Acta Crystallographica Section D
Biological Crystallography
ISSN 0907-4449

Simple algorithm for a maximum-likelihood SAD function

Airlie J. McCoy, Laurent C. Storoni and Randy J. Read*

University of Cambridge, Department of Haematology, Cambridge Institute for Medical Research, Wellcome Trust/MRC Building, Hills Road, Cambridge CB2 2XY, England

Correspondence e-mail: rjr27@cam.ac.uk

Received 18 December 2003
Accepted 23 April 2004

Recently, the multivariate complex normal distribution has been used to develop a maximum-likelihood probability function for single-wavelength anomalous diffraction phasing and refinement of heavy-atom parameters [Pannu & Read (2004), *Acta Cryst.* D**60**, 22–27]. The function accounts explicitly for the correlations between the observed and calculated Friedel mates and their errors. However, the method of derivation of the equation described by Pannu & Read (2004) leads to a complicated likelihood expression that suffers from a number of algorithmic limitations. Here, an alternative derivation of the $P_{\rm SAD}$ function is described that leads to simplified algorithmic requirements and that allows an intuitive understanding of the expression.

1. Introduction

The availability of tuneable synchrotron sources allowed the development of multiple-wavelength anomalous diffraction (MAD; Hendrickson, 1991) phasing experiments, which today underpin many high-throughput structural biology efforts around the world. With improvements in synchrotron sources, cryocooling of crystals and increased detector sensitivity, phasing by single-wavelength anomalous diffraction (SAD) has become not only feasible, but in some cases preferable to phasing by MAD, particularly where radiation damage is significant (Rice *et al.*, 2000; Dodson, 2003) or where the absorption edge for the anomalous scatterer is not accessible (*e.g.* sulfur, xenon). However, until recently technical improvements in the SAD experiment had not been matched by corresponding improvements in the theory for obtaining phases from SAD.

A maximum-likelihood treatment of the SAD phasing problem describes the probability distribution $P_{\rm SAD}$ of the (unphased) model structure factors F^+ and F^- given the (phased) calculated heavy-atom structure factors $\mathbf{H}^+$ and $\mathbf{H}^-$,

$$P_{\rm SAD} = P(F^+, F^-|\mathbf{H}^+, \mathbf{H}^{-*}),$$

where $F^+ = |\mathbf{F}^+|$ and $F^- = |\mathbf{F}^-|$. F^+ and F^- are highly correlated and so $P_{\rm SAD}$ cannot be approximated by a product of independent probabilities for the two observations F^+ and F^-. Also highly correlated are the substructure-model errors contributing to the conditional probability distribution of F^+ and F^-, since they are generated by the same set of anomalous scatterers. These correlations must be included in the probability distribution for a complete analysis.

Traditional methods for SAD phasing have avoided the complication of including the correlations by using the mean F and the Bijvoet difference (F and $\Delta F^{\pm}$) rather than F^+ and F^-, as these are relatively independent and have relatively inde-

pendent errors. In these treatments, the distribution of Bijvoet differences has been assumed to be Gaussian (North, 1965; Matthews, 1966; de La Fortelle & Bricogne, 1997). More recently, joint probability distributions for F^+ and F^- have been described that go some way towards addressing the problem (Hauptman, 1982; Giacovazzo, 1983; Burla *et al.*, 2002; Giacovazzo & Siliqi, 2001*a*,*b*; Terwilliger & Eisenberg, 1987), but it was not until Pannu & Read (2004) that a $P_{\rm SAD}$ function was described that accounted explicitly for the correlations in the SAD experiment,

$$\begin{aligned} P_{\rm SAD} = {} & \frac{2F^+F^-|\Sigma_2|}{\pi|\Sigma_4|}\exp[-a_{11}F^{+2} - a_{22}F^{-2} \\ & - (a_{33}-c_{33})H^{+2} - (a_{44}-c_{44})H^{-2}] \\ & \times \exp\{-2H^+H^-[(a_{34}-c_{34})\cos(\alpha_H^+ - \alpha_H^-) \\ & - (b_{34}-d_{34})\sin(\alpha_H^+ - \alpha_H^-)]\} \\ & \times \int_0^{2\pi} \big(\exp\{-2F^-H^+[a_{23}\cos(\alpha^- - \alpha_H^+) \\ & - b_{23}\sin(\alpha^- - \alpha_H^+)]\} \\ & \times \exp\{-2F^-H^-[a_{24}\cos(\alpha^- - \alpha_H^-) \\ & - b_{24}\sin(\alpha^- - \alpha_H^-)]\} I_0(\xi^{1/2})\, {\rm d}\alpha^-\big), \end{aligned} \tag{1}$$

$$\begin{aligned} \xi = {} & 4F^{+2}[a_{12}F^-\cos(\alpha^-) + b_{12}F^-\sin(\alpha^-) \\ & + a_{13}H^+\cos(\alpha_H^+) + b_{13}H^+\sin(\alpha_H^+) \\ & + a_{14}H^-\cos(\alpha_H^-) + b_{14}H^-\sin(\alpha_H^-)]^2 \\ & + 4F^{+2}[a_{12}F^-\sin(\alpha^-) - b_{12}F^-\cos(\alpha^-)\,' \\ & + a_{13}H^+\sin(\alpha_H^+) - b_{13}H^+\cos(\alpha_H^+) \\ & + a_{14}H^-\sin(\alpha_H^-) - b_{14}H^-\cos(\alpha_H^-)]^2 \end{aligned}$$

$$\Sigma_4^{-1} = \begin{pmatrix} a_{11} & a_{12}+ib_{12} & a_{13}+ib_{13} & a_{14}+ib_{14} \\ a_{12}-ib_{12} & a_{22} & a_{23}+ib_{23} & a_{24}+ib_{24} \\ a_{13}-ib_{13} & a_{23}-ib_{23} & a_{33} & a_{34}+ib_{34} \\ a_{14}-ib_{14} & a_{24}-ib_{24} & a_{34}-ib_{34} & a_{44} \end{pmatrix},$$

$$\Sigma_2^{-1} = \begin{pmatrix} c_{11} & c_{12}+id_{12} \\ c_{12}-id_{12} & c_{22} \end{pmatrix},$$

where Σ_4 is the (Hermitian) covariance matrix of the tetravariate complex Gaussian distribution $P(\mathbf{F}^+, \mathbf{F}^{-*}, \mathbf{H}^+, \mathbf{H}^{-*})$, Σ_2 is the (Hermitian) covariance matrix of the bivariate Gaussian complex distribution $P(\mathbf{H}^+, \mathbf{H}^{-*})$ and α^-, α_H^+ and α_H^- are the phases of $\mathbf{F}^{-*}$, $\mathbf{H}^+$ and $\mathbf{H}^{-*}$, respectively. It is assumed that the reflections are independent, so the total likelihood is the product of the reflection likelihoods.

The complexity of (1) is immediately apparent. There are 20 different coefficients arising from the inverse of the covariance matrices Σ_4 (ten real, six imaginary) and Σ_2 (three real, one imaginary). During refinement Σ_4 and Σ_2 must be kept positive definite and in the implementation of the $P_{\rm SAD}$ function described by Pannu & Read (2004) this was performed by setting negative eigenvalues to zero during calculation of their inverses by singular value decomposition. The derivatives of the function become even more verbose. In the implementation described by Pannu & Read (2004), derivatives were not calculated analytically. Instead, an automatic differentiation method (*ADOLC*; Griewank *et al.*, 1996) was used to obtain the gradient vectors. The complex functional form of (1) makes it difficult to get an intuitive feel for the effects of the different parameters or the physical meaning of the terms.

Here, we present an alternative derivation of a maximum-likelihood $P_{\rm SAD}$ function that has only three unique error parameters, does not involve matrix inversion, allows analytic derivatives to be calculated easily and provides an intuitive understanding of the SAD experiment.

2. Results

2.1. SAD likelihood function

Equation (1) was derived by finding the expression for $P(\mathbf{F}^+, \mathbf{F}^{-*}, \mathbf{H}^+, \mathbf{H}^{-*})$, integrating out the unknown phases to obtain the joint probability $P(F^+, F^-, \mathbf{H}^+, \mathbf{H}^{-*})$ and then fixing the calculated structure factors and renormalizing to obtain the desired conditional probability $P(F^+, F^-|\mathbf{H}^+, \mathbf{H}^{-*})$. If, instead, the order of the operations is reversed and the conditional probability $P(\mathbf{F}^+, \mathbf{F}^{-*}|\mathbf{H}^+, \mathbf{H}^{-*})$ is formed before integrating out the unknown phases, we obtain (Appendix *A*) the expression

$$\begin{aligned} P_{\rm SAD} = {} & \frac{2F^+F^-}{\pi\varepsilon^2(1-D_\Phi^2)\sigma_\Delta^4}\int_0^{2\pi}\exp\Bigg[-\frac{|F^-\exp(i\alpha^-) - D\mathbf{H}^{-*}|^2}{\varepsilon\sigma_\Delta^2} \\ & - \frac{F^{+2}+F_C^{+2}}{\varepsilon(1-D_\Phi^2)\sigma_\Delta^2}\Bigg] I_0\Bigg[\frac{2F^+F_C^+}{\varepsilon(1-D_\Phi^2)\sigma_\Delta^2}\Bigg]\,{\rm d}\alpha^-, \end{aligned} \tag{2}$$

where

$$F_C^+ = |D\mathbf{H}^+ + D_\Phi\exp(i\alpha_\Phi)[F^-\exp(i\alpha^-) - D\mathbf{H}^{-*}]|.$$

This equation contains three error parameters derived from the initial covariance matrix (σ_Δ, D_Φ and α_Φ). Again, it is assumed that the reflections are independent so that the total likelihood is the product of the reflection probabilities.

(2) was derived by integrating out the phase α^+ analytically, leaving the integration over α^- to be performed numerically. Equivalently, the phase α^- could have been integrated out analytically, leaving the integration over α^+ to be performed numerically. Numerical integration tests comparing these two forms of the equation confirm that they give the same values for $P_{\rm SAD}$ (data not shown).

2.2. Phase probabilities and maps

$P_{\rm SAD}$ is obtained by integrating $P(F^+, F^-, \alpha^-|\mathbf{H}^+, \mathbf{H}^{-*})$ over α^-. The conditional probability distribution of α^- can be obtained by fixing F^+ and F^- in the joint distribution $P(F^+, F^-, \alpha^-|\mathbf{H}^+, \mathbf{H}^{-*})$ and renormalizing to obtain $P(\alpha^-|F^+, F^-, \mathbf{H}^+, \mathbf{H}^{-*})$. In other words, the probability distribution for this phase is proportional to the integrand in (2). The roles of F^+ and F^- can be reversed to obtain the probability distribution for α^+.

For building an atomic model into electron density one is generally most interested in the map representing the normal (real) scattering component, although the map representing the imaginary component is often useful as well. When the

Table 1
Statistics for SAD refinement and phasing of a Z-form DNA hexamer duplex.

	MLPHARE†	*SOLVE*‡	*SHARP*§	*PHASER*¶
360° pass				
Map correlation††	0.607	0.588	0.722	0.723
Reported figure of merit††	0.587	0.492	0.575	0.650
Mean cos(phase error)††	0.500	0.553	0.634	0.643
Mean phase error††	53.53	50.52	42.90	41.64
90° pass				
Map correlation††	0.500	0.487	0.643	0.649
Reported figure of merit††	0.405	0.352	0.443	0.561
Mean cos(phase error)††	0.416	0.484	0.548	0.568
Mean phase error††	59.67	55.23	49.49	47.55

† Coordinates and isotropic *B* factors were refined. Occupancies were not refined. ‡ Coordinates, isotropic *B* factors and occupancies were refined. The minimum allowed *B* factor was zero. § Coordinates, isotropic *B* factors and the global and local imperfection parameters on anomalous differences were refined. ¶ Coordinates, isotropic *B* factors, occupancies and variance parameters were refined. †† Statistics calculated with *SFTOOLS* (B. Hazes, unpublished work; Collaborative Computational Project, Number 4, 1994). Map correlation compared the figure-of-merit-weighted map from experimental phasing with the figure-of-merit-weighted *SIGMAA* (Read, 1986) map calculated with phases from the final model.

relative contribution of the imaginary component of the anomalous scatterers is small, a map computed using either the centroid (figure-of-merit-weighted) estimate of $\mathbf{F}^{+}$ or the centroid estimate of $\mathbf{F}^{-*}$ (making the usual assumption in the map calculation that Friedel's law applies) will differ little from the map representing the real component of the electron density. However, in the presence of very strong anomalous scatterers the phases of $\mathbf{F}^{+}$ and $\mathbf{F}^{-*}$ will differ significantly. Therefore, for generality it is better either to compute a complex electron-density map by providing separate coefficients for $\mathbf{F}^{+}$ and $\mathbf{F}^{-}$ or to compute separate real and imaginary electron-density maps with coefficients obtained from figure-of-merit-weighted $(\mathbf{F}^{+} + \mathbf{F}^{-*})/2$ and $\exp(-\pi i/2)(\mathbf{F}^{+} - \mathbf{F}^{-*})/2$, respectively.

2.3. Implementation and test cases

The $P_{\rm SAD}$ function described above, with slight modifications for numerical stability and the inclusion of the effect of experimental errors (Appendix *B*), was implemented in the program *PHASER*. Analytic derivatives were used to calculate the gradients. Optimal anomalous scatterer and error parameters were found by minimizing the minus log-likelihood.

Results of the implementation in *PHASER* were compared with results from the programs *MLPHARE* (version 4.0; Otwinowski, 1991; Collaborative Computational Project, Number 4, 1994), *SOLVE* (version 2.02; Terwilliger & Berendzen, 1997) and *SHARP* (version 2.0; de La Fortelle & Bricogne, 1997). Tests were performed with the two publicly available data sets used by Pannu & Read (2004): the 90° and the 360° pass data sets of a Z-form DNA hexamer duplex phased on ten intrinsic P atoms (Dauter & Adamiak, 2001). The results (Table 1) for *MLPHARE* and *SOLVE* were comparable to those reported by Pannu & Read (2004), but the results for *SHARP* were significantly better, as instead of using the default refinement protocol, the refinement protocol was customized to the test case. Statistics for the implementation of $P_{\rm SAD}$ in *PHASER* were not significantly different from those reported for the $P_{\rm SAD}$ function implemented in Pannu & Read (2004), confirming that when the parameters have been optimized (1) and (2) give very similar final phase distributions.

3. Discussion

The $P_{\rm SAD}$ expression described in (2) is simpler than that in (1). It has several algorithmic advantages: the parameterization is compact, refinement of heavy-atom parameters does not involve the inversion of covariance matrices, and analytic derivatives can be determined easily. It is thus likely to be much more robust when applied to a wide range of SAD data sets.

In general, a maximum-likelihood approach in crystallography is of greatest benefit when the data and the model are poor. This is clearly seen in the test cases, where including the correlations has a significant influence on the determination of the figure of merit in the poorer (90°) data set, but little effect in the better (360°) data set. The figure of merit reported by *PHASER* for the poorer (90°) data set is closer to the mean cosine of the phase error than that produced by the other three programs. This suggests that the $P_{\rm SAD}$ function gives better phase probability distribution estimates for use in density modification (required to break the phase ambiguity present in SAD phasing) when the phasing is marginal.

The $P_{\rm SAD}$ function can also be used for the refinement of models containing anomalous scatterers (Garib Murshudov, personal communication). In model refinement, fast calculation of the target function is of key importance as other aspects of the algorithm are already time-consuming given the large number of atomic parameters (*e.g.* the structure-factor calculation). The reduced parameterization for $P_{\rm SAD}$ should also be helpful for this application.

The new formulation of $P_{\rm SAD}$ also allows a more intuitive understanding of the SAD likelihood function. As shown in the appendices, $P_{\rm SAD}$ can be expressed as the integral of the product of two functions,

$$\begin{aligned} P_{\rm SAD} &= P(F_O^{+}, F_O^{-}|\mathbf{H}^{+}, \mathbf{H}^{-*}) \\ &= \int_0^{2\pi} P(F_O^{-}, \alpha^{-}|\mathbf{H}^{+}, \mathbf{H}^{-*}) P(F_O^{+}|F_O^{-}, \alpha^{-}, \mathbf{H}^{+}, \mathbf{H}^{-*})\, {\rm d}\alpha^{-}, \end{aligned} \tag{3}$$

where

$$\begin{aligned} P(F_O^{-}, \alpha^{-}|\mathbf{H}^{+}, \mathbf{H}^{-*}) &= P(F_O^{-}, \alpha^{-}|\mathbf{H}^{-*}) \\ &= \frac{F_O^{-}}{\pi\Sigma^{-}} \exp\left[\frac{-|F_O^{-}\exp(i\alpha^{-}) - D\mathbf{H}^{-*}|^2}{\Sigma^{-}}\right], \end{aligned}$$

$$P(F_O^{+}|F_O^{-}, \alpha^{-}, \mathbf{H}^{+}, \mathbf{H}^{-*}) = \frac{2F_O^{+}}{\Sigma^{+}} \exp\left[-\frac{(F_O^{+} - F_C^{+})^2}{\Sigma^{+}}\right] e I_0\left(\frac{2F_O^{+}F_C^{+}}{\Sigma^{+}}\right),$$

$$F_C^{+} = |D\mathbf{H}^{+} + D_{\Phi}\exp(i\alpha_{\Phi})[F_O^{-}\exp(i\alpha^{-}) - D\mathbf{H}^{-*}]|.$$

In this version of the expression for $P_{\rm SAD}$, the variances Σ^+ and Σ^- have been inflated (as discussed in Appendix *B*) to account for the effect of experimental error. The first distribution in the product expresses what is known about one observation, F_O^-, when only the corresponding calculated structure factor $\mathbf{H}^-$ is given; accordingly, its variance Σ^- accounts for what is left unexplained by $\mathbf{H}^-$. (Once $\mathbf{H}^-$ is known, no further information about F_O^- is added by the knowledge of $\mathbf{H}^+$, so this part of the distribution does not depend on $\mathbf{H}^+$.) The second distribution expresses what is

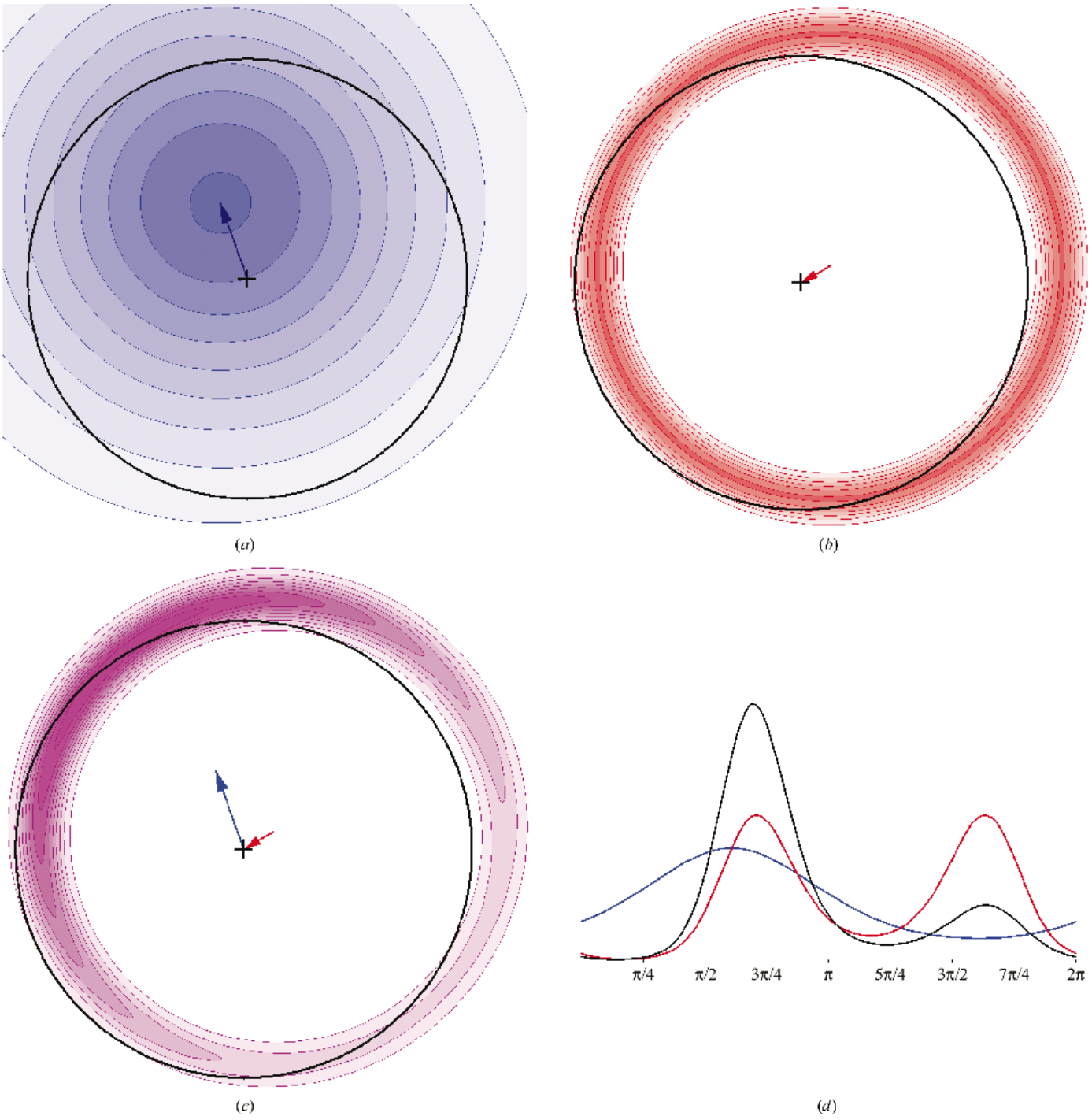

Figure 1
Schematic illustration of $P_{\rm SAD}$ for the case of SAD phasing. The three contour plots (*a*)–(*c*) are shown as a function of the assumed complex value of $\mathbf{F}^{-*}$; in each contour plot, the cross indicates the origin and the black circle indicates the measured value of F_O^- for which the function values shown in (*d*) are taken. (*a*) The first (Sim) component of $P_{\rm SAD}$, $P(\mathbf{F}^{-*}|\mathbf{H}^{-*})$, is shown in blue contours centred on $\mathbf{H}^{-*}$ (blue arrow). (*b*) The second component of $P_{\rm SAD}$, $P(F_O^+|\mathbf{F}^{-*}, \mathbf{H}^+, \mathbf{H}^{-*})$, is shown in red contours centred on the expected vector difference between $\mathbf{F}^+$ and $\mathbf{F}^{-*}$ (tail of red arrow). (*c*) The product of the two components of $P_{\rm SAD}$ is shown in magenta contours. $P_{\rm SAD}$ is given by the integral of this surface under the black circle. (*d*) The components of $P_{\rm SAD}$ are shown as a function of the assumed value of α^-, with $P(F_O^-, \alpha^- |\mathbf{H}^+, \mathbf{H}^{-*})$ shown in blue, $P(F_O^+|F_O^-, \alpha^-, \mathbf{H}^+, \mathbf{H}^{-*})$ shown in red and their product in black. The three distributions have been normalized to place them on a common scale.

known about the second observation, F_O^+, when F_O^- (phased by some value of the variable of integration, α^-) and both calculated structure factors are given; accordingly, its variance Σ^+ accounts for what is left unexplained by the value of $\mathbf{F}^+$ predicted from the other three structure factors. To a good approximation, the first distribution provides a 'Sim factor' to account for the information given by the partial structure (primarily normal scattering), while the second distribution

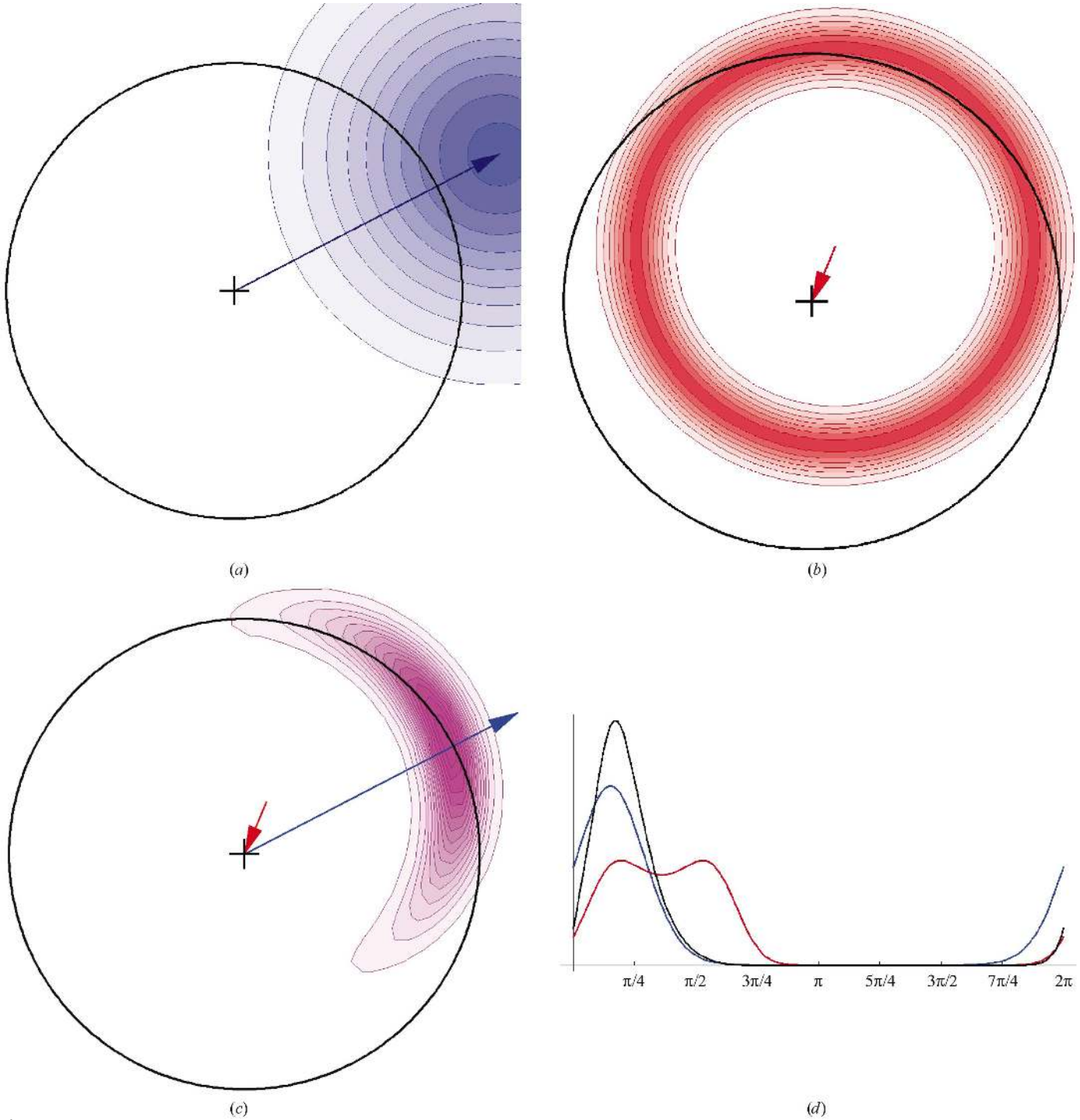

Figure 2
Schematic illustration of $P_{\rm SAD}$ for the case of model refinement against the SAD function. The three contour plots (*a*)–(*c*) are shown as a function of the assumed complex value of $\mathbf{F}^{-*}$; in each contour plot the cross indicates the origin and the black circle indicates the measured value of F_Q^- for which the function values shown in (*d*) are taken. (*a*) The first (Sim) component of $P_{\rm SAD}$, $P(\mathbf{F}^{-*}|\mathbf{H}^{-*})$, is shown in blue contours centred on $\mathbf{H}^{-*}$ (blue arrow). Compared with the SAD phasing case, the full scattering model is more complete, which increases the magnitude of $\mathbf{H}^{-*}$ and decreases the variance in this distribution. (*b*) The second component of $P_{\rm SAD}$, $P(F_O^+|\mathbf{F}^{-*}, \mathbf{H}^+, \mathbf{H}^{-*})$, is shown in red contours centred on the expected vector difference between $\mathbf{F}^+$ and $\mathbf{F}^{-*}$ (tail of red arrow). (*c*) The product of the two components of $P_{\rm SAD}$ is shown in magenta contours. $P_{\rm SAD}$ is given by the integral of this surface under the black circle. (*d*) The components of $P_{\rm SAD}$ are shown as a function of the assumed value of α^-, with $P(F_O^-, \alpha^- |\mathbf{H}^+, \mathbf{H}^{-*})$ shown in blue, $P(F_O^+|F_O^-, \alpha^-, \mathbf{H}^+, \mathbf{H}^{-*})$ shown in red and their product in black. The three distributions have been normalized to place them on a common scale.

takes account of the anomalous difference. While the mathematical details differ considerably, the SAD phasing function presented recently by Giacovazzo *et al.* (2003) also combines a term arising from anomalous differences with a Sim-like term. Note that when expressed using the exponential Bessel function (eI_0), the second distribution in (3) has the same exponential term as a Gaussian distribution. The exponential Bessel function will tend to be flatter than the Gaussian component and so the Gaussian component will dominate the shape of the distribution. This resemblance to a Gaussian distribution explains why the Gaussian approximation, comparing the calculated and observed anomalous differences, is reasonably successful.

The influence of the two components of $P_{\rm SAD}$ is shown in Figs. 1 and 2. Fig. 1 illustrates the situation characteristic of SAD phasing, in which the model consists of only the strong anomalous scatterers. In this case, the model of the normal scattering component is very incomplete, so the first (Sim) distribution is very broad and serves primarily to break the phase ambiguity of the second (anomalous difference) distribution. By contrast, Fig. 2 illustrates the situation that would occur in full model refinement against SAD data, where the model of the normal scattering component is nearly complete so the Sim distribution will tend to dominate, while the anomalous difference distribution will provide a weak bimodal indication of the correct phase.

(3) bears a close resemblance to the phased MLHL target (Pannu *et al.*, 1998) for model refinement, so one would expect refinement of a full model against the MLHL target (if appropriately implemented) to yield similar results to refinement against the SAD target. In the MLHL target, an integration over possible phases in the Sim probability distribution is weighted by prior knowledge of the phase probability distribution. If no significant improvement were made in the anomalous scatterer model, the second (anomalous difference) component of (3) would not change during the course of refinement, so it could be used as a constant source of prior phase information in the MLHL target. Note, however, that it would not be appropriate to provide prior phase information to MLHL in the form of the full phase probability distribution obtained by normalizing the integrand of $P_{\rm SAD}$, because the normal scattering from the anomalous scatterers would then appear twice, in both the Sim component of $P_{\rm SAD}$ and the Sim component of MLHL. When the imaginary (f'') contribution to the structure factor is weak compared with the real ($f + f'$) contribution, the amplitude of the real scattering component can be approximated reasonably well by the mean of F_O^+ and F_O^-. Typically, such a mean amplitude would be used in the MLHL target. However, in the presence of very strong anomalous scatterers this approximation breaks down. By analogy with (3), the Sim component of the MLHL target should then compare the observed value of one of the Friedel mates with its corresponding calculated value (including the imaginary contribution). Compared with such an implementation of the MLHL target, any improvement from using the SAD function for model refinement would only arise through improvements in the anomalous scatterering model during the course of refinement. The model of strong anomalous scatterers is unlikely to change substantially during subsequent full model refinement, so the main potential for improvement with the SAD function will come from accounting for partially occupied sites and the weak anomalous scattering from the rest of the structure, such as C, N and O atoms.

APPENDIX *A*
Derivation of SAD likelihood function

A1. General SAD likelihood function

For our maximum-likelihood $P_{\rm SAD}$ function we obtain first the probability of the true F^+ and F^- (unphased) given the heavy-atom structure factors $\mathbf{H}^+$ and $\mathbf{H}^{-*}$ (phased). (A correction for the effect of measurement error will be introduced later; see Appendix *B*). We derive this expression from the probability of the true phased structure factors $\mathbf{F}^+$ and $\mathbf{F}^{-*}$ given the calculated heavy-atom structure factors $\mathbf{H}^+$ and $\mathbf{H}^{-*}$ and then integrate out the phases. Complex conjugates are used for $\mathbf{F}^-$ and $\mathbf{H}^-$ because these are much more highly correlated with their Friedel mates, $\mathbf{F}^+$ and $\mathbf{H}^+$,

$$P(F^+, F^-|\mathbf{H}^+,\mathbf{H}^{-*}) = \int_0^{2\pi}\int_0^{2\pi} P(F^+, \alpha^+, F^-, \alpha^-|\mathbf{H}^+, \mathbf{H}^{-*})\, {\rm d}\alpha^+\, {\rm d}\alpha^-. \quad (4)$$

The conditional probability within the integral can be expressed as a product of two conditional probabilities, only one of which is dependent on α^+,

$$P(F^+, \alpha^+, F^-, \alpha^-|\mathbf{H}^+, \mathbf{H}^{-*}) = P(F^-, \alpha^-|\mathbf{H}^+, \mathbf{H}^{-*})P(F^+, \alpha^+|F^-, \alpha^-, \mathbf{H}^+, \mathbf{H}^{-*}). \quad (5)$$

Substituting (5) into (4) we obtain

$$P(F^+, F^-|\mathbf{H}^+, \mathbf{H}^{-*}) = \int_0^{2\pi} P(F^-, \alpha^-|\mathbf{H}^+, \mathbf{H}^{-*}) \times \left[\int_0^{2\pi} P(F^+, \alpha^+|F^-, \alpha^-, \mathbf{H}^+, \mathbf{H}^{-*})\, {\rm d}\alpha^+\right] {\rm d}\alpha^-. \quad (6)$$

The integral within the square brackets can be performed analytically to obtain a Rice distribution (§*A*4). The integration over α^- must be performed numerically,

$$P(F^+,F^-|\mathbf{H}^+, \mathbf{H}^{-*}) = \int_0^{2\pi} P(F^-, \alpha^-|\mathbf{H}^+, \mathbf{H}^{-*})P(F^+|F^-, \alpha^-, \mathbf{H}^+, \mathbf{H}^{-*})\, {\rm d}\alpha^-. \quad (7)$$

A2. Multivariate complex normal distribution of $\{F^+, F^{-*}, \mathbf{H}^+, \mathbf{H}^{-*}\}$

In order to obtain the probability functions in (2), we start from a multivariate complex normal distribution of structure factors $\{\mathbf{F}^+, \mathbf{F}^{-*}, \mathbf{H}^+, \mathbf{H}^{-*}\}$. There is no prior information before fixing the heavy-atom model and so the expected values are zero.

$$P(\mathbf{F}^{+}, \mathbf{F}^{-*}, \mathbf{H}^{+}, \mathbf{H}^{-*}) = \frac{1}{|\pi\Sigma_{FFHH}|}\exp\left[-\begin{pmatrix}\mathbf{F}^{+}\\ \mathbf{F}^{-*}\\ \mathbf{H}^{+}\\ \mathbf{H}^{-*}\end{pmatrix}^{H}\Sigma_{FFHH}^{-1}\begin{pmatrix}\mathbf{F}^{+}\\ \mathbf{F}^{-*}\\ \mathbf{H}^{+}\\ \mathbf{H}^{-*}\end{pmatrix}\right], \quad (8)$$

where

$$\Sigma_{FFHH} = \begin{pmatrix}\Sigma_{11} & \Sigma_{12}\\ \Sigma_{21} & \Sigma_{22}\end{pmatrix}$$

and

$$\Sigma_{11} = \begin{pmatrix}\langle\mathbf{F}^{+}\mathbf{F}^{+*}\rangle & \langle\mathbf{F}^{+}\mathbf{F}^{-}\rangle\\ \langle\mathbf{F}^{+}\mathbf{F}^{-}\rangle^{*} & \langle\mathbf{F}^{-}\mathbf{F}^{-*}\rangle\end{pmatrix},$$

$$\Sigma_{12} = \begin{pmatrix}\langle\mathbf{F}^{+}\mathbf{H}^{+*}\rangle & \langle\mathbf{F}^{+}\mathbf{H}^{-}\rangle\\ \langle\mathbf{F}^{-}\mathbf{H}^{+}\rangle^{*} & \langle\mathbf{F}^{-*}\mathbf{H}^{-}\rangle\end{pmatrix},$$

$$\Sigma_{21} = \Sigma_{12}^{H},$$

$$\Sigma_{22} = \begin{pmatrix}\langle\mathbf{H}^{+}\mathbf{H}^{+*}\rangle & \langle\mathbf{H}^{+}\mathbf{H}^{-}\rangle\\ \langle\mathbf{H}^{+}\mathbf{H}^{-}\rangle^{*} & \langle\mathbf{H}^{-}\mathbf{H}^{-*}\rangle\end{pmatrix}.$$

The covariance matrix Σ_{FFHH} is shown in terms of submatrices (Σ_{11}, Σ_{12}, Σ_{21} and Σ_{22}) that will be manipulated when the conditional variables are fixed. A superscript H is used here and elsewhere to denote the Hermitian transpose of a matrix.

In defining $\mathbf{F}^{+}$, $\mathbf{F}^{-*}$, $\mathbf{H}^{+}$ and $\mathbf{H}^{-*}$, we use $\mathbf{f}$ and $\mathbf{g}$ to represent atomic scattering factors and $\mathbf{x}$ and $\mathbf{y}$ to represent coordinates for the corresponding crystal and model. In general, the scattering factors are complex to allow for the effects of anomalous scattering so that, for instance, $\mathbf{f}_k = f_k + if_k''$. For simplicity, the model can be considered to contain all the atoms present in the crystal (N), but with zero scattering factor for atoms that are not present in the model. The sums can then be divided into contributions from unmodelled (NU atoms) and modelled atoms.

$$\begin{aligned}\mathbf{F}^{+} &= \sum_{k=1}^{N}\mathbf{f}_k\exp(2\pi i\mathbf{h}\cdot\mathbf{x}_k)\\ &= \sum_{k=1}^{NU}\mathbf{f}_k\exp(2\pi i\mathbf{h}\cdot\mathbf{x}_k) + \sum_{k=NU+1}^{N}\mathbf{f}_k\exp(2\pi i\mathbf{h}\cdot\mathbf{x}_k),\\ \mathbf{F}^{-*} &= \sum_{k=1}^{N}\mathbf{f}_k^{*}\exp(2\pi i\mathbf{h}\cdot\mathbf{x}_k)\\ &= \sum_{k=1}^{NU}\mathbf{f}_k^{*}\exp(2\pi i\mathbf{h}\cdot\mathbf{x}_k) + \sum_{k=NU+1}^{N}\mathbf{f}_k^{*}\exp(2\pi i\mathbf{h}\cdot\mathbf{x}_k),\end{aligned}$$

$$\begin{aligned}\mathbf{H}^{+} &= \sum_{k=1}^{N}\mathbf{g}_k\exp(2\pi i\mathbf{h}\cdot\mathbf{y}_k)\\ &= \sum_{k=NU+1}^{N}\mathbf{g}_k\exp(2\pi i\mathbf{h}\cdot\mathbf{y}_k)\\ \mathbf{H}^{-*} &= \sum_{k=1}^{N}\mathbf{g}_k^{*}\exp(2\pi i\mathbf{h}\cdot\mathbf{y}_k),\\ &= \sum_{k=NU+1}^{N}\mathbf{g}_k^{*}\exp(2\pi i\mathbf{h}\cdot\mathbf{y}_k).\end{aligned} \quad (9)$$

Following the reasoning outlined in Pannu *et al.* (2003), the submatrix Σ_{22} can be filled in as follows:

$$\Sigma_{22} = \begin{pmatrix}\varepsilon\Sigma_H & \varepsilon\sigma_{\mathbf{H}^{+}\mathbf{H}^{-}}\\ \varepsilon\sigma_{\mathbf{H}^{+}\mathbf{H}^{-}}^{*} & \varepsilon\Sigma_H\end{pmatrix}, \quad (10)$$

where

$$\Sigma_H = \sum_{k=NU+1}^{N}|\mathbf{g}_k|^2,$$

$$\sigma_{\mathbf{H}^{+}\mathbf{H}^{-}} = \sum_{k=NU+1}^{N}\mathbf{g}_k^2 = \sum_{k=NU+1}^{N}g_k^2 - g_k''^2 + 2ig_kg_k''.$$

The factor ε accounts for the statistical effect of symmetry. The submatrix Σ_{11} is completed similarly,

$$\Sigma_{11} = \begin{pmatrix}\varepsilon\Sigma_N & \varepsilon\sigma_{\mathbf{F}^{+}\mathbf{F}^{-}}\\ \varepsilon\sigma_{\mathbf{F}^{+}\mathbf{F}^{-}}^{*} & \varepsilon\Sigma_N\end{pmatrix}, \quad (11)$$

where

$$\Sigma_N = \sum_{k=1}^{N}|\mathbf{f}_k|^2,$$

$$\sigma_{\mathbf{F}^{+}\mathbf{F}^{-}} = \sum_{k=1}^{N}\mathbf{f}_k^2 = \sum_{k=1}^{N}f_k^2 - f_k''^2 + 2if_kf_k''.$$

The submatrix Σ_{12} includes the effects of coordinate error and of differences between the true and modelled atomic scattering factors. In a fashion similar to that described in Read (2003), in the context of multiple isomorphous replacement, the elements of Σ_{12} can be described in terms of the elements of Σ_{22}. Consider one element of the matrix Σ_{12},

$$\langle\mathbf{F}^{+}\mathbf{H}^{-}\rangle = \varepsilon\sum_{k=NU+1}^{N}\langle\mathbf{f}_k\mathbf{g}_k\exp[2\pi i(\mathbf{x}_k - \mathbf{y}_k)]\rangle = \varepsilon D\sigma_{\mathbf{H}^{+}\mathbf{H}^{-}}. \quad (12)$$

Here, it is assumed that differences in position are uncorrelated with differences in scattering factor. The factor D accounts for the overall effect of the phase-shift term arising from coordinate errors and absorbs any overall difference in scale between $\mathbf{f}$ and $\mathbf{g}$. The same considerations apply to other elements of Σ_{12}, so that

$$\Sigma_{12} = D\Sigma_{22}.$$

As discussed in Read (2003), after the maximum-likelihood refinement of occupancies and B factors, the model atomic scattering factors $\mathbf{g}_k$ should be approximately equal to $\mathbf{f}_k\langle\exp[2\pi i(\mathbf{x}_k - \mathbf{y}_k)]\rangle$, so that the phase shift and scale components of D will cancel and D will be equal to one.

A3. Conditional distribution $P(\mathbf{F}^+, \mathbf{F}^{-*} \mid \mathbf{H}^+, \mathbf{H}^{-*})$

The conditional distribution $P(\mathbf{F}^+, \mathbf{F}^{-*}|\mathbf{H}^+, \mathbf{H}^{-*})$ has a mean and covariance matrix given by standard manipulation (Johnson & Wichern, 1998) of the above covariance elements,

$$P(\mathbf{F}^+, \mathbf{F}^{-*}|\mathbf{H}^+, \mathbf{H}^{-*}) = \frac{1}{|\pi\Sigma_{FF}|}\exp\left\{-\left[\begin{pmatrix}\mathbf{F}^+\\ \mathbf{F}^{-*}\end{pmatrix} - \boldsymbol{\mu}_{FF}\right]^H \times \Sigma_{FF}^{-1}\left[\begin{pmatrix}\mathbf{F}^+\\ \mathbf{F}^{-*}\end{pmatrix} - \boldsymbol{\mu}_{FF}\right]\right\} \quad (13)$$

where

$$\boldsymbol{\mu}_{FF} = \Sigma_{12}\Sigma_{22}^{-1}\begin{pmatrix}\mathbf{H}^+\\ \mathbf{H}^{-*}\end{pmatrix} = D\begin{pmatrix}\mathbf{H}^+\\ \mathbf{H}^{-*}\end{pmatrix}$$

and

$$\begin{aligned}\Sigma_{FF} &= \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}\\ &= \Sigma_{11} - D^2\Sigma_{22}\\ &= \begin{pmatrix}\varepsilon\sigma_\Delta^2 & \varepsilon\sigma_\Phi\\ \varepsilon\sigma_\Phi^* & \varepsilon\sigma_\Delta^2\end{pmatrix},\\ \sigma_\Delta^2 &= \Sigma_N - D^2\Sigma_H,\\ \sigma_\Phi &= \sigma_{\mathbf{F}^+\mathbf{F}^-} - D^2\sigma_{\mathbf{H}^+\mathbf{H}^-}.\end{aligned}$$

The phase component of σ_Φ arises both from errors in the model of anomalous scatterers and from the (perhaps weak) anomalous scattering from atoms not included in the model. It represents the systematic phase shift between the parts of $\mathbf{F}^+$ and $\mathbf{F}^{-*}$ that are not explained by the model. If the model includes most of the significant anomalous scatterers, the phase shift will be very small and could probably be ignored.

A4. Conditional distributions $P(F^-, \alpha^- \mid \mathbf{H}^+, \mathbf{H}^{-*})$ and $P(F^+, \alpha^+ \mid F^-, \alpha^-, \mathbf{H}^+, \mathbf{H}^{-*})$

Again, with standard manipulations (including a change of variable from complex to polar coordinates) we can form the two conditional distributions in (7). For convenience in notation, we define $\mathbf{F}^{-*} = F^-\exp(i\alpha^-)$, *i.e.* α^- is the phase of the complex conjugate of $\mathbf{F}^-$,

$$\begin{aligned}P(F^-, \alpha^-|\mathbf{H}^+, \mathbf{H}^{-*}) &= P(F^-, \alpha^-|\mathbf{H}^{-*}) \quad (14)\\ &= \frac{F^-}{\pi\varepsilon\sigma_\Delta^2}\exp\left[\frac{-|F^-\exp(i\alpha^-) - D\mathbf{H}^{-*}|^2}{\varepsilon\sigma_\Delta^2}\right],\end{aligned}$$

$$P(F^+, \alpha^+|F^-, \alpha^-, \mathbf{H}^+, \mathbf{H}^{-*}) = \frac{F^+}{\pi\varepsilon\left(\sigma_\Delta^2 - \frac{|\sigma_\Phi|^2}{\sigma_\Delta^2}\right)}\exp\left[\frac{-|F^+\exp(i\alpha^+) - \mathbf{F}_C^+|^2}{\varepsilon\left(\sigma_\Delta^2 - \frac{|\sigma_\Phi|^2}{\sigma_\Delta^2}\right)}\right], \quad (15)$$

where

$$\mathbf{F}_C^+ = D\mathbf{H}^+ + \frac{\sigma_\Phi}{\sigma_\Delta^2}[F^-\exp(i\alpha^-) - D\mathbf{H}^{-*}].$$

The phase α^+ can be integrated out analytically to obtain the Rice distribution, which appears frequently in crystallographic literature (*e.g.* Sim, 1959),

$$P(F^+|F^-, \alpha^-, \mathbf{H}^+, \mathbf{H}^{-*}) = \frac{2F^+}{\varepsilon\left(\sigma_\Delta^2 - \frac{|\sigma_\Phi|^2}{\sigma_\Delta^2}\right)} \times \exp\left[-\frac{F^{+2} + F_C^{+2}}{\varepsilon\left(\sigma_\Delta^2 - \frac{|\sigma_\Phi|^2}{\sigma_\Delta^2}\right)}\right] I_0\left[\frac{2F^+F_C^+}{\varepsilon\left(\sigma_\Delta^2 - \frac{|\sigma_\Phi|^2}{\sigma_\Delta^2}\right)}\right]. \quad (16)$$

A5. Conditional distribution $P(F^+, F^- \mid \mathbf{H}^+, \mathbf{H}^{-*})$

Using the probabilities (14) and (16) in (7) and making the substitution

$$\sigma_\Phi = \sigma_\Delta^2 D_\Phi\exp(i\alpha_\Phi) \text{ where } 0 \le D_\Phi \le 1,$$

we obtain (2) as presented above,

$$P_{\rm SAD} = \frac{2F^+F^-}{\pi\varepsilon^2(1 - D_\Phi^2)\sigma_\Delta^4}\int_0^{2\pi}\exp\left[-\frac{|F^-\exp(i\alpha^-) - D\mathbf{H}^{-*}|^2}{\varepsilon\sigma_\Delta^2} - \frac{F^{+2} + F_C^{+2}}{\varepsilon(1 - D_\Phi^2)\sigma_\Delta^2}\right] I_0\left[\frac{2F^+F_C^+}{\varepsilon(1 - D_\Phi^2)\sigma_\Delta^2}\right] \mathrm{d}\alpha^-,$$

where

$$F_C^+ = |D\mathbf{H}^+ + D_\Phi\exp(i\alpha_\Phi)[F^-\exp(i\alpha^-) - D\mathbf{H}^{-*}]|.$$

APPENDIX *B*
Implementation of SAD likelihood function

For numerical stability it is convenient to express (2) in terms of the exponential Bessel function $eI_0(x) = \exp(-x)I_0(x)$ (Cody & Stoltz, 1989). During refinement of the heavy-atom parameters, the D values are absorbed by the occupancies and B factors of the heavy atoms and are therefore not included. The term $(1 - D_\Phi^2)\sigma_\Delta^2$ can be problematic during refinement because D_Φ and σ_Δ are on very different scales (D_Φ is very close to 1, while σ_Δ is large) and $(1 - D_\Phi^2)$ must remain positive (*i.e.* D_Φ^2 must remain between 0 and 1). In order to avoid these problems, we introduce a parameter σ_+ to replace this term. This removes the problem of scale and simplifies the constraint to one in which σ_+ must remain positive.

Up to this point, we have derived the function in terms of the true values of F^+ and F^-. If we use the experimental observations of their values, F_O^+ and F_O^-, we need to consider the experimental errors, which will be described by variance parameters $\sigma_{F_O^+}^2$ and $\sigma_{F_O^-}^2$. In the case of MIR phasing, the effect of measurement error in the observed amplitude can be approximated by inflating the corresponding variance element of the covariance matrix (Pannu *et al.*, 2003), as suggested by others (Green, 1979; de La Fortelle & Bricogne, 1997; Murshudov *et al.*, 1997). The increment to the variance ends up in the variance of the Rice distribution for each observed amplitude. However, if this approach is taken for the SAD function, the variances for the component distributions of $P_{\rm SAD}$ become unnecessarily complicated. Rather than inflating the diagonal elements of the covariance matrix, we have chosen instead to inflate the variances of the conditional

distributions for each observation that are the components of $P_{\rm SAD}$. The variance term for $P(F_O^-, \alpha^-|\mathbf{H}^+, \mathbf{H}^{-*})$ only needs to account for errors in the measurement of F_O^-, but the variance term for $P(F_O^+, \alpha^+|F_O^-, \alpha^-, \mathbf{H}^+, \mathbf{H}^{-*})$ needs to account for errors in both measurements, as the expected value of F_O^+ is computed using the measured value of F_O^-, weighted by D_Φ. The magnitude of D_Φ will typically be very close to one, so the weighting factor on the variance of F_O^- can be ignored; the very small systematic decrease in the contribution from the experimental error in F_O^- owing to D_Φ can be absorbed by σ_+. Numerical simulations show that this approximation to the effect of measurement error gives almost identical results to those obtained by inflating the diagonal elements of the covariance matrix.

The target function for anomalous scatterer refinement in *PHASER* is thus given by

$$-\ln(P_{\rm SAD}) = -\ln\left\{\frac{2F_O^+F_O^-}{\pi\Sigma^+\Sigma^-}\int_0^{2\pi}\exp\left[-\frac{|F_O^-\exp(i\alpha^-)-\mathbf{H}^{-*}|^2}{\Sigma^-} - \frac{(F_O^+-F_C^+)^2}{\Sigma^+}\right]eI_0\left(\frac{2F_O^+F_C^+}{\Sigma^+}\right)\,{\rm d}\alpha^-\right\}, \qquad (17)$$

where

$$\Sigma^- = \varepsilon\sigma_\Delta^2 + \sigma_{F_O^-}^2,$$
$$\Sigma^+ = \varepsilon\sigma_+ + \sigma_{F_O^+}^2 + \sigma_{F_O^-}^2,$$
$$F_C^+ = |\mathbf{H}^+ + D_\Phi\exp(i\alpha_\Phi)[F_O^-\exp(i\alpha^-) - \mathbf{H}^{-*}]|.$$

Initial estimates for σ_Δ^2 can be obtained for each resolution shell by subtracting the mean value of $|\mathbf{H}^-|^2$ from the mean value of F_O^{-2}. Initial estimates for σ_+ could in principle be obtained as a weighted average of $(F_O^+ - F_C^+)^2$ over the phase integral, weighted by the phase probability distribution. In practice, σ_+ will be comparable in size to the contributions from measurement errors and can be readily refined from an initial estimate given by the mean value of $\sigma_{F_O^+}^2$.

We thank Z. Dauter and M. Turkenburg for the diffraction data sets used in the test cases. This work was funded by NIH/NIGMS under grant No. 1P01GM063210 and by a Principal Research Fellowship from the Wellcome Trust (RJR).

References

Burla, M. C., Carrazzini, B., Cascarano, G. L., Giacovazzo, C., Polidori, G. & Siliqi, G. (2002). *Acta Cryst.* D**58**, 928–935.
Cody, W. J. & Stoltz, L. (1989). *ACM TOMS*, **15**, 41–48.
Collaborative Computational Project, Number 4 (1994). *Acta Cryst.* D**50**, 760–763.
Dauter, Z. & Adamiak, D. A. (2001). *Acta Cryst.* D**57**, 990–995.
Dodson, E. (2003). *Acta Cryst.* D**59**, 1958–1965.
Giacovazzo, C. (1983). *Acta Cryst.* A**39**, 585–592.
Giacovazzo, C., Ladisa, M. & Siliqi, D. (2003). *Acta Cryst.* A**59**, 262–265.
Giacovazzo, C. & Siliqi, D. (2001*a*). *Acta Cryst.* A**57**, 40–46.
Giacovazzo, C. & Siliqi, D. (2001*b*). *Acta Cryst.* A**57**, 414–419.
Green, E. A. (1979). *Acta Cryst.* A**35**, 351–359.
Griewank, A., Juedes, D., Mitev, H., Utke, J., Vogel, O. & Walther, A. (1996). *ACM TOMS*, **22**, 131–167.
Hauptman, H. (1982). *Acta Cryst.* A**38**, 632–641.
Hendrickson, W. A. (1991). *Science*, **254**, 51–58.
Johnson, R. A. & Wichern, D. W. (1998). *Applied Multivariate Statistical Analysis*, 4th ed. New Jersey: Prentice–Hall.
La Fortelle, E. de & Bricogne, G. (1997). *Methods Enzymol.* **276**, 472–494.
Matthews, B. W. (1966). *Acta Cryst.* **20**, 82–86.
Murshudov, G. N., Vagin, A. A. & Dodson, E. J. (1997). *Acta Cryst.* D**53**, 240–255.
North, A. C. T. (1965). *Acta Cryst.* **18**, 212–216.
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., McCoy, A. J. & Read, R. J. (2003). *Acta Cryst.* D**59**, 1801–1808.
Pannu, N. S., Murshudov, G. N., Dodson, E. J. & Read, R. J. (1998). *Acta Cryst.* D**54**, 1285–1294.
Pannu, N. S. & Read, R. J. (2004). *Acta Cryst.* D**60**, 22–27.
Read, R. J. (1986). *Acta Cryst.* A**42**, 140–149.
Read, R. J. (2003). *Acta Cryst.* D**59**, 1891–1902.
Rice, L. M., Earnest, T. N. & Brünger, A. T. (2000). *Acta Cryst.* D**56**, 1413–1420.
Sim, G. A. (1959). *Acta Cryst.* **12**, 813–815.
Terwilliger, T. C. & Berendzen, J. (1997). *Acta Cryst.* D**53**, 571–579.
Terwilliger, T. C. & Eisenberg, D. (1987). *Acta Cryst.* A**43**, 6–13.