{if } i \neq j, \\ \sum\limits_{k=1}^{N_i} \varepsilon f_{ki}^2 + \sigma_{F_i}^2 & \text{if } i = j. \end{cases} \tag{8}$$

The covariance terms are real and the covariance matrix is symmetric. d_{kij} is shown as a function of each atom; however, in practice, it is estimated as a function of resolution and will be denoted by D_{ij}. The parameter D_{ij} accounts for non-isomorphism and also absorbs errors in overall scale and temperature factors. When considering the covariances between derivative structure factors, the number of atoms in common will be the same as that in the native crystal (or the 'zeroth derivative'). When considering the covariance between the derivative structure factor F_i and heavy-atom structure H_j, the atoms in common will be those contained in the heavy-atom model. Furthermore, it will be assumed that there are no common sites between models in different derivatives. Therefore, the covariances become the following

$$\langle F_iF_j^*\rangle = \begin{cases} \varepsilon D_{ij}\Sigma_N & \text{if } i \neq j, \\ \varepsilon \Sigma_{N_j} + \sigma_{F_j}^2 & \text{if } i = j, \end{cases} \tag{9}$$

$$\langle F_iH_j^*\rangle = \begin{cases} 0 & \text{if } i \neq j, \\ \varepsilon D_{H_j}\Sigma_{H_j} & \text{if } i = j, \end{cases} \tag{10}$$

where Σ_N is the variance parameter for the dummy variable F and is estimated as the sum of the scattering factors squared for the 'zeroth derivative' or native crystal, Σ_{N_0}. The parameters D_{H_j} account for errors in the heavy-atom model for derivative j, Σ_{H_j} is the sum of the scattering factors squared of the heavy-atom model and Σ_{N_j} is the sum of the scattering factors squared for derivative crystal j. Using the above covariances, the partitioned matrices can be filled in:

$$\Sigma_{11} = \begin{pmatrix} \varepsilon\Sigma_{N_0} + \sigma_{F_0}^2 & \varepsilon D_{01}\Sigma_N & \dots & \varepsilon D_{0n}\Sigma_N \\ \varepsilon D_{01}\Sigma_N & \varepsilon\Sigma_{N_1} + \sigma_{F_1}^2 & \dots & \varepsilon D_{1n}\Sigma_N \\ \vdots & \vdots & \ddots & \vdots \\ \varepsilon D_{0n}\Sigma_N & \varepsilon D_{1n}\Sigma_N & \dots & \epsilon\Sigma_{N_n} + \sigma_{F_n}^2 \end{pmatrix}, \tag{11}$$

$$\Sigma_{12} = \begin{pmatrix} \varepsilon D_0\Sigma_N & \varepsilon D_{H_0}\Sigma_{H_0} & 0 & \dots & 0 \\ \varepsilon D_1\Sigma_N & 0 & \varepsilon D_{H_1}\Sigma_{H_1} & \dots & 0 \\ \vdots & \vdots & 0 & \ddots & \vdots \\ \varepsilon D_n\Sigma_N & 0 & 0 & \dots & \varepsilon D_{H_n}\Sigma_{H_n} \end{pmatrix}, \tag{12}$$

$$\Sigma_{22} = \begin{pmatrix} \varepsilon\Sigma_N & 0 & \dots & 0 \\ 0 & \varepsilon\Sigma_{H_0} & \dots & 0 \\ 0 & 0 & \ddots & 0 \\ 0 & \dots & \dots & \varepsilon\Sigma_{H_n} \end{pmatrix}. \tag{13}$$

The distribution of $P(F_0, F_1, \dots, F_n; F, H_0, H_1, \dots, H_n)$ is also Gaussian with mean $\mu = \Sigma_{12}\Sigma_{22}^{-1}(F, H_0, H_1, \dots, H_n)$ and covariance matrix $\Sigma = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}$ (*e.g.* Johnson & Wichern, 1992). Performing the above operations gives the following:

$$\mu = \begin{pmatrix} D_0F + D_{H_0}H_0 \\ \vdots \\ D_nF + D_{H_n}H_n \end{pmatrix}. \tag{14}$$

$$\Sigma = \begin{pmatrix} \varepsilon\sigma_{\Delta_0}^2 + \sigma_{F_0}^2 & \omega_{10} & \dots & \omega_{0n} \\ \omega_{10} & \varepsilon\sigma_{\Delta_1}^2 + \sigma_{F_1}^2 & \dots & \omega_{1n} \\ \vdots & \vdots & \ddots & \vdots \\ \omega_{n0} & \omega_{n1} & \dots & \varepsilon\sigma_{\Delta_N}^2 + \sigma_{F_N}^2 \end{pmatrix}. \tag{15}$$

In the above equation, $\sigma_{\Delta_i}^2 = \Sigma_{N_i} - D_i^2\Sigma_N - D_{H_i}^2\Sigma_{H_i}$ and $\omega_{ij} = \varepsilon(D_{ij} - D_iD_j)\Sigma_N$. If it is assumed that non-isomorphism errors of each derivative with respect to the true structure factor are independent (*i.e.* there is no correlated lack of isomorphism), then $D_iD_j = D_{ij}$. This relationship is a property of independent variables, for which the product of expected values is equal to the expected value of a product (*e.g.* Grimmett & Stirzaker, 1992). Thus, the covariance matrix of the conditional matrix is diagonal and the resulting distribution is a product of univariate complex Gaussian distributions whose jth term has mean $D_jF + D_{H_j}H_j$ and variance $\varepsilon(\Sigma_{N_j} - D_j^2\Sigma_N - D_{H_j}^2\Sigma_{H_j}) + \sigma_{F_j}^2$. The likelihood function required for multiple isomorphous replacement phasing can be obtained from this product of univariate complex Gaussians and is derived in Appendix *C*.

In contrast to earlier publications on similar distributions (Bricogne, 1991; Read, 1991), the variance term is expressed explicitly in terms of elements derived from Wilson structure-factor variances: in *SHARP* (de La Fortelle & Bricogne, 1997), $\sigma_{\Delta_j}^2$ is defined in terms of a 'global component of non-isomorphism' attenuation factor D_j and a 'local component of non-isomorphism' attenuation factor D_{lj}, which are each parameterized by single B factors,

$$\sigma^2_{\Delta_j} = \Sigma_N(1 - D_j^2) + C_j^{\rm loc}\langle(|F_{j=1}| - |F_{j\neq1}|)^2\rangle D_{lj}^2, \qquad (16)$$

where $C_j^{\rm loc}$ is an heuristic proportionality fraction and $\langle(|F_{j\,=\,1}| - |F_{j\,\neq\,1}|)^2\rangle$ are the mean-squared isomorphous differences estimated at the start in resolution bins between the reference (*i.e.* $j = 1$) data set and the other data sets (*i.e.* $j \neq 1$) collected.

4. Implementation

The likelihood functions discussed in Appendix *C* that numerically integrate the unknown native phase angle only, denoted below as ONED, and that numerically integrate both the native amplitude and phase, denoted TWOD, with the variance term $\sigma^2_\Delta = \varepsilon(\Sigma_{N_j} - D_j^2\Sigma_N - D_{H_j}^2\Sigma_{H_j}) + \sigma^2_{F_j}$, can be applied to the refinement and phasing of heavy atoms assuming no anomalous diffraction. For the ONED function, in the case where a reflection has no native structure-factor measurement, the TWOD function was used to calculate the likelihood for the reflection.

The parameters in the likelihood functions are the atomic parameters (*i.e.* x, y, z, occupancy and isotropic B factors), an overall scale and temperature-factor parameters to scale the derivative observations relative to the native data set and the non-isomorphism D_j and heavy-atom model error D_{H_j} parameters described above, which are both a function of resolution bins.

In the tests shown below, extra care is given not to simultaneously refine parameters that are highly correlated. For example, the scale and non-isomorphism D_j parameters are highly correlated and are not refined together. Furthermore, the heavy-atom model error parameter D_{H_j} and the individual atomic B factors are highly correlated: tests (not shown) indicate that the same results are obtained if the D_{H_j} parameter is held constant and variation of the atomic B factors is allowed to account for the coordinate errors (Read, 1990).

5. Test cases

The maximum-likelihood targets of one- (ONED) and two-dimensional (TWOD) numerical integrations were compared against the programs *MLPHARE* version 4.0 (Otwinowski, 1991) from *CCP*4 (Collaborative Computational Project, Number 4, 1994), *SHARP* version 1.3.18 (denoted below as S 1.3.18), which produced better results over *SHARP* version 1.4.0 (Pannu & Read, 2002) and *SHARP* version 2.0.1 (denoted below as S 2.0.1). In all tests, the default or example scripts were used in running each program. The first test shows an SIR experiment with an isomorphous derivative, while the second test is MIR involving two relatively weaker derivatives. The results of the phasing packages were compared with amplitudes and phases obtained from the final model obtained from the Protein Data Bank.

Table 1
Statistics for SIR phasing of troponin-C using the osmium-derivative data set.

	MLPHARE	S 1.3.18	S 2.0.1	ONED	TWOD
Map correlation	0.406	0.409	0.404	0.409	0.409
Reported FOM	0.383	0.393	0.397	0.385	0.387
Mean cos(phase error)	0.303	0.307	0.306	0.308	0.308
Mean phase error (°)	66.85	66.44	66.53	66.38	66.39

Table 2
Statistics for MIR refinement and phasing of α-dendrotoxin using the mercury- and iodine-derivative data sets.

	MLPHARE	S 1.3.18	S 2.0.1	ONED	TWOD
Map correlation	0.353	0.312	0.380	0.399	0.399
Reported FOM	0.365	0.260	0.346	0.333	0.334
Mean cos(phase error)	0.282	0.261	0.315	0.323	0.322
Mean phase error (°)	68.02	69.55	65.43	64.49	64.56
Reflections phased	1992	1899	2002	2002	2002

5.1. Troponin-C

The first test system used was troponin-C, which was originally solved at 2.8 Å resolution using multiple isomorphous replacement (MIR) phasing from 11 derivatives (Herzberg & James, 1985). Of these 11, a single osmium derivative was chosen. The final atomic coordinates from the one osmium site obtained during the structure solution were used as a starting point for refinement and phasing in all programs with the occupancy set to 1 and the B factor set to 25 Å^2. Data for the native and derivative crystal were available to 2.8 Å resolution. Table 1 shows that all methods performed similarly when they are compared on the results for the same subset of 3514 reflections that are phased by *SHARP* version 1.3.18: the program left 149 reflections unphased because they had a 'variance too large for integration'. In this test, for reasons that are not yet understood, all programs overestimate phase quality significantly.

5.2. α-Dendrotoxin

The structure of α-dendrotoxin was solved by the MIRAS method (Skarzynski, 1992). In this test case, the iodine and mercury derivatives are used in addition to the native data set. Both derivative data sets were available to 2.7 Å resolution, while the native data set was collected to 2.3 Å resolution. The native and the two derivative data sets were put on an approximate absolute scale with the program *WILSON* (E. Dodson and P. Evans, unpublished work) and the two derivatives were scaled relative to the native with the program *SCALEIT* (P. Evans, E. Dodson and R. Dodson, unpublished work) both from *CCP*4 (Collaborative Computational Project, Number 4, 1994). The atomic coordinates used in the refinement in all programs were the same as those in the example script of *MLPHARE* included in the *CCP*4 distribution, with the occupancies set to 0.5 and the B factors set to 30 Å^2 for the one iodine site and one mercury site that were refined. Results from this test case are shown in Table 2. Again, the results of the phasing packages were compared with amplitudes and phases generated from the final model obtained from the Protein Data Bank and *SHARP* version 1.3.18 left some reflections unphased because they had a 'variance too large for integration'.

This test case heightens the differences between the various likelihood targets. The ONED, TWOD and *SHARP* version 2.0.1 functions perform the best in terms of map correlation, phase error and in obtaining realistic estimates of the phase probability distribution. *MLPHARE* performs better than *SHARP* version 1.3.18 in terms of phase difference, but significantly overestimates the accuracy of the phases. It should be noted though that *SHARP* version 1.3.18 can produce better phases compared with the final model phases if the 'global' and 'local' non-isomorphism parameters are just estimated and not refined. However, not refining the non-isomorphism parameters comes at a cost: the figure of merit and corresponding phase probability distributions are significantly overestimated to a similar degree as with *MLPHARE*.

6. Discussion

The multivariate derivation of the likelihood function for MIR phasing described above gives a better understanding of the approximations made by making the assumptions explicit and points the way to further improvement. In this work, it has been assumed that there is no correlation between the native structure factor and any of the structure factors from heavy-atom models (*i.e.* isomorphous addition rather than isomorphous replacement) and that there are no common sites among the derivatives. It has also been assumed that the lack-of-isomorphism errors for the different derivatives are independent, an assumption that has been recognized in the past (Bricogne, 1991; Read, 1991).

The two test cases performed exhibit currently held beliefs in phasing: in the first test, involving a (relatively) strong derivative, all programs tested perform equally well. In the second, more marginal, test case involving weaker derivatives, *MLPHARE* overestimates phase accuracy, while the likelihood treatments, with the refinement of error parameters, produce more reliable phase probability statistics. The ONED and TWOD functions perform satisfactorily in terms of map correlation and phase errors compared with the other target functions in the test cases shown.

The initial results are promising, but more tests need to be performed in the future. The above cases also showed that there was not much difference between the ONED and TWOD functions. This agrees with the observation by de La Fortelle & Bricogne (1997) that a one-dimensional integration over just the native phase would suffice for just isomorphous replacement without anomalous diffraction. Indeed, numerical tests (N. S. Pannu, not shown) indicate that the ONED and TWOD functions evaluate roughly to the same numbers in the case when the native $|F|/\sigma(F)$ is greater than four in the above test cases.

The equations can be generalized for correlated MAD and MIR phasing, accounting for correlations in lack-of-isomorphism errors or in errors of the heavy-atom models that are currently neglected in likelihood-based phasing methods (de La Fortelle & Bricogne, 1997). Equations have been previously developed for correlated MIR (Terwilliger & Berendzen, 1996) and MAD (Terwilliger & Berendzen, 1997) phasing, but differ in some important respects to the approach described above: Terwilliger and Berendzen derive the error model using correlations between structure factors but, to make the equations more tractable, they assume correlation in structure-factor amplitudes and a Gaussian error model in structure-factor amplitudes, as first introduced by Blow & Crick (1959).

The disadvantage of the general equations described above is that they require N integrations for the N data sets. Recently, Bricogne (2000) has presented a solution to this problem in the form of 'Generalized Bessel Functions'. In the specific case of SAD phasing, where there are only two phase integrals, an implementation resulting in only one numerical integration has been performed (Pannu & Read, 2003).

APPENDIX *A*
Covariance of real and imaginary parts of *F*

To construct a multivariate Gaussian joint probability distribution for the real and imaginary parts of the structure factors F, only the expected values and the covariance terms are needed. In the calculation of the moments, no prior knowledge of any of the atomic positions will be assumed. Thus, the expected values will be zero (*i.e.* $\langle A_i \rangle = \langle B_i \rangle = 0, \forall i$). With zero mean vectors, the covariances simplify to the following: $\langle A_i A_j \rangle$, $\langle B_i B_j \rangle$, $\langle A_i B_j \rangle$, $\langle A_j B_i \rangle$. In the analysis below, suppose that A_i and B_i are calculated from N_i atoms, A_j and B_j are determined from N_j atoms and the scattering factors, $f_j = |f_j| \exp(i\beta_j)$, are complex.

The exact expression for $\langle A_i A_j \rangle$ will be calculated first and is shown below.

$$\begin{aligned}\Big\langle \sum_{n=1}^{N_i} |f_{ni}| \cos(2\pi h \cdot x_{ni} + \beta_{ni}) \sum_{k=1}^{N_j} |f_{kj}| \cos(2\pi h \cdot x_{kj} + \beta_{kj}) \Big\rangle \\ = \frac{1}{2} \Big\langle \sum_{n=1}^{N_i} \sum_{k=1}^{N_j} |f_{ni}||f_{kj}| \cos[2\pi h \cdot (x_{ni} - x_{kj}) + \beta_{ni} - \beta_{kj}] \Big\rangle \\ + \frac{1}{2} \Big\langle \sum_{n=1}^{N_i} \sum_{k=1}^{N_j} |f_{ni}||f_{kj}| \cos[2\pi h \cdot (x_{ni} + x_{kj}) + \beta_{ni} + \beta_{kj}] \Big\rangle.\end{aligned} \tag{17}$$

The expression for $\langle B_i B_j \rangle$ is

$$\begin{aligned}\Big\langle \sum_{n=1}^{N_i} |f_{ni}| \sin(2\pi h \cdot x_{ni} + \beta_{ni}) \sum_{k=1}^{N_j} |f_{kj}| \sin(2\pi h \cdot x_{kj} + \beta_{kj}) \Big\rangle \\ = \frac{1}{2} \Big\langle \sum_{n=1}^{N_i} \sum_{k=1}^{N_j} |f_{ni}||f_{kj}| \cos[2\pi h \cdot (x_{ni} - x_{kj}) + \beta_{ni} - \beta_{kj}] \Big\rangle \\ - \frac{1}{2} \Big\langle \sum_{n=1}^{N_i} \sum_{k=1}^{N_j} |f_{ni}||f_{kj}| \cos[2\pi h \cdot (x_{ni} + x_{kj}) + \beta_{ni} + \beta_{kj}] \Big\rangle.\end{aligned} \tag{18}$$

Finally, the calculation of $\langle A_iB_j\rangle$, followed by $\langle A_jB_i\rangle$, is performed:

$$\begin{aligned}\left\langle \sum_{n=1}^{N_i} |f_{ni}| \cos(2\pi h\cdot x_{ni} + \beta_{ni}) \sum_{k=1}^{N_j} |f_{kj}| \sin(2\pi h\cdot x_{kj}) + \beta_{kj}\right\rangle \\ = \frac{1}{2}\left\langle \sum_{n=1}^{N_i}\sum_{k=1}^{N_j} |f_{ni}||f_{kj}| \sin[2\pi h\cdot(x_{ni}+x_{kj}) + \beta_{ni} + \beta_{kj}]\right\rangle \\ + \frac{1}{2}\left\langle \sum_{n=1}^{N_i}\sum_{k=1}^{N_j} |f_{ni}||f_{kj}| \sin[2\pi h\cdot(x_{ni}-x_{kj}) + \beta_{ni} - \beta_{kj}]\right\rangle,\end{aligned} \tag{19}$$

$$\begin{aligned}\left\langle \sum_{n=1}^{N_i} |f_{ni}| \sin(2\pi h\cdot x_{ni} + \beta_{ni}) \sum_{k=1}^{N_j} |f_{kj}| \cos(2\pi h\cdot x_{kj} + \beta_{kj})\right\rangle \\ = \frac{1}{2}\left\langle \sum_{n=1}^{N_i}\sum_{k=1}^{N_j} |f_{ni}||f_{kj}| \sin[2\pi h\cdot(x_{ni}+x_{kj}) + \beta_{ni} + \beta_{kj}]\right\rangle \\ - \frac{1}{2}\left\langle \sum_{n=1}^{N_i}\sum_{k=1}^{N_j} |f_{ni}||f_{kj}| \sin[2\pi h\cdot(x_{ni}-x_{kj}) + \beta_{ni} - \beta_{kj}]\right\rangle.\end{aligned} \tag{20}$$

No approximations were made in the above derivations of the covariances: only trigonometric identities and the properties of expectations were used. Thus, these expressions are exact and can be used for higher-level approximations. Hauptman & Karle (1953), Bertaut (1955) and Bricogne (1988) derived similar expressions to those shown above.

To a good approximation, all of the cross-terms cancel in the summation if the atomic positions are considered to be independent (Luzzati, 1952; Read, 1990). However, the contributions of symmetry-related atoms are identical for certain classes of reflections. The statistical effect of symmetry is accounted for by the expected intensity factor ε (Stewart & Karle, 1976; Read, 1990). If it is assumed further that there is no prior knowledge of the atomic positions, the covariances can be approximated as below:

$$\begin{aligned}\langle A_iA_j\rangle &\approx \frac{1}{2}\varepsilon \sum_{n=1}^{N_j} |f_{ni}||f_{nj}|\langle\cos[2\pi h\cdot(x_{ni}-x_{nj}) + \beta_{ni} - \beta_{nj}]\rangle, \\ \langle B_iB_j\rangle &\approx \frac{1}{2}\varepsilon \sum_{n=1}^{N_j} |f_{ni}||f_{nj}|\langle\cos[2\pi h\cdot(x_{ni}-x_{nj}) + \beta_{ni} - \beta_{nj}]\rangle, \\ \langle A_iB_j\rangle &\approx \frac{1}{2}\varepsilon \sum_{n=1}^{N_j} |f_{ni}||f_{nj}|\langle\sin[2\pi h\cdot(x_{ni}-x_{nj}) + \beta_{ni} - \beta_{nj}]\rangle, \\ \langle A_jB_i\rangle &\approx -\frac{1}{2}\varepsilon \sum_{n=1}^{N_j} |f_{ni}||f_{nj}|\langle\sin[2\pi h\cdot(x_{ni}-x_{nj}) + \beta_{ni} - \beta_{nj}]\rangle.\end{aligned} \tag{21}$$

APPENDIX *B* The covariance matrix Σ

The covariances for the complex multivariate normal distribution of structure factors, σ_{ij}, will be calculated from first principles. For applications to experimental phasing, molecular replacement and refinement, it will be assumed that structure factors with different values of Miller indices *hkl* are independent. Probability distributions relating arbitrary sets of structure factors can be cast in the formulation described and the covariances between structure factors with different Miller indices capture the information exploited by density modification and direct methods. However, these covariances are generally much smaller than those between structure factors with the same Miller indices.

In this analysis, differences in the atomic parameters and errors in the measurement of structure factors will be considered. A complex measurement error in the structure factors will be assumed (Green, 1979). Although it seems unintuitive to ascribe a complex error to a scalar quantity, this approximation has proven to work well in phasing (de La Fortelle & Bricogne, 1997) and model refinement (Murshudov *et al.*, 1997). Suppose that F_i is calculated from N_i atoms and F_j is determined from N_j atoms. As performed above, without loss of generality, assume $N_j \geq N_i$, because the *j* structure can be considered to contain null atoms. The complex measurement error will be represented by γ,

$$\begin{aligned}F_i^+ &= \sum_{n=1}^{N_i}(f_{ni} + if''_{ni})\exp(2\pi ih\cdot x_{ni}) + \gamma_i^+, \\ (F_j^-)^* &= \sum_{n=1}^{N_j}(f_{nj} - if''_{nj})\exp(2\pi ih\cdot x_{nj}) + \gamma_j^-.\end{aligned} \tag{22}$$

The complex conjugate of the F^- term is chosen just so that the same Miller indices are compared between structure factors. F_j^+ and $(F_i^-)^*$ are defined in an analogous fashion. Again, common atoms between the two structures will be represented with the same first subscript and it will be assumed that the measurement error for a particular structure factor is independent of all other measurement errors and independent of model errors. Furthermore, the variance $\langle\gamma_i^2\rangle$ will be denoted by $\sigma_{F_i}^2$.

First the covariance between F_i^+ and F_j^+ is calculated and shown below.

$$\begin{aligned}\left\langle\left[\sum_{n=1}^{N_i}(f_{ni} + if''_{ni})\exp(2\pi ih\cdot x_{ni}) + \gamma_i^+\right] \right.\\ \left.\times \left[\sum_{k=1}^{N_j}(f_{kj} - if''_{kj})\exp(-2\pi ih\cdot x_{kj}) + \gamma_j^+\right]\right\rangle \\ = \sum_{n=1}^{N_i}\sum_{k=1}^{N_j}(f_{ni} + if''_{ni})(f_{kj} - if''_{kj}) \\ \times \langle\exp[2\pi ih\cdot(x_{ni} - x_{kj})]\rangle + \langle\gamma_i^+\gamma_j^+\rangle.\end{aligned} \tag{23}$$

[Note that all the cross-terms involving $\langle\gamma_i^+\exp(-2\pi ih{\cdot}x_{kj})\rangle$ and $\langle\gamma_j^+\exp(-2\pi ih{\cdot}x_{ni})\rangle$ cancel, because of independence.] As discussed above, the cross-terms between independent atoms will tend to cancel and the expected intensity factor ε can be introduced to account for the statistical effects of symmetry.

$$\langle\exp[2\pi ih\cdot(x_{ni} - x_{kj})]\rangle = \begin{cases}\varepsilon d_{kij} & \text{if } n = k\ , \\ 0 & \text{if } n \neq k\ .\end{cases} \tag{24}$$

d_{kij} is the Fourier transform of the probability distribution for differences in atomic positions and in general can be complex. However, by assuming a centre of symmetry in the probability distribution for coordinate differences, the simplest example being an isotropic Gaussian (Luzzati, 1952; Read, 1990), the imaginary term disappears. Thus, in practice d_{kij} will be treated as real-valued. Therefore, $\langle F_i^+(F_j^+)^*\rangle$ reduces to

$$\sum_{k=1}^{N_i} \varepsilon d_{kij}[f_{ki}f_{kj} + f''_{ki}f''_{kj} + i(f''_{ki}f_{kj} - f_{ki}f''_{kj})] + \langle \gamma_i^+\gamma_j^+\rangle. \qquad (25)$$

When $i \neq j$, the term $\langle \gamma_i^+\gamma_j^+\rangle$ disappears, because the measurement errors are assumed to be independent, but when $i = j$ the above expression simplifies to

$$\langle |F_i^+|^2\rangle = \sum_{k=1}^{N_i} \varepsilon(f_{ki}^2 + f''^2_{ki}) + \sigma^2_{F_i^+}. \qquad (26)$$

Similarly, $\langle (F_i^-)^* F_j^-\rangle$ can be found to be

$$\begin{cases} \sum_{k=1}^{N_i} \varepsilon d_{kij}[f_{ki}f_{kj} + f''_{ki}f''_{kj} + i(f''_{ki}f_{kj} - f_{ki}f''_{kj})] & \text{if } i \neq j, \\ \sum_{k=1}^{N_i} \varepsilon(f_{ki}^2 + f''^2_{ki}) + \sigma^2_{F_i^-} & \text{if } i = j. \end{cases} \qquad (27)$$

Finally, the calculation of $\langle F_i^+[(F_j^-)^*]^*\rangle = \langle F_i^+(F_j^-)\rangle$:

$$\Bigg\langle\Bigg(\sum_{n=1}^{N_i}(f_{ni} + if''_{ni})\exp(2\pi ih\cdot x_{ni}) + \gamma_i^+\Bigg) \times \Bigg(\sum_{k=1}^{N_j}(f_{kj} + if''_{kj})\exp(-2\pi ih\cdot x_{kj}) + \gamma_j^-\Bigg)\Bigg\rangle = \sum_{k=1}^{N_i} \varepsilon d_{kij}[f_{ki}f_{kj} - f''_{ki}f''_{kj} + i(f''_{ki}f_{kj} + f_{ki}f''_{kj})] + \langle \gamma_i^+\gamma_j^-\rangle. \qquad (28)$$

Because the measurement errors are uncorrelated, the term $\langle \gamma_i^+\gamma_j^-\rangle$ disappears in the above equation. In general, the covariance matrix has real diagonal terms (*i.e.* $\sigma_{ii} \in \mathbb{R}, \forall i$) and complex off-diagonal terms.

The approach described above differs from the treatment of Giacovazzo & Siliqi (2001*a*,*b*) in that individual sources of errors (*e.g.* lack of isomorphism, atomic parameter and measurement errors) are modelled separately (as opposed to a cumulative error; Terwilliger & Eisenberg, 1987) and the effect of correlated errors is considered in all but the measurement errors.

APPENDIX *C*
The MIR likelihood function

The required likelihood target for multiple isomorphous phasing is the joint distribution of the measured amplitudes given the heavy-atom model and parameters describing the variance. To obtain this from the joint distribution of structure factors, both the unknown phase angles and the dummy structure factor must be eliminated. After a change to polar coordinates in each of the univariate distributions in the product, the unknown phase is integrated to obtain the following:

$$P(|F_j|;|F|,\alpha,H_j) = \frac{2|F_j|}{\varepsilon\sigma^2_{\Delta_j} + \sigma^2_{F_j}}\exp\left(-\frac{|F_j|^2 + |F_{jc}|^2}{\varepsilon\sigma^2_{\Delta_j} + \sigma^2_{F_j}}\right) I_0\left(\frac{2|F_j||F_{jc}|}{\varepsilon\sigma^2_{\Delta_j} + \sigma^2_{F_j}}\right). \qquad (29)$$

In the above equation, F_{jc} is the expected value of F_j. To eliminate the dummy structure factor from the conditional distribution, we first convert this to a joint distribution by multiplying by the prior probability of the dummy structure factor:

$$P(|F|,\alpha,\{|F_j|\}_{j=0\ldots n};\{H_j\}_{j=0\ldots n}) = P(|F|,\alpha)\prod_{j=0}^{n} P(|F_j|;\,|F|,\alpha,H_j). \qquad (30)$$

In this equation, $P(|F|, \alpha)$ is a Wilson (1949) distribution expressed in terms of amplitude and phase. Finally, the required likelihood target is obtained by integrating over the amplitude and phase of the dummy structure factor,

$$P(\{|F_j|\}_{j=0\ldots n};\,\{H_j\}_{j=0\ldots n}) = \int_0^\infty\int_0^{2\pi} P(|F|,\alpha)\prod_{j=0}^{n} P(|F_j|;\,|F|,\alpha,H_j)\,\mathrm{d}\alpha\,\mathrm{d}|F|. \qquad (31)$$

The centric case can be treated in a similar fashion, but starting from a multivariate normal distribution of real structure factors and ending with the one-dimensional integral

$$P(\{|F_j|\}_{j=0\ldots n};\,\{H_j\}_{j=0\ldots n}) = \int_0^\infty \sum_{\alpha=\alpha_r,\alpha_r+\pi} P(|F|,\alpha)\prod_{j=0}^{n} P(|F_j|;\,|F|,\alpha,H_j)\,\mathrm{d}|F|. \qquad (32)$$

In the above equation, α_r and $\alpha_r = \pi$ are the two phases the centric reflection is restricted to and $P(F_j|;\,|F|,\alpha,H_j)$ is

$$P(|F_j|;\,|F|,\alpha,H_j) = \left[\frac{2}{\pi(\varepsilon\sigma^2_{\Delta_j} + \sigma^2_{F_j})}\right]^{1/2}\exp\left[-\frac{|F_j|^2 + |F_{jc}|^2}{2(\varepsilon\sigma^2_{\Delta_j} + \sigma^2_{F_j})}\right] \times \cosh\left(\frac{|F_j||F_{jc}|}{\varepsilon\sigma^2_{\Delta_j} + \sigma^2_{F_j}}\right). \qquad (33)$$

(31) and (32) form the TWOD likelihood function discussed in the paper. The integrals over the structure-factor amplitude in both the acentric and centric cases can be avoided by assuming the observed native structure-factor amplitude is exact with no measurement error (*i.e.* $|F| = |F_o|$), thus resulting in only a phase integration (acentric case) or summation (centric case). The likelihood function avoiding the integral over the structure-factor amplitude is denoted ONED in the paper.

We thank J. P. Abrahams, R. A. G. de Graaff, A. G. W. Leslie, J. R. Plaisier, T. Terwilliger, the referees and the editor for valuable suggestions. Funding was provided by Leiden University, Trinity College (Cambridge, UK) and the Wellcome Trust.

References

Bertaut, E. F. (1955). *Acta Cryst.* **8**, 537–543.
Blow, D. M. & Crick, F. H. C. (1959). *Acta Cryst.* **12**, 794–802.
Bricogne, G. (1984). *Acta Cryst.* A**40**, 410–445.

Bricogne, G. (1988). *Acta Cryst.* A**44**, 517–545.

Bricogne, G. (1991). *Isomorphous Replacement and Anomalous Scattering. Proceedings of the CCP4 Study Weekend*, edited by W. Wolf, P. R. Evans and A. G. W. Leslie, pp. 60–68. Warrington: Daresbury Laboratory.

Bricogne, G. (1997). *Methods Enzymol.* **276**, 361–423.

Bricogne, G. (2000). *Advanced Special Functions and Applications: Proceedings of the Melfi School on Advanced Topics in Mathematics and Physics*, edited by D. Cocolicchio, G. Dattoli & H. M. Srivastava, pp. 315–323. Rome: Aracne Editrice.

Burla, M. C., Carrozzini, B., Cascarano, G. L., Giacovazzo, C., Polidori, G. & Siliqi, G. (2002). *Acta Cryst.* D**58**, 928–935.

Collaborative Computational Project, Number 4 (1994). *Acta Cryst.* D**50**, 760–763.

Giacovazzo, C. (1998). *Direct Phasing in Crystallography: Fundamentals and Applications.* Oxford University Press.

Giacovazzo, C. & Siliqi, G. (1983). *Acta Cryst.* A**39**, 585–592.

Giacovazzo, C. & Siliqi, G. (2001*a*). *Acta Cryst.* A**57**, 40–46.

Giacovazzo, C. & Siliqi, G. (2001*b*). *Acta Cryst.* A**57**, 700–707.

Green, E. A. (1979). *Acta Cryst.* A**35**, 351–359.

Grimmett, G. R & Stirzaker, D. R.(1992). *Probability and Random Processes.* Oxford: Clarendon Press.

Hauptman, H. (1982). *Acta Cryst.* A**38**, 632–641.

Hauptman, H. & Karle, J. (1953). *The Solution of the Phase Problem. I: The Centrosymmetric Crystal. ACA Monograph No. 3.* New York: Polycrystal.

Herzberg, O. & James, M. N. G. (1985). *Nature (London)*, **313**, 653–659.

Johnson, R. A. & Wichern, D. W. (1992). *Applied Multivariate Statistical Analysis.* New Jersey: Prentice–Hall Inc.

Klug, A. (1958). *Acta Cryst.* **11**, 515–543.

La Fortelle, E. de & Bricogne, G. (1997). *Methods Enzymol.* **276**, 472–494.

Lunin, V. Y., Lunina, N. L., Petrova, T. E., Urzhumtsev, A. G. & Podjarny, A. D. (1998). *Acta Cryst.* D**54**, 726–734.

Luzzati, V. (1952). *Acta Cryst.* **5**, 802–810.

Murshudov, G. N., Vagin, A. A. & Dodson, E. J. (1997). *Acta Cryst.* D**53**, 240–255.

Otwinowski, Z. (1991). *Proceedings of the CCP4 Study Weekend. Isomorphous Replacement and Anomalous Scattering*, edited by W. Wolf, P. R. Evans & A. G. W. Leslie, pp. 80–86. Warrington: Daresbury Laboratory.

Pannu, N. S. (1997). Masters thesis. University of Alberta.

Pannu, N. S. & Read, R. J. (2002). *Acta Cryst.* A**58**, C239.

Pannu, N. S. & Read, R. J. (2003). Submitted.

Read, R. J. (1990). *Acta Cryst.* A**46**, 900–912.

Read, R. J. (1991). *Proceedings of the CCP4 Study Weekend. Isomorphous Replacement and Anomalous Scattering*, edited by W. Wolf, P. R. Evans & A. G. W. Leslie, pp. 69–79. Warrington: Daresbury Laboratory.

Read, R. J. (2001). *Acta Cryst.* D**57**, 1373–1382.

Shmueli, U. & Weiss, G. H. (1991). *Introduction to Crystallographic Statistics.* Oxford University Press.

Skarzynski, T. (1992). *J. Mol. Biol.* **224**, 671–683.

Stewart, J. M. & Karle, J. (1976). *Acta Cryst.* A**32**, 1005–1007.

Terwilliger, T. C. & Berendzen, J. (1996). *Acta Cryst.* D**52**, 749–757.

Terwilliger, T. C. & Berendzen, J. (1997). *Acta Cryst.* D**53**, 571–579.

Terwilliger, T. C. & Eisenberg, D. (1987). *Acta Cryst.* A**43**, 6–13.

Wilson, A. J. C. (1949). *Acta Cryst.* **2**, 318–321.

Wooding, R. A. (1956). *Biometrika*, **43**, 212-215.