Acta Cryst. (1988). A44, 517–545

A Bayesian Statistical Theory of the Phase Problem. I. A Multichannel Maximum-Entropy Formalism for Constructing Generalized Joint Probability Distributions of Structure Factors

BY G. BRICOGNE

LURE, Bâtiment 209D, 91405 Orsay, France

(*Received* 23 *December* 1987; *accepted* 24 *March* 1988)

Abstract

In this first of three papers on a full Bayesian theory of crystal structure determination, it is shown that all currently used sources of phase information can be represented and combined through a universal expression for the joint probability distribution of structure factors. Particular attention is given to situations arising in macromolecular crystallography, where the proper treatment of non-uniform distributions of atoms is absolutely essential. A procedure is presented, in stages of gradually increasing complexity, for constructing the joint probability distribution of an arbitrary collection of structure factors. These structure factors may be gathered from one or several crystal forms of an unknown molecule, each comprising one or several isomorphous structures related by substitution operations, possibly containing solvent regions and known fragments, and/or obeying a set of non-crystallographic symmetries. This universal joint probability distribution can be effectively approximated by the saddlepoint method, using maximum-entropy distributions of atoms [Bricogne (1984). *Acta Cryst.* A40, 410–445] and a generalization of structure-factor algebra. Atomic scattering factors may assume arbitrary complex values, so that this formalism applies to neutron as well as to X-ray diffraction methods. This unified procedure will later be extended by the construction of conditional distributions allowing phase extension, and of likelihood functions capable of detecting and characterizing all potential sources of phase information considered so far, thus completing the formulation of a full Bayesian inference scheme for crystal structure determination.

Introduction

The determination of crystal structures by X-ray or neutron diffraction, *i.e.* the transition from integrated intensity measurements to a refined atomic model, has often been likened, explicitly or implicitly, to a process of statistical inference. In particular, French (1978) has given a detailed overview of the Bayesian approach to the theory of inference, and a persuasive account of its relevance to crystallographic methodology. His paper, however, is mostly of an expository nature and concerns itself with advocating the use of Bayesian concepts, rather than with actually developing the specific tools needed to bring about their implementation in crystallography. The main practical outcomes of his study are procedures for the treatment of negative-intensity observations (French & Wilson, 1978) and for profile fitting in diffractometer-data analysis (Oatley & French, 1982). The present work is concerned with the derivation of precisely those specific analytical and probabilistic results which are needed to turn Bayesian concepts into effective computational tools for crystal structure determination, *i.e.* for solving the phase problem. An introductory survey of these developments has been given elsewhere (Bricogne, 1987, 1988).

Various instances besides the actual knowledge or legitimate assumption of certain phase values have been described as 'sources of phase information': (*a*) the atomicity of crystal structures and the statistical phase relations that ensue; (*b*) the availability of an isomorphous substitution series; (*c*) the presence of known molecular fragments; (*d*) the existence of solvent regions and/or of non-crystallographic symmetries; and (*e*) the availability of several crystal forms of the same molecule. Each such instance has given rise to a distinct phase-determination methodology in which the nature, the relative importance and the degree of sophistication of the statistical methods employed vary greatly. Certain structure determination procedures, such as correlation methods based on Patterson functions, even seem at first not to involve any statistics at all; but assessing the level of significance of their results *is* a statistical problem. Clearly, a unified statistical treatment of all these methods would be most valuable, and the Bayesian approach to inference (*e.g.* Box & Tiao, 1973) is a natural candidate for such a theory.

The most fundamental prerequisite to the construction of such a Bayesian theory is as follows. If we assume a given combination of instances (*a*)–(*e*) to hold, this assumption being described by a suitable collection of parameters, then the assumption should be translated into a procedure for calculating the *joint probability of all the diffraction intensities* which can be experimentally measured on the structure under

0108-7673/88/040517-29$03.00

© 1988 International Union of Crystallography

study. Once this is achieved, different assumptions may be ranked in the light of the observations according to their *likelihood*, *i.e.* to the probability they had assigned to the actual outcome of the measurements. Choices may then be made, on the basis of prior probabilities summarizing previous experience, that minimize the cost of possibly making wrong decisions. The overall process of crystal structure determination thus becomes a *game* in the sense of statistical decision theory (Blackwell & Girshick, 1954; De Groot, 1970), to which the tree-directed multisolution strategy using a combination of entropy and likelihood as a heuristic criterion (Bricogne, 1984, 1988) can be applied.

This first paper presents a general procedure for carrying out the basic task of deriving joint probability distributions of structure factors capable of accommodating all current used sources of phase information. Our starting point (§ 0) is a self-contained recapitulation of the procedure proposed in a previous paper (Bricogne, 1984, hereafter referred to as MEFDM) for constructing the saddlepoint approximation to the joint probability distribution (j.p.d.) of any prescribed collection of structure factors belonging to a crystal structure made of unit-weight point atoms distributed identically and independently with a non-uniform density. The first generalization (§ 1) extends this treatment to the case where all atoms remain identical but have an arbitrary complex scattering factor. The second generalization (§ 2) considers a heterogeneous structure made up of several species of atoms, each species having its own scattering factor and its own non-uniform prior distribution; this affords a treatment not only of the chemical heterogeneity of most crystalline compounds, but also of solvent regions in macromolecular crystals. The third generalization (§ 3) considers not *one* structure, but a *family* of isomorphous heterogeneous structures, where each atom contributes simultaneously to all structures of the family but may do so with different scattering factors; this model can now accommodate a statistical formulation of all substitution methods: isomorphous replacement (single or multiple), anomalous scattering (at one or several wavelengths) and solvent-contrast variation. The fourth generalization (§ 4) deals with cases where part of a structure, or a common part of all structures in an isomorphous family, consists of a known molecular fragment; it yields a statistical formulation of the molecular replacement method and of the process of completing a partial structure which has several advantages over current methodology. The fifth generalization (§ 5) allows for the possible existence of local non-crystallographic symmetry within the asymmetric unit of a crystal or of an isomorphous family of crystal structures, whose effect is to alter profoundly the covariance structure of the j.p.d. and thus to create very strong phase relations. The sixth generalization (§ 6) further allows for the possible availability of several crystal forms of the same molecule, each built with a different lattice and a different space group, and possibly comprising an isomorphous family, with known fragments and non-crystallographic symmetries; if the various crystal forms differ only by slight lattice deformations, a treatment of non-isomorphous heavy-atom derivatives is obtained. Remarkably, *the formal structure of the j.p.d. remains the same* throughout these successive generalizations. These developments are summarized in § 7.

The outcome of this paper is thus a procedure for optimally combining, in the form of a universal expression for the j.p.d. of relevant structure factors, all currently used sources of phase information. Paper II will proceed to the local study of the dependence of this universal form for joint and conditional distributions on the initial assumptions, and to the construction of likelihood functions for testing these assumptions on the basis of various types of observations. These will usually be single-crystal diffraction intensities, but may also be fibre or powder data, and the treatment will include the effect of measurement errors. Finally, paper III will illustrate this unified approach by applying it to a number of classical problems and comparing its results with those of the standard methods available in each case.

0. Unit point-atom structures

This preliminary section aims at giving a self-contained presentation of a new analytical approach – the saddlepoint method – proposed in MEFDM for approximating j.p.d.'s of structure factors when large moduli are present. For simplicity this approach is first formulated here for point atoms of unit weight, *i.e.* for the *purely trigonometric* part of atomic contributions to structure factors. This has the advantage of displaying the mathematical basis of the method without the complications introduced by atomic scattering factors, which are dealt with in § 1. The transition to conditions distributions and to likelihood functions is then sketched in qualitative terms, to serve as a pointer to their later use in paper II and to illustrate the scope of each of the generalizations presented here.

0.0. *Definitions and conventions*

Let H be a set of unique non-origin reflexions $\mathbf{h}$ for a crystal with lattice $\mathscr{R}$ and space group $G = \{S_g \mid g \in G\}$, where

$$S_g: \quad \mathbf{x} \rightarrow \mathbf{R}_g\mathbf{x} + \mathbf{t}_g.$$

The asymmetric unit D of this crystal may contain special positions $\mathbf{x}$ which are invariant under some of the S_g; for such points $\mathbf{x}$, $G_\mathbf{x}$ will denote the

isotropy subgroup of $\mathbf{x}$:

$$G_{\mathbf{x}} = \{g \in G \,|\, \mathbf{R}_g\mathbf{x} + \mathbf{t}_g \equiv \mathbf{x} \bmod \mathcal{R}\},$$

and $|G_{\mathbf{x}}|$ will denote the number of its elements.

Let H contain n_a acentric and n_c centric reflexions. Structure factor values attached to all reflexions in H will comprise $n = 2n_a + n_c$ real numbers. For $\mathbf{h}$ acentric, $\alpha_{\mathbf{h}}$ and $\beta_{\mathbf{h}}$ will be the real and imaginary parts of the complex structure factor; for $\mathbf{h}$ centric, $\gamma_{\mathbf{h}}$ will be the real coordinate of the (possibly complex) structure factor measured along a real axis rotated by one of the two angles $\theta_{\mathbf{h}}$, π apart, to which the phase is restricted modulo 2π. These n real coordinates will be arranged (MEFDM, § 7.1.1) as a column vector containing the acentric then the centric data, *i.e.* in the order

$$\alpha_1, \beta_1, \alpha_2, \beta_2, \ldots, \alpha_{n_a}, \beta_{n_a}, \gamma_1, \gamma_2, \ldots, \gamma_{n_c}.$$

0.1. *Vectors of trigonometric structure-factor expressions*

Let $\boldsymbol{\xi}(\mathbf{x})$ denote the vector of trigonometric structure-factor expressions associated with $\mathbf{x} \in D$. These are defined as follows:

$$\alpha_{\mathbf{h}}(\mathbf{x}) + i\beta_{\mathbf{h}}(\mathbf{x}) = \Xi(\mathbf{h}, \mathbf{x}) \qquad \text{for } \mathbf{h} \text{ acentric} \tag{0.1a}$$

$$\gamma_{\mathbf{h}}(\mathbf{x}) = \exp(-i\theta_{\mathbf{h}})\Xi(\mathbf{h}, \mathbf{x}) \qquad \text{for } \mathbf{h} \text{ centric} \tag{0.1b}$$

where

$$\Xi(\mathbf{h}, \mathbf{x}) = |G_{\mathbf{x}}|^{-1} \sum_{g \in G} \exp[2\pi i \mathbf{h}.(S_g\mathbf{x})], \tag{0.2}$$

since each point in the orbit of $\mathbf{x}$ under the action of G is repeated $|G_{\mathbf{x}}|$ times.

According to the convention of § 0.0, the coordinates of $\boldsymbol{\xi}(\mathbf{x})$ in $\mathbb{R}^n$ will be arranged in a column vector as

$$\xi_{2r-1}(\mathbf{x}) = \alpha_{\mathbf{h}(r)}(\mathbf{x}) \qquad r = 1, \ldots, n_a \tag{0.3a}$$

$$\xi_{2r}(\mathbf{x}) = \beta_{\mathbf{h}(r)}(\mathbf{x}) \qquad r = 1, \ldots, n_a \tag{0.3b}$$

$$\xi_{n_a+r}(\mathbf{x}) = \gamma_{\mathbf{h}(r)}(\mathbf{x}) \qquad r = n_a+1, \ldots, n_a+n_c \tag{0.3c}$$

0.2. *Distributions of random atoms and moment-generating functions*

Let position $\mathbf{x}$ in D now become a random vector with probability distribution $m(\mathbf{x})$. Then $\boldsymbol{\xi}(\mathbf{x})$ becomes itself a random vector in $\mathbb{R}^n$, whose distribution $p(\boldsymbol{\xi})$ is the image of distribution $m(\mathbf{x})$ through the mapping $\mathbf{x} \to \boldsymbol{\xi}(\mathbf{x})$ described by (0.1), (0.2) and (0.3). The locus of $\boldsymbol{\xi}(\mathbf{x})$ in $\mathbb{R}^n$ is an algebraic manifold (the multidimensional analogue of a Lissajous curve), and hence p is a singular measure; nevertheless, the average with respect to p of any function Ω in $\mathbb{R}^n$ may be calculated as an average with respect to m over D by the 'induction formula'

$$\int_{\mathbb{R}^n} p(\boldsymbol{\xi})\Omega(\boldsymbol{\xi})\, \mathrm{d}^n\boldsymbol{\xi} = \int_D m(\mathbf{x})\Omega[\boldsymbol{\xi}(\mathbf{x})]\, \mathrm{d}^3\mathbf{x}. \tag{0.4}$$

In particular, one can calculate the moment generating function (m.g.f.) M for distribution p as

$$M(\mathbf{t}) \equiv \int_{\mathbb{R}^n} p(\boldsymbol{\xi})\, e^{\mathbf{t}.\boldsymbol{\xi}}\, \mathrm{d}^n\boldsymbol{\xi} = \int_D m(\mathbf{x})\, e^{\mathbf{t}.\boldsymbol{\xi}(\mathbf{x})}\, \mathrm{d}^3\mathbf{x} \tag{0.5}$$

and hence calculate the moments μ (cumulants κ) of p by differentiation of M ($\log M$) at $\mathbf{t} = \mathbf{0}$:

$$\mu_{r_1 r_2 \ldots r_n} \equiv \int_D m(\mathbf{x})\xi_1^{r_1}(\mathbf{x}) \ldots \xi_n^{r_n}(\mathbf{x})\, \mathrm{d}^3\mathbf{x} = \left.\frac{\partial^{r_1+\ldots+r_n} M}{\partial t_1^{r_1} \ldots \partial t_n^{r_n}}\right|_{\mathbf{t}=\mathbf{0}} \tag{0.6}$$

$$\kappa_{r_1 r_2 \ldots r_n} = \left.\frac{\partial^{r_1+\ldots+r_n} (\log M)}{\partial t_1^{r_1} \ldots \partial t_n^{r_n}}\right|_{\mathbf{t}=\mathbf{0}}. \tag{0.7}$$

The structure-factor algebra for group g (see Appendix) then allows one to express products of ξ's as linear combinations of other ξ's, and hence to express all moments and cumulants of distribution $p(\boldsymbol{\xi})$ as linear combinations of real and imaginary parts of Fourier coefficients of the prior distribution of atoms $m(\mathbf{x})$ (MEFDM, § 7.1.1). This is the key element in the use of non-uniform distributions of atoms throughout this work.

An important property of the cumulant-generating function $\log M$ is that it is *strictly convex*, *i.e.* that its Hessian matrix $\nabla^2(\log M)$ is everywhere positive definite, provided that the set of vectors

$$\{\boldsymbol{\xi}(\mathbf{x}) \,|\, \mathbf{x} \in D, m(\mathbf{x}) \neq 0\}$$

spans the whole of $\mathbb{R}^n$. Partial loss of this property in later extension of the theory will require that certain regularization procedures be used to recover it in a subspace.

0.3. *The joint probability distribution of structure factors*

In the random atom model of an equal-atom structure, N atoms are placed randomly, independently of each other, in the asymmetric unit D of the crystal with probability density $m(\mathbf{x})$. For point atoms of unit weight, the vector $\mathbf{F}$ of structure-factor values for reflexions $\mathbf{h} \in H$ may be written as

$$\mathbf{F} = \sum_{I=1}^{N} \boldsymbol{\xi}^{[I]} \tag{0.8}$$

where the N copies $\boldsymbol{\xi}^{[I]}$ of random vector $\boldsymbol{\xi}$ are independent and have the same distribution $p(\boldsymbol{\xi})$.

The joint probability distribution $\mathcal{P}(\mathbf{F})$ is then the Nth convolution power of p:

$$\mathcal{P}(\mathbf{F}) = p^{*N}(\mathbf{F}) \tag{0.9}$$

and hence is the Fourier transform of the Nth power of its characteristic function $M(i\mathbf{t})$, which may be written [MEFDM, equation (5.11)] as

$$\mathscr{P}(\mathbf{F}) = (2\pi)^{-n} \int_{\mathbb{R}^n} M^N(i\mathbf{t})\, e^{-i\mathbf{t}.\mathbf{F}}\, \mathrm{d}^n\mathbf{t} \qquad (0.10a)$$

$$= (2\pi)^{-n} \int_{\mathbb{R}^n} \exp\{N[\log M(i\mathbf{t}) - i\mathbf{t}.(\mathbf{F}/N)]\}\, \mathrm{d}^n\mathbf{t}. \qquad (0.10b)$$

For low dimensionality n it is possible to carry out the Fourier transformation (0.10a) numerically, provided $M(i\mathbf{t})$ is sampled sufficiently finely that no aliasing results from taking its Nth power (Barakat, 1974). This exact approach, which can also fully cope with atomic heterogeneity, was first used in the field of intensity statistics (Shmueli, Weiss, Kiefer & Wilson, 1984; Shmueli, Weiss & Kiefer, 1985; Shmueli & Weiss, 1987), then in the study of the $\sum_1$ and $\sum_2$ relations in triclinic space groups (Shmueli & Weiss, 1985, 1986). It could be extended to the construction of any j.p.d. in any space group by using the generic expression for the m.g.f. derived in MEFDM § 3.5.2. It is, however, limited to small values of n by the necessity to carry out n-dimensional FFT's on large arrays of sample values: in all other situations, some approximation method must be used.

The asymptotic expansions of Gram–Charlier (Bertaut, 1955a, b) and Edgeworth (Klug, 1958) are obtained by expanding log M as a Taylor series near $\mathbf{t} = \mathbf{0}$, then integrating termwise after a rearrangement. As pointed out in MEFDM (§ 2), these expansions have good convergence properties only if $F_\mathbf{h}$ lies in the vicinity of $\langle F_\mathbf{h}\rangle = N\mathscr{F}^{-1}[m](\mathbf{h})$ for all $\mathbf{h} \in H$. All previous work on the j.p.d. of structure factors has used for $m(\mathbf{x})$ a *uniform* distribution, so that $\langle\mathbf{F}\rangle = \mathbf{0}$; as a result, the corresponding expansions are accurate only if all moduli $|F_\mathbf{h}|$ are small, in which case the j.p.d. contains little phase information.

The recent work of Castleden (1987) and of Peschar & Schenk (1987) on the general formal structure of j.p.d.'s and on the computer-aided construction of the Edgeworth series still uses a uniform prior distribution of atoms, and thus still does not address the problem of 'recentring' this asymptotic expansion near a point $\mathbf{F}$ far from $\langle\mathbf{F}\rangle = \mathbf{0}$. The work on exact methods also uses uniform prior distributions of atoms, although in this case this does not, of course, cause convergence difficulties. Only exceptionally (Gramlich, 1984) has the problem of a proper treatment of non-uniform distributions of atoms in direct methods been examined.

Since the major thrust of this work is the construction of joint probability distributions of structure factors in a large variety of situations where the non-uniformity of prior distributions of atoms is an essential feature, the present approach will use a different method of approximation in which non-uniformity is not only *accommodated*, but is *used actively* as a recentring device: the saddlepoint method.

0.4. *The saddlepoint approximation*

The convergence difficulties encountered with classical asymptotic expansions are easily understood: one is substituting a *local* approximation to log M, in the form of a Taylor series expansion valid near $\mathbf{t} = \mathbf{0}$, into an integral (0.10); whereas integration is a *global* process which consults values of log M far from $\mathbf{t} = \mathbf{0}$.

It is possible, however, to let the point $\mathbf{t}$ where log M is expanded as a Taylor series depend on the particular value $\mathbf{F}^*$ of $\mathbf{F}$ for which an accurate evaluation of $\mathscr{P}(\mathbf{F})$ is desired. This is the essence of the saddlepoint method (Fowler, 1936; Khinchin 1949; Daniels, 1954; MEFDM, § 5), which uses an analytical continuation of $M(\mathbf{t})$ from a function over $\mathbb{R}^n$ to a function over $\mathbb{C}^n$ (Paley & Wiener, 1934; Schwartz, 1966). With the substitution $\mathbf{t} = \mathbf{s} - i\boldsymbol{\tau}$, the $\mathbb{C}^n$ version of Cauchy's theorem (Hörmander, 1973) gives rise to the identity

$$\mathscr{P}(\mathbf{F}^*) = (2\pi)^{-n} \exp(-\boldsymbol{\tau}.\mathbf{F}^*) \times \int_{\mathbb{R}^n} \exp\{N[\log M(\boldsymbol{\tau} + i\mathbf{s}) - i\mathbf{s}.(\mathbf{F}^*/N)]\}\, \mathrm{d}^n\mathbf{s} \qquad (0.11)$$

for all $\boldsymbol{\tau} \in \mathbb{R}^n$. By a convexity argument, which depends crucially on the fact that the vectors $\boldsymbol{\xi}(\mathbf{x})$ span the whole of $\mathbb{R}^n$ as $\mathbf{x}$ runs through D, there is a unique value of $\boldsymbol{\tau}$ such that

$$\boldsymbol{\nabla}(\log M)|_{\mathbf{t}=\mathbf{0}-i\boldsymbol{\tau}} = \mathbf{F}^*/N. \qquad (0.12)$$

At the *saddlepoint* $\mathbf{t}^* = \mathbf{0} - i\boldsymbol{\tau}$, the modulus of the integrand in (0.11) is a maximum and its phase is stationary with respect to the integration variable $\mathbf{s}$: as N tends to infinity, all contributions to the integral cancel because of rapid oscillation, except those coming from the immediate vicinity of $\mathbf{t}^*$ where there is no oscillation. A Taylor expansion of log M^N to second order with respect to $\mathbf{s}$ at $\mathbf{t}^*$ then gives

$$\log M^N(\boldsymbol{\tau} + i\mathbf{s}) \simeq \log M^N(\boldsymbol{\tau}) + i\mathbf{s}.\mathbf{F}^* - \tfrac{1}{2}[\mathbf{s}^T\boldsymbol{\nabla}^2(\log M^N)\mathbf{s}] \qquad (0.13)$$

and substitution into (0.11) leads to

$$\mathscr{P}(\mathbf{F}^*) \simeq \exp[\log M^N(\boldsymbol{\tau}) - \boldsymbol{\tau}.\mathbf{F}^*] \times (2\pi)^{-n} \int_{\mathbb{R}^n} \exp\{-\tfrac{1}{2}[\mathbf{s}^T\boldsymbol{\nabla}^2(\log M^N)\mathbf{s}]\}\, \mathrm{d}^n\mathbf{s}. \qquad (0.14)$$

The last integral is elementary and gives the 'saddlepoint approximation'

$$\mathscr{P}^{\mathrm{SP}}(\mathbf{F}^*) = e^{\mathscr{S}}[\det(2\pi\mathbf{Q})]^{-1/2} \qquad (0.15)$$

where

$$\mathscr{S} = \log M^N(\boldsymbol{\tau}) - \boldsymbol{\tau}.\mathbf{F}^* \qquad (0.16)$$

and where

$$\mathbf{Q} = \nabla^2(\log M^N). \tag{0.17}$$

As shown previously (MEFDM, § 5.4) this approximation amounts to using the exponentially modulated 'conjugate distribution' (Khinchin, 1949)

$$p_\tau(\xi) = p(\xi)\, e^{\tau.\xi}/M(\tau) \tag{0.18}$$

instead of the original distribution $p(\xi) = p_0(\xi)$ for the distribution of random vector ξ. By (0.4), p_τ is induced from the modified distribution of atoms

$$q_\tau(\mathbf{x}) = q(\mathbf{x})\, e^{\tau.\xi(\mathbf{x})}/M(\tau) \tag{SP1}$$

with, by (0.5),

$$M(\tau) = \int_D m(\mathbf{x})\, e^{\tau.\xi(\mathbf{x})}\, \mathrm{d}^3\mathbf{x}, \tag{SP2}$$

and where τ is the unique solution of (0.12), which we may write

$$\nabla_\tau(\log M^N) = \mathbf{F}^*. \tag{SP3}$$

Finally, the elements of the Hessian matrix $\nabla^2(\log M)$ are just the trigonometric second-order cumulants of distribution p, and hence can be calculated *via* structure-factor algebra (§ 0.2) from the Fourier coefficients of $q_\tau(\mathbf{x})$. All the quantities involved in expression (0.15) for $\mathscr{P}^{\mathrm{SP}}(\mathbf{F}^*)$ are therefore effectively computable from the initial data $m(\mathbf{x})$ and $\mathbf{F}^*$.

It was further shown in MEFDM, § 6, that the positive definiteness of $\nabla^2(\log M)$, which follows from the positivity of $m(\mathbf{x})$ and of $q_\tau(\mathbf{x})$ for all τ, is the source of all the determinantal inequalities (Toeplitz, 1911; Carathéodory, 1911) used in direct methods (Harker & Kasper, 1948; Karle & Hauptman, 1950; Tsoucaris, 1970).

0.5. *Maximum-entropy distributions of atoms*

One of the main results of the previous paper (MEFDM, § 5.4) is that the modified distribution $q_\tau(\mathbf{x})$ in (SP1) is the unique distribution which has *maximum entropy* $\mathscr{S}_m(q)$ (*i.e.* contains the least amount of added information) relative to $m(\mathbf{x})$, where

$$\mathscr{S}_m(q) = -\int_D q(\mathbf{x}) \log[q(\mathbf{x})/m(\mathbf{x})]\, \mathrm{d}^3\mathbf{x}, \tag{0.19}$$

under the constraint that $\mathbf{F}^*$ be the centroid vector of the corresponding conjugate distribution $\mathscr{P}_\tau(\mathbf{F})$. The coordinates τ of the saddlepoint are then the Lagrange multipliers λ for these constraints. The traditional notation of maximum-entropy (ME) theory (Jaynes, 1957, 1968) for this case (MEFDM, § 3) is

$$q^{\mathrm{ME}}(\mathbf{x}) = [m(\mathbf{x})/Z(\lambda)]\, e^{\lambda.\xi(\mathbf{x})} \tag{ME1}$$

$$Z(\lambda) = \int_D m(\mathbf{x})\, e^{\lambda.\xi(\mathbf{x})}\, \mathrm{d}^3\mathbf{x} \tag{ME2}$$

$$\nabla_\lambda(\log Z^N) = \mathbf{F}^*, \tag{ME3}$$

so that Z is identical to the m.g.f. M. In what follows we will use a mixed notation, in which $M(\mathbf{t})$ will denote the m.g.f. of distribution $p(\xi)$, ξ being the vector of *trigonometric structure-factor expressions* for unit point atoms; while $Z(\mathbf{u})$ will denote the m.g.f. of distribution $P(\mathbf{X})$, $\mathbf{X}$ being the vector of *actual structure-factor contributions* after the atomic scattering factor has been applied.

Jaynes's ME theory also gives an estimate for $\mathscr{P}(\mathbf{F}^*)$,

$$\mathscr{P}^{\mathrm{ME}}(\mathbf{F}^*) \simeq e^{\mathscr{S}}, \tag{0.20}$$

where

$$S = \log Z^N - \lambda.\mathbf{F}^* = N\mathscr{S}_m(q^{\mathrm{ME}}) \tag{0.21}$$

is the total entropy and is the counterpart to (0.16) under the equivalence just established.

$\mathscr{P}^{\mathrm{ME}}$ is similar to $\mathscr{P}^{\mathrm{SP}}$, but lacks the denominator. The latter, which is the normalization factor of a multivariate Gaussian with covariance matrix $\mathbf{Q}$, may easily be seen to arise through Szegö's (1920) theorem from the extra logarithmic term in Stirling's formula

$$\log(n!) \simeq n \log n - n + \tfrac{1}{2}\log(2\pi n) \tag{0.22}$$

(*e.g.* Lebedev, 1972) beyond the first two terms which serve to define entropy. It is thus clear that the saddlepoint approximation (0.15) is superior to the estimate (0.20) provided by ME theory, since the effect of the extra normalization factor will have to be taken into account as soon as the ratio n/N ceases to be negligible. By the same theorem of Szegö (Bricogne, 1982*b*; Britten & Collins, 1982; Narayan & Nityananda, 1982), the logarithm of this normalization factor is related to the Burg entropy of q^{ME}. These matters will be developed further in paper II.

The above relation between entropy maximization and the saddlepoint approximation is fundamental to the general approach which is about to be presented. The ME criterion intervenes only in the construction of $q^{\mathrm{ME}}(\mathbf{x})$ under constraint values $\mathbf{F}^*$, and the distribution $q^{\mathrm{ME}}(\mathbf{x})$ is merely a computational intermediate in obtaining the approximate j.p.d. $\mathscr{P}^{\mathrm{SP}}(\mathbf{F}^*)$ and its associated conditional distributions and likelihood functions.

0.6. *Conditional distributions, likelihood functions and phase refinement*

Although this is the main topic of paper II, we will introduce here the basic ideas, terminology and notation needed to indicate in general terms the purpose of the extensive generalizations of § 0.4 which are about to be carried out, and thus to establish in advance a logical link with the corresponding sections of paper II.

Let H and K denote two disjoint subsets of unique non-origin reflexions, and let $\mathbf{F}_H$ and $\mathbf{F}_K$ denote the vectors of associated structure-factor values. The conditional probability distribution of $\mathbf{F}_K$, given that $\mathbf{F}_H$

has the value $\mathbf{F}_H^*$, is defined as

$$\mathscr{P}(\mathbf{F}_K|\mathbf{F}_H=\mathbf{F}_H^*)=\mathscr{P}(\mathbf{F}_H^*,\mathbf{F}_K)/\mathscr{P}(\mathbf{F}_H^*). \qquad (0.23)$$

In MEFDM, § 4, it was shown that the ME distribution of atoms $q^{\mathrm{ME}}(\mathbf{x})$ used to construct $\mathscr{P}^{\mathrm{SP}}(\mathbf{F}^*)$ provides a means of approximating this conditional distribution by a multivariate Gaussian with centre $\mathbf{F}_K^{\mathrm{ME}}$ and covariance matrix $\mathbf{Q}_{KK}$. Here, $\mathbf{F}_K^{\mathrm{ME}}$ is the vector of Fourier coefficients of q^{ME} for reflexions in K, which have been extrapolated from the data $\mathbf{F}_H^*$ by the process of entropy maximization; while the covariance matrix $\mathbf{Q}_{KK}$ is constructed *via* structure-factor algebra from the spectrum of q^{ME} in the same way as matrix $\mathbf{Q}_{HH}=\mathbf{Q}$ [(0.17)] was for the 'basis' reflexions in H. This approximate conditional distribution will be denoted $\mathscr{P}^{\mathrm{SP}}(\mathbf{F}_K|\mathbf{F}_H=\mathbf{F}_H^*)$. It differs from the usual Wilson (1949) distribution, which would correspond to a uniform q^{ME}, in two respects:

(1) it is centred around $\mathbf{F}_K=\mathbf{F}_K^{\mathrm{ME}}$, not $\mathbf{F}_K=\mathbf{0}$;

(2) its covariance matrix is $\mathbf{Q}_{KK}$, not a multiple of the identity matrix.

Therefore, integration of $\mathscr{P}^{\mathrm{SP}}(\mathbf{F}_K|\mathbf{F}_H=\mathbf{F}_H^*)$ with respect to the phases in $\mathbf{F}_K$ will yield a conditional marginal distribution of moduli $\mathscr{P}^{\mathrm{SP}}(|\mathbf{F}_K||\mathbf{F}_H=\mathbf{F}_H^*)$ which will differ from the standard Wilson distribution of moduli, the latter being actually $\mathscr{P}^{\mathrm{SP}}(|\mathbf{F}_K||\mathbf{F}_H=\mathbf{0})$. In summary we may say that ME extrapolation acts as a 'transducer' converting hypotheses about the initial distribution of atoms $m(\mathbf{x})$ and about phase values in $\mathbf{F}_H^*$ into hypotheses about a change in the distribution of the moduli $|\mathbf{F}_K|$ (or intensities $|\mathbf{F}_K|^2$).

The first type of hypothesis cannot be tested directly from measured data, but the second one can. For this purpose [assuming for the moment that $m(\mathbf{x})$ is initially uniform] we define the *likelihood of the phase choices* in $\mathbf{F}_H^*$ as the *conditional marginal probability of the observed moduli* $|\mathbf{F}_K|^{\mathrm{obs}}$,

$$\Lambda(\mathbf{F}_H^*)=\mathscr{P}^{\mathrm{SP}}(|\mathbf{F}_K|^{\mathrm{obs}}|\mathbf{F}_H=\mathbf{F}_H^*). \qquad (0.24)$$

It follows from the fundamental work of Neyman & Pearson (1933*a*, *b*) that the most powerful statistical test for picking the 'best' set of phases for $\mathbf{F}_H^*$ consists of examining the value of the *likelihood ratio* $\Lambda(\mathbf{F}_H^*)/\Lambda(\mathbf{0})$ as a function of these phases, which measures the extent to which the observed values of the yet unphased structure-factor moduli $|\mathbf{F}_K|^{\mathrm{obs}}$ have been made more probable by the assumption that $\mathbf{F}_H=\mathbf{F}_H^*$ rather than $\mathbf{F}_H=\mathbf{0}$. Such tests may be regarded as a generalization of standard tests based on intensity statistics, but they are potentially more powerful because likelihood involves *joint intensity statistics* rather than the usual marginal statistics.

Furthermore, the *a priori* probabilities $\mathscr{P}^{\mathrm{SP}}(\mathbf{F}_H^*)$ may be combined with the likelihoods $\Lambda(\mathbf{F}_H^*)$ through Bayes's theorem to yield the *a posteriori* probability

$$\mathscr{P}_{\mathrm{post}}(\mathbf{F}_H^*)\propto\mathscr{P}^{\mathrm{SP}}(\mathbf{F}_H^*)\times\Lambda(\mathbf{F}_H^*), \qquad (0.25)$$

providing the basis for a full-scale Bayesian approach to phase refinement. Thus the most powerful criterion for measuring the 'goodness' of a set of phases is a linear combination of the Jaynes entropy $\mathscr{S}$, of the Burg entropy $\mathscr{T}=\log\det\mathbf{Q}$, and of the log-likelihood $\log\Lambda$, with relative weights involving the dimensions of the vectors of observations and the number N of atoms; the value of N actually used should be the 'effective N' (MEFDM § 8.3), which may be determined by maximizing the posterior probability (0.25) with respect to N. The simplest instance of this procedure (MEFDM, § 4.2.2; Bricogne, 1988) yields as an approximation the quartet formula of Hauptman (1975), but has considerably greater scope and generality.

In the rest of this paper, the assumptions on the basis of which joint distributions of structure factors will be sought will incorporate many new ingredients (such as molecular boundaries, isomorphous substitutions, known fragments, non-crystallographic symmetries, multiple crystal forms) besides trial phase choices for basis reflexions. The procedure just outlined for deriving likelihood functions will be systematically developed in paper II, and will lead in paper III to *likelihood ratio tests* for the detection and characterization of these new elements, and for refining the initial trial phases under these extra assumptions. This task will be facilitated by the fact that, however complex these assumptions become, the j.p.d. of structure factors will always be obtained in the *same* functional form summarized in (0.15), (0.16) and (0.17).

1. Homogeneous structures

The simplest manifestation of the chemical identity of atoms is that the observable structure factors are sums not of the trigonometric contributions themselves, but of their product by atomic scattering factors. This section will extend the derivation of $\mathscr{P}^{\mathrm{SP}}(\mathbf{F})$ for homogeneous (*i.e.* equal-atom) structures to the case of an arbitrary complex scattering factor. While being relatively straightforward, this first generalization already brings to light interesting phenomena concerning normalization and determinantal inequalities.

1.1. *Normal scatterers*

Let us consider a homogeneous structure in which each of the N atoms has a scattering factor $f=f(\mathbf{h})$ which may be an arbitrary *real* number, positive or negative. Let a set of unique non-origin reflexions H be chosen, and let the observable structure-factor values attached to them be arranged into a vector $\mathbf{F}$ of dimension $n=2n_a+n_c$ as in § 0.0.

Let $\mathbf{X}(\mathbf{x})$ be the random vector of contributions to $\mathbf{F}$ originating from an atom placed at $\mathbf{x}$. Then

$$\mathbf{X}(\mathbf{x}) = \mathbf{f}\boldsymbol{\xi}(\mathbf{x}) \tag{1.1}$$

where $\boldsymbol{\xi}(\mathbf{x})$ is defined by (0.3) and $\mathbf{f}$ is the (diagonal) matrix of scattering factors for the various reflexions $\mathbf{h} \in H$. The distribution $P(\mathbf{X})$ of $\mathbf{X}$ has a m.g.f. Z given by

$$Z(\mathbf{u}) = \int_D m(\mathbf{x})\, e^{\mathbf{u}.\mathbf{X}(\mathbf{x})}\, \mathrm{d}^3\mathbf{x}. \tag{1.2}$$

But $\mathbf{u}.\mathbf{X} = \mathbf{u}.(\mathbf{f}\boldsymbol{\xi}) = (\mathbf{f}^T\mathbf{u}).\boldsymbol{\xi}$, so that

$$Z(\mathbf{u}) = M(\mathbf{t}) \text{ with } \mathbf{t} = \mathbf{f}^T\mathbf{u}, \tag{1.3}$$

M being the m.g.f. of distribution $p(\boldsymbol{\xi})$ defined in (0.5). The transposition operation in (1.3) may seem superfluous at this stage, since $\mathbf{f}$ is diagonal, but this expression has the advantage that it will remain valid in this form for general $\mathbf{f}$ later. The result (1.3) was obtained by Klug (1958), the rearrangement on page 518 of his paper being equivalent to this transposition.

Since there are N independent identical random atoms, the m.g.f. for the j.p.d. $\mathscr{P}(\mathbf{F})$ is

$$\mathscr{Z}(\mathbf{u}) = Z^N(\mathbf{u}) \tag{1.4}$$

and the standard saddlepoint argument may be invoked since, as $\mathbf{x}$ runs through D, the various vectors $\mathbf{X}(\mathbf{x})$ span the whole of $\mathbb{R}^n$. Thus, recentring $\mathscr{P}$ near $\mathbf{F}^*$ amounts to updating the prior distribution of atoms $m(\mathbf{x})$ to

$$q^{\mathrm{ME}}(\mathbf{x}) = [m(\mathbf{x})/Z(\boldsymbol{\lambda})]\, e^{\boldsymbol{\lambda}.\mathbf{f}\boldsymbol{\xi}(\mathbf{x})} \tag{1.5}$$

with

$$Z(\boldsymbol{\lambda}) = \int_D m(\mathbf{x})\, e^{\boldsymbol{\lambda}.\mathbf{f}\boldsymbol{\xi}(\mathbf{x})}\, \mathrm{d}^3\mathbf{x} \tag{1.6}$$

and

$$\boldsymbol{\nabla}_{\boldsymbol{\lambda}}(\log \mathscr{Z}) = \mathbf{F}^*. \tag{1.7}$$

The saddlepoint approximation to $\mathscr{P}(\mathbf{F}^*)$ is then

$$\mathscr{P}^{\mathrm{SP}}(\mathbf{F}^*) = e^{\mathscr{S}}[\det(2\pi\mathbf{Q})]^{-1/2}, \tag{1.8}$$

where

$$\mathscr{S} = \log \mathscr{Z} - \boldsymbol{\lambda}.\mathbf{F}^* = N\mathscr{S}_m(q^{\mathrm{ME}}) \tag{1.9}$$

is the total relative entropy, and $\mathbf{Q}$ is the Hessian matrix of $\log \mathscr{Z}$ at the saddlepoint

$$\mathbf{Q} = \boldsymbol{\nabla}^2(\log \mathscr{Z}). \tag{1.10}$$

It may be noted that, if $\boldsymbol{\tau} = \mathbf{f}^T\boldsymbol{\lambda}$, then

$$\boldsymbol{\nabla}_{\boldsymbol{\lambda}}(\cdot) = \mathbf{f}\boldsymbol{\nabla}_{\boldsymbol{\tau}}(\cdot) \tag{1.11}$$

$$\boldsymbol{\nabla}^2_{\boldsymbol{\lambda}\boldsymbol{\lambda}}(\cdot) = \mathbf{f}\boldsymbol{\nabla}^2_{\boldsymbol{\tau}\boldsymbol{\tau}}(\cdot)\mathbf{f}^T, \tag{1.12}$$

so that (1.10) may be written

$$\mathbf{Q} = \mathbf{f}\boldsymbol{\nabla}^2_{\boldsymbol{\tau}\boldsymbol{\tau}}(\log M)\mathbf{f}^T, \tag{1.10a}$$

showing that $\mathbf{Q}$ may be calculated from the knowledge of the trigonometric covariance matrix $\boldsymbol{\nabla}^2(\log M)$ (§ 0.2 and Appendix) and of the scattering-factor matrix $\mathbf{f}$.

Furthermore, (1.7) may then be rewritten

$$\boldsymbol{\nabla}_{\boldsymbol{\tau}}(\log M) = (N\mathbf{f})^{-1}\mathbf{F}^* = \mathbf{U}^*, \tag{1.7a}$$

where $\mathbf{U}^*$ is the vector of *normalized* structure factors. Similarly,

$$\mathscr{S} = N(\log M - \boldsymbol{\tau}.\mathbf{U}^*) \tag{1.9a}$$

is the same entropy as would be calculated by constructing q^{ME} from the $\mathbf{U}^*$ data rather than the $\mathbf{F}^*$ data. Finally, since

$$\mathrm{d}^n\mathbf{F} = [\det(N\mathbf{f})]\, \mathrm{d}^n\mathbf{U}$$

$$\det[\boldsymbol{\nabla}^2_{\boldsymbol{\lambda}\boldsymbol{\lambda}}(\log Z^N)] = (\det \mathbf{f})^2 \det[\boldsymbol{\nabla}^2_{\boldsymbol{\tau}\boldsymbol{\tau}}(\log M^N)],$$

we may write

$$\mathscr{P}^{\mathrm{SP}}(\mathbf{F}^*)\, \mathrm{d}^n\mathbf{F} = \mathscr{P}^{\mathrm{SP}}(\mathbf{U}^*)\, \mathrm{d}^n\mathbf{U},$$

where the latter is given by (0.15). Thus we might have dealt with this case by first 'normalizing' $\mathbf{F}^*$ to $\mathbf{U}^*$, then solving the problem for unit point atoms for the $\mathbf{U}^*$ data by the methods of § 0.

It may also be noted that, because the trigonometric covariance matrix $\boldsymbol{\nabla}^2_{\boldsymbol{\tau}\boldsymbol{\tau}}(\log M)$ is positive definite (see § 0.2), it follows by (1.10) and (1.12) that $\mathbf{Q}$ has the same property even if $\mathbf{f}$ contains negative values. This implies that determinantal inequalities will exist among the structure factors in spite of the fact that the electron (or scattering-length) density may be negative. It is thus the positivity of the probability distribution of scatterers $m(\mathbf{x})$, *not* of the electron density $\rho(\mathbf{x})$, which is the basis for determinantal inequalities.

1.2. *Anomalous scatterers*

Let us now assume that each of the N equal atoms has a complex scattering factor $f = f^R + if^I$, f^R and f^I being known functions of $\mathbf{h}$, as can occur with X-rays or with neutrons. The vector $\mathbf{F}$ of symmetry-unique observable structure factors attached to reflexions $\mathbf{h} \in H$ now has dimension $4n_a + n_c$, and its components may be ordered by placing first, in an obvious notation, the real components $(A^+_{\mathbf{h}}, B^+_{\mathbf{h}}, A^-_{\mathbf{h}}, B^-_{\mathbf{h}})$ for acentric reflexions, followed by the real components $C_{\mathbf{h}}$ for the centric reflexions.

Let $\mathbf{X}(\mathbf{x})$ be the random vector of contributions to $\mathbf{F}$ from an atom placed at random position $\mathbf{x} \in D$, and let its components a^+, b^+, a^-, b^- and c be ordered as those of $\mathbf{F}$. The relation giving $\mathbf{X}(\mathbf{x})$ in terms of $\boldsymbol{\xi}(\mathbf{x})$ is now

$$\begin{bmatrix} a^+(\mathbf{x}) \\ b^+(\mathbf{x}) \\ a^-(\mathbf{x}) \\ b^-(\mathbf{x}) \end{bmatrix}_{\mathbf{h}} = \begin{bmatrix} f^R & -f^I \\ f^I & f^R \\ f^R & f^I \\ f^I & -f^R \end{bmatrix}_{\mathbf{h}} \begin{bmatrix} \alpha(\mathbf{x}) \\ \beta(\mathbf{x}) \end{bmatrix}_{\mathbf{h}} \tag{1.13a}$$

for all acentric reflexions, and

$$c_{\mathbf{h}}(\mathbf{x}) = f^R(\mathbf{h})\gamma_{\mathbf{h}}(\mathbf{x}) \qquad (1.13b)$$

for the centric reflexions, where $\alpha_{\mathbf{h}}(\mathbf{x})$, $\beta_{\mathbf{h}}(\mathbf{x})$ and $\gamma_{\mathbf{h}}(\mathbf{x})$ are defined in (0.1a, b). This relation may be written more compactly as

$$\mathbf{X}(\mathbf{x}) = \mathbf{f}\boldsymbol{\xi}(\mathbf{x}) \qquad (1.14)$$

where $\mathbf{f}$ is the scattering-factor matrix built from the blocks defined in (1.13a, b). It is a *rectangular* $(4n_a + n_c) \times (2n_a + n_c)$ matrix, which converts the $(2n_a + n_c)$ vector $\boldsymbol{\xi}$ into the $(4n_a + n_c)$ vector $\mathbf{X}$; therefore, as $\mathbf{x}$ runs through D, the vectors $\mathbf{X}(\mathbf{x})$ only span a subspace of dimension $n = 2n_a + n_c$ of the $(4n_a + n_c)$-dimensional space in which they are defined. This phenomenon requires that some caution be exercised prior to using the saddlepoint method since the global log-convexity of the m.g.f. $\mathscr{Z}$, which is essential to guarantee the existence and uniqueness of the saddlepoint, is now in jeopardy. Since this difficulty will recur later, we will go to some length in showing how to overcome it.

We may decompose the total space as an orthogonal direct sum of the image space of $\mathbf{f}$ (the 'allowed' subspace, of dimension n) and of its orthogonal complement (the 'forbidden' subspace, which is also the null subspace of $\mathbf{f}^T$, of dimension $2n_a$), denoting by $\mathrm{Pr}_{\|}$ and $\mathrm{Pr}_{\perp}$ the associated projection operators. A generic vector $\mathbf{X}$ in the total space may be written

$$\mathbf{X} = \mathrm{Pr}_{\|}(\mathbf{X}) \oplus \mathrm{Pr}_{\perp}(\mathbf{X}) = \mathbf{X}_{\|} + \mathbf{X}_{\perp}. \qquad (1.15)$$

If one uses the generalized inverse $\mathbf{f}^{\#} = (\mathbf{f}^T\mathbf{f})^{-1}\mathbf{f}^T$ of $\mathbf{f}$ [see, for instance, Graybill (1969), Chapter 6], these projectors may be 'diagonalized' as

$$\mathrm{Pr}_{\|}(\mathbf{X}) = \mathrm{Pr}_1[\mathbf{f}\mathbf{f}^{\#}\mathbf{X}] \qquad (1.16a)$$

$$\mathrm{Pr}_{\perp}(\mathbf{X}) = \mathrm{Pr}_2[(\mathbf{I} - \mathbf{f}\mathbf{f}^{\#})\mathbf{X}], \qquad (1.16b)$$

where $\mathbf{I}$ is the identity matrix in the total space, and Pr_1 (Pr_2) is a diagonal projector which projects an $(n + 2n_a)$ vector by taking its first n (last $2n_a$) coordinates. The integration volume form may then be written

$$\mathrm{d}^{n+2n_a}\mathbf{X} = \mathrm{d}^n\mathbf{X}_{\|}\,\mathrm{d}^{2n_a}\mathbf{X}_{\perp}. \qquad (1.17)$$

The fact that a vector $\mathbf{X}(\mathbf{x})$ defined by (1.14) lies in the image space of $\mathbf{f}$ may be written, in a basis adapted to this decomposition,

$$\mathbf{X}(\mathbf{x}) = \begin{bmatrix} \mathbf{X}_{\|}(\mathbf{x}) \\ \vdots \\ \mathbf{0} \end{bmatrix} \quad \text{or} \quad \mathbf{X}_{\perp}(\mathbf{x}) = \mathbf{0}, \qquad (1.18)$$

so that the distribution P of $\mathbf{X}(\mathbf{x})$ is of the form

$$P(\mathbf{X}) = P_{\|}(\mathbf{X}_{\|}) \otimes \delta(\mathbf{X}_{\perp}) \qquad (1.19)$$

where the Dirac distribution $\delta(\mathbf{X}_{\perp})$ [see Schwartz (1966) as a general reference] has the effect of concentrating P in the allowed subspace. Using the dual decomposition $\mathbf{u} = \mathbf{u}_{\|} + \mathbf{u}_{\perp}$ for the vector $\mathbf{u}$ of carrying variables, we may write the m.g.f. Z of P as a tensor product

$$Z(\mathbf{u}) = Z_{\|}(\mathbf{u}_{\|}) \otimes \mathbf{1}(\mathbf{u}_{\perp}) \qquad (1.20)$$

where $\mathbf{1}(\mathbf{u}_{\perp})$ is a function having identically the value 1 in the forbidden subspace, and where

$$Z_{\|}(\mathbf{u}_{\|}) = \int_D m(\mathbf{x}) \exp[\mathbf{u}_{\|}.\mathbf{X}_{\|}(\mathbf{x})]\,\mathrm{d}^3\mathbf{x}. \qquad (1.21)$$

Since the collection of all vectors $\mathbf{X}_{\|}(\mathbf{x})$ as $\mathbf{x}$ runs through D now spans the whole of $\mathbb{R}^n$, $\log Z_{\|}$ has recovered the strict convexity property which $\log Z$ had lost.

The m.g.f. $\mathscr{Z}$ for $\mathscr{P}(\mathbf{F})$ is then

$$\mathscr{Z}(\mathbf{u}) = [Z_{\|}(\mathbf{u}_{\|})]^N \otimes \mathbf{1}(\mathbf{u}_{\perp}), \qquad (1.22)$$

and Fourier-transforming back will turn the factor $\mathbf{1}(\mathbf{u}_{\perp})$ into a factor $\delta(\mathbf{F}_{\perp})$. Therefore, as expected, $\mathscr{P}$ is concentrated in the same subspace as P, and hence may be written

$$\mathscr{P}(\mathbf{F}) = \mathscr{P}_{\|}(\mathbf{F}_{\|}) \otimes \delta(\mathbf{F}_{\perp}) \qquad (1.23)$$

where

$$\mathscr{P}_{\|}(\mathbf{F}_{\|}) = (2\pi)^{-n} \int_{\mathbb{R}^n} \exp\{N[\log Z_{\|}(i\mathbf{u}_{\|}) - i\mathbf{u}_{\|}.(\mathbf{F}_{\|}/N)]\}\,\mathrm{d}^n\mathbf{u}_{\|}. \qquad (1.24)$$

The saddlepoint argument may now be invoked to approximate $\mathscr{P}_{\|}$ at $\mathbf{F}_{\|}$. It suffices to complexify $\mathbf{u}_{\|}$ into $\mathbf{v}_{\|} - i\boldsymbol{\lambda}_{\|}$, then to choose for $\boldsymbol{\lambda}_{\|}$ the unique value such that

$$\boldsymbol{\nabla} \log[(Z_{\|})^N] = \mathbf{F}_{\|}^*, \qquad (1.25)$$

thus causing the integrand to have both a maximum modulus and a stationary phase at $\mathbf{u}_{\|}^*$. The rest of the calculation follows the same course as in § 1.1 and yields results of identical formal appearance, but two points of difference must be emphasized.

(1) Equation (1.7) can only be solved if $\mathbf{F}^*$ lies in the allowed subspace, *i.e.* by (1.16a) if

$$(\mathbf{f}\mathbf{f}^{\#})\mathbf{F}^* = \mathbf{F}^*, \qquad (1.26)$$

in which case it is equivalent to (1.25); expressions (1.5) to (1.9) are then valid, but the vector $\boldsymbol{\lambda}$ involved is such that $\boldsymbol{\lambda}_{\perp} = \mathbf{0}$.

(2) The matrix $\mathbf{Q}$ whose determinant intervenes in the expression (1.8) for $\mathscr{P}^{\mathrm{SP}}$ is the Hessian matrix of $\log[(Z_{\|})^N]$ with respect to the coordinates $\boldsymbol{\lambda}_{\|}$ in the allowed subspace:

$$\mathbf{Q} = \boldsymbol{\nabla}^2_{\boldsymbol{\lambda}_{\|}\boldsymbol{\lambda}_{\|}}(\log \mathscr{Z}_{\|}). \qquad (1.27)$$

It is positive definite, while $\boldsymbol{\nabla}^2_{\boldsymbol{\lambda}\boldsymbol{\lambda}}(\log \mathscr{Z})$ is only positive semi-definite. From a computational point of view, $\mathbf{Q}$ (and hence $\det \mathbf{Q}$) may be obtained by applying eigenvalue filtering to $\boldsymbol{\nabla}^2_{\boldsymbol{\lambda}\boldsymbol{\lambda}}(\log \mathscr{Z})$ so as to isolate its

'regular part'. We will therefore introduce the notation

$$\mathbf{Q} = \text{reg}\, \boldsymbol{\nabla}^2_{\lambda\lambda}(\log \mathscr{Z}) \qquad (1.28a)$$

$$\det \mathbf{Q} = \det\,[\text{reg}\, \boldsymbol{\nabla}^2_{\lambda\lambda}(\log \mathscr{Z})], \qquad (1.28b)$$

which will be useful later in denoting a general prescription to deal with degeneracies such as those encountered in this section.

We may now examine these results from the viewpoint of normalization procedures. The equivalent of (1.3) is here

$$Z_{\|}(\mathbf{u}_{\|}) = M(\mathbf{t}) \quad \text{with} \quad \mathbf{t} = \mathbf{f}^T\mathbf{u}, \qquad (1.29)$$

while the vector $\mathbf{u}_{\|}$ itself may be written $\mathbf{u}_{\|} = (\mathbf{f}^{I})^{\#}\mathbf{t}$. Thus we could have applied the saddlepoint argument to M rather than Z, and obtained the saddlepoint condition (1.25) in the form

$$\boldsymbol{\nabla}_{\tau}(\log M) = \mathbf{f}^{\#}(\mathbf{F}^*_{\|}/N) = (N\mathbf{f})^{\#}\mathbf{F}^* = \mathbf{U}^*, \quad (1.30)$$

where the vector $\mathbf{U}^*$ of normalized structure factors, obtained by means of the generalized inverse $(N\mathbf{f})^{\#}$, automatically lies in the allowed subspace. Furthermore,

$$\mathrm{d}^n\mathbf{F}_{\|} = \{\det\,[(N\mathbf{f})^T(N\mathbf{f})]\}^{1/2}\,\mathrm{d}^n\mathbf{U},$$

while

$$\det \mathbf{Q} = \det\,\{\,\text{reg}\,[\mathbf{f}\boldsymbol{\nabla}^2_{\tau\tau}(\log M^N)\mathbf{f}^T]\};$$

hence the effects of the scattering factors on the normalization factor and on the volume integration form cancel, leaving

$$\mathscr{P}^{\mathrm{SP}}(\mathbf{F}^*)\,\mathrm{d}^n\mathbf{F} = \mathscr{P}^{\mathrm{SP}}(\mathbf{U}^*)\,\mathrm{d}^n\mathbf{U}.$$

Thus again we might have dealt with this case by normalizing $\mathbf{F}^*$ to $\mathbf{U}^*$ and then proceeding as if the structure were made of unit point atoms. The normalization formula

$$\mathbf{U}^* = (N\mathbf{f})^{\#}\mathbf{F}^*$$

is, however, less obvious than previously: it is actually equivalent to a least-squares fit to the data $\mathbf{F}^*$, enforcing the known value f^I/f^R for the ratio of their Hermitian-antisymmetric to their Hermitian-symmetric parts. In particular, this normalization cannot be performed on the moduli data alone, since it acts on the phase information as well.

Finally, the positive definiteness of $\mathbf{Q}$ as given by (1.28*a*) points to the existence of determinantal inequalities among structure factors, even though the electron (or neutron scattering-length) density $\rho(\mathbf{x})$ now has *complex* values, once more emphasizing the fact that these inequalities follow from the positivity of $m(\mathbf{x})$, not of $\rho(\mathbf{x})$.

2. Heterogeneous structures

Crystal structures usually contain several chemical types of atoms, with different numbers of electrons or different neutron scattering lengths. In small-molecule direct methods this heterogeneity is dealt with through the process of normalization. The procedure used, however, assumes a uniform distribution of all atom types, which is often unsatisfactory and requires *ad hoc* amendments.

Another, even more extreme, type of heterogeneity, which might be termed physical rather than chemical, is encountered with the existence of solvent channels in macromolecular crystals: ordinary 'sharp' atoms are found in the macromolecule itself, while the solvent regions are filled with 'fuzzy' atoms having a large thermal parameter, the prior estimates for the distributions of the two categories of atoms being strongly non-uniform and mutually non-overlapping.

Thus a proper statistical model for heterogeneous structures should accommodate not only different types of atoms, but also different non-uniform spatial distributions for each atom type. It is the purpose of this section to develop such a model within the framework of the saddlepoint approximation.

2.1. *General theory*

Let us consider a heterogeneous structure made of c distinct species of atoms, containing N_j atoms of species j distributed identically and independently with prior distribution $m_j(\mathbf{x})$, for each $j = 1, \ldots, c$. Atoms of different species are also assumed to be statistically independent. The scattering factor for species j is a known function $f_j(\mathbf{h})$ of $\mathbf{h}$, which may take arbitrary complex values. Finally, let a set H of n_a acentric and n_c centric unique non-origin reflexions be chosen, with the same conventions as in §§ 0 and 1 for ordering into vectors the real components of structure-factor data attached to them.

Let $\boldsymbol{\xi}(\mathbf{x})$ denote the random vector of trigonometric structure-factor expressions for a random position $\mathbf{x} \in D$. For each species j, the prior distribution $m_j(\mathbf{x})$ induces a different distribution $p_j(\boldsymbol{\xi})$ for this vector, whose m.g.f. M_j is given by

$$M_j(\mathbf{t}) \equiv \int_{\mathbb{R}^n} p_j(\boldsymbol{\xi})\, e^{\mathbf{t}.\boldsymbol{\xi}}\,\mathrm{d}^n\boldsymbol{\xi} = \int_D m_j(\mathbf{x})\, e^{\mathbf{t}.\boldsymbol{\xi}(\mathbf{x})}\,\mathrm{d}^3\mathbf{x}. \qquad (2.1)$$

Let $\mathbf{X}_j(\mathbf{x})$ now denote the random vector of contributions of an atom of species j placed at $\mathbf{x} \in D$ to the observable structure-factor vector $\mathbf{F}$ for the whole structure. As in § 1,

$$\mathbf{X}_j(\mathbf{x}) = \mathbf{f}_j\boldsymbol{\xi}(\mathbf{x}), \qquad (2.2)$$

where $\mathbf{f}_j$ is the scattering-factor matrix for atomic species j. The m.g.f. Z_j for the distribution P_j of $\mathbf{X}_j$ is

$$Z_j(\mathbf{u}) = M_j(\mathbf{t}) \quad \text{with} \quad \mathbf{t} = \mathbf{f}_j^T\mathbf{u}. \qquad (2.3)$$

The initial assumptions on the heterogeneous composition of the structure may be reflected by writing

$$\mathbf{F} = \sum_{j=1}^{c} \sum_{l_j=1}^{N_j} \mathbf{X}_j^{[l_j]}, \tag{2.4}$$

where all the summands are statistically independent and each $\mathbf{X}_j^{[l_j]}$ is distributed according to P_j. The global m.g.f. $\mathscr{Z}$ of the joint distribution $\mathscr{P}(\mathbf{F})$ is therefore

$$\mathscr{Z}(\mathbf{u}) = \prod_{j=1}^{c} [Z_j(\mathbf{u})]^{N_j} = \prod_{j=1}^{c} [M_j(\mathbf{f}_j^T\mathbf{u})]^{N_j}. \tag{2.5}$$

At this stage, we encounter the same difficulty as in § 1.2, namely that, as shown by (2.3), each $Z_j(\mathbf{u})$ depends on $\mathbf{u}$ only through $\mathbf{f}_j^T\mathbf{u}$, so that $\log Z_j$ has vanishing curvature in the orthogonal complement of the image subspace of $\mathbf{f}_j$. As the difference matrices $\mathbf{f}_j$ may have different image subspaces in $\mathbb{R}^{4n_a+n_c}$, the curvature of the weighted sum $\log \mathscr{Z}$ will be positive in the vector sum of the various image subspaces, which may be directly obtained as the image subspace of the average scattering matrix $\mathbf{f}$ defined by

$$N\mathbf{f} = \sum_{j=1}^{c} N_j\mathbf{f}_j. \tag{2.6}$$

We may therefore use (1.15) and (1.16*a*, *b*) with this matrix $\mathbf{f}$ to define a decomposition of the total space in such a way that $\nabla^2(\log \mathscr{Z})$ be positive definite in the allowed subspace labelled $\|$, while the zero curvature occurs in the forbidden subspace labelled $\perp$. In contrast to the case of a single atom type, the dimension n of the allowed subspace may be any integer between $2n_a + n_c$ and $4n_a + n_c$, depending on the degree of heterogeneity of the structure.

Having performed this decomposition, we have, as in § 1.2,

$$\mathscr{Z}(\mathbf{u}) = \mathscr{Z}_{\|}(\mathbf{u}_{\|}) \otimes \mathbf{1}(\mathbf{u}_{\perp}), \tag{2.7}$$

with

$$\mathscr{Z}_{\|}(\mathbf{u}_{\|}) = \prod_{j=1}^{c} [Z_{j\|}(\mathbf{u}_{\|})]^{N_j}; \tag{2.8}$$

hence, after Fourier inversion

$$\mathscr{P}(\mathbf{F}) = \mathscr{P}_{\|}(\mathbf{F}_{\|}) \otimes \delta(\mathbf{F}_{\perp}) \tag{2.9}$$

where, if we put $\zeta_j = N_j/N$,

$$\mathscr{P}_{\|}(\mathbf{F}_{\|}) = (2\pi)^{-n} \int_{\mathbb{R}^n} \exp\left\{ N\left[\sum_{j=1}^{c} \zeta_j \log Z_{j\|}(i\mathbf{u}_{\|}) - i\mathbf{u}_{\|}\cdot(\mathbf{F}_{\|}/N) \right] \right\} \mathrm{d}^n\mathbf{u}_{\|}. \tag{2.10}$$

The saddlepoint argument may now be applied by letting N and all the N_j tend to infinity while keeping the ratios ζ_j constant: complexifying $\mathbf{u}_{\|}$ to $\mathbf{v}_{\|} - i\boldsymbol{\lambda}_{\|}$ shows that all contributions to $\mathscr{P}_{\|}(\mathbf{F}_{\|})$ will come from the immediate vicinity of $\mathbf{u}_{\|}^* = \mathbf{0} - i\boldsymbol{\lambda}_{\|}$, where $\boldsymbol{\lambda}_{\|}$ is the unique point at which the saddlepoint condition

$$\nabla \log [(Z_{\|})^N] = \mathbf{F}_{\|}^* \tag{2.11}$$

holds. This will lead to the desired approximation of $\mathscr{P}(\mathbf{F})$ at $\mathbf{F} = \mathbf{F}^*$ if and only if $\mathbf{F}^*$ lies in the allowed subspace, which is equivalent to the condition

$$(\mathbf{f}\mathbf{f}^{\#})\mathbf{F}^* = \mathbf{F}^*. \tag{2.12}$$

In that case, the desired approximation formula will involve updated distributions q_j^{ME} for each atom type j defined by

$$q_j^{\mathrm{ME}}(\mathbf{x}) = [m_j(\mathbf{x})/Z_j(\boldsymbol{\lambda})]\, e^{\boldsymbol{\lambda}\cdot\mathbf{f}_j\boldsymbol{\xi}(\mathbf{x})} \tag{2.13}$$

with

$$Z_j(\boldsymbol{\lambda}) = \int_D m_j(\mathbf{x})\, e^{\boldsymbol{\lambda}\cdot\mathbf{f}_j\boldsymbol{\xi}(\mathbf{x})}\, \mathrm{d}^3(\mathbf{x}), \tag{2.14}$$

where the vector $\boldsymbol{\lambda}$ (which is also the vector of Lagrange multipliers for the entropy maximization problem) is *common* to all the atom types. This vector is of the form

$$\boldsymbol{\lambda} = \boldsymbol{\lambda}_{\|} + \mathbf{0}_{\perp}, \tag{2.15}$$

$\boldsymbol{\lambda}_{\|}$ being defined by (2.11), and hence lies in the allowed dual subspace.

The saddlepoint approximation to $\mathscr{P}(\mathbf{F}^*)$ is then

$$\mathscr{P}^{\mathrm{SP}}(\mathbf{F}^*) = e^{\mathscr{S}}[\det(2\pi\mathbf{Q})]^{-1/2}, \tag{2.16}$$

where

$$\mathscr{S} = \log \mathscr{Z} - \boldsymbol{\lambda}.\mathbf{F}^* = \sum_{j=1}^{c} N_j \mathscr{S}_{m_j}(q_j^{\mathrm{ME}}) \tag{2.17}$$

is the total relative entropy (the N-weighted sum of individual relative entropies), and where

$$\mathbf{Q} = \mathrm{reg}\,[\nabla_{\boldsymbol{\lambda}\boldsymbol{\lambda}}^2(\log \mathscr{Z})] \tag{2.18a}$$

$$= \mathrm{reg}\left\{ \sum_{j=1}^{c} N_j[\mathbf{f}_j \nabla_{\boldsymbol{\tau}\boldsymbol{\tau}}^2(\log M_j)\mathbf{f}_j^T] \right\} \tag{2.18b}$$

is the regular part of the Hessian matrix of $\log \mathscr{Z}$ at the saddlepoint in the allowed subspace.

It is instructive to re-examine these results from the viewpoint of normalization procedures. By virtue of (2.5) and (1.11), the saddlepoint condition (2.11) may be written in the equivalent form

$$\sum_{j=1}^{c} N_j\mathbf{f}_j\nabla_{\boldsymbol{\tau}}(\log M_j) = \mathbf{F}^*. \tag{2.19}$$

If all the prior distributions $m_j(\mathbf{x})$ are the same [say $m(\mathbf{x})$], then all the m.g.f.'s M_j are the same (say M), and the saddlepoint condition may then be cast in the form

$$\nabla_{\boldsymbol{\tau}}(\log M) = \mathbf{U}^*, \tag{2.20}$$

where

$$\mathbf{U}^* = \left(\sum_{j=1}^{c} N_j\mathbf{f}_j \right)^{\#} \mathbf{F}^* = (N\mathbf{f})^{\#}\mathbf{F}^* \tag{2.21}$$

is the vector of unitary structure factors. Thus the normalization operation (2.21) does again reduce the problem to the case of unit point atoms, but at the cost of having to impose throughout the condition that all atom types should be identically distributed. As is well known, this assumption is often inappropriate, and special corrections must be made which have been the subject of extensive investigations, using the basic methods developed by Foster & Hargreaves (1963), Srinivasan & Parthasarathy (1976), Shmueli (1979, 1982), Shumeli & Wilson (1981, 1983) to derive intensity statistics for heterogeneous structures. The approach just presented is much more radical: instead of attempting to overcome the problem of heterogeneity by a *single* normalization operation, it allows a *different* prior distribution for each type of scatterer, and updates it *differently* by (2.13) as phase information is being introduced to recentre $\mathscr{P}$; this explains the term 'multichannel' in the title of this paper. Since all these distributions $m_j(\mathbf{x})$ and $q^{\mathrm{ME}}(\mathbf{x})$ may each range from complete uniformity to various degrees of peakiness, they can accommodate within a single scheme the myriads of specific situations (heavy atoms in general or special positions, *etc.*) which, in various combinations, have been studied in the literature. We may therefore conclude that *normalization should be abandoned for heterogeneous structures*: instead, the j.p.d. of raw (unnormalized) structure factors $\mathbf{F}$ should be calculated directly by the multichannel procedure just presented. The problem of initially estimating the absolute scale factor of the data when non-uniform prior distributions of atoms are assumed will be dealt with by a maximum likelihood method in the sequel to this paper.

Finally, the positive-definite character of matrix $\mathbf{Q}$ (2.18) signals once more that certain determinantal inequalities continue to hold between structure factors, even when the latter emanate from an arbitrary mixture of normal and anomalous scatterers.

2.2. *Effect of heavy atoms*

The difference between the present 'multichannel' approach on the one hand, and the use of a single channel with normalized data on the other, may be illustrated most simply by considering a structure in $P1$ made of two types of atoms with uniform prior distributions $m_1(\mathbf{x}) = m_2(\mathbf{x}) = 1/V$.

Let the value of one structure factor $F_{\mathbf{h}} = |F_{\mathbf{h}}| e^{i\varphi_{\mathbf{h}}}$ be specified. Then the method of MEFDM, § 3.4.2 may be followed to give the single-channel result

$$q^{\mathrm{ME}}(\mathbf{x}) = [VI_0(\kappa)]^{-1} \exp[\kappa \cos(2\pi \mathbf{h}.\mathbf{x} - \varphi_{\mathbf{h}})],$$

where κ satisfies

$$I_1(\kappa)/I_0(\kappa) = |F_{\mathbf{h}}|/(N_1 f_1 + N_2 f_2) = |U_{\mathbf{h}}|;$$

on the other hand, the two-channel calculation runs as follows:

$$Z_1 = I_0(f_1\lambda),\ Z_2 = I_0(f_2\lambda),\ \mathscr{Z} = Z_1^{N_1} Z_2^{N_2},$$

whence

$$q_1^{\mathrm{ME}}(\mathbf{x}) = [VI_0(f_1\lambda)]^{-1} \exp[f_1\lambda \cos(2\pi \mathbf{h}.\mathbf{x} - \varphi_{\mathbf{h}})]$$
$$q_2^{\mathrm{ME}}(\mathbf{x}) = [VI_0(f_2\lambda)]^{-1} \exp[f_2\lambda \cos(2\pi \mathbf{h}.\mathbf{x} - \varphi_{\mathbf{h}})],$$

where λ satisfies

$$N_1 f_1 [I_1(f_1\lambda)/I_0(f_1\lambda)] + N_2 f_2 [I_1(f_2\lambda)/I_0(f_2\lambda)] = |F_{\mathbf{h}}|.$$

The solutions of the equations for κ and λ are summarized in Table 1, which shows that if $f_1 \gg f_2$ then the updated distribution q_1^{ME} can be much sharper than q_2^{ME} because f_i intervenes *multiplicatively in the exponent* of q_j^{ME}. Indeed, introducing the 'titres'

$$w_j = N_j f_j/(N_1 f_1 + N_2 f_2),$$

we may rewrite the two-channel equation as

$$w_1 [I_1(f_1\lambda)/I_0(f_1\lambda)] + w_2 [I_1(f_2\lambda)/I_0(f_2\lambda)] = |U_{\mathbf{h}}|,$$

and Table 1 shows that, as $|U_{\mathbf{h}}|$ approaches or exceeds w_1, the ME updating will suddenly decide to attribute this major non-uniformity mostly to the heavy atom. This behaviour is easily rationalized in terms of entropy, since 'nailing down' one heavy atom costs less entropy than nailing down several light atoms to produce a feature of the same height.

Clearly, the multichannel approach using unnormalized data can lead to substantially different results from those of the single-channel approach using normalized data. Since the ME-updated distributions $q(\mathbf{x})$ are different in the two approaches, they will lead to different approximations for conditional distributions, and hence to different likelihood functions.

The saddlepoint method of approximation was compared with the Edgeworth expansion by Weiss, Shmueli, Kiefer & Wilson (1985) with regard to the effectiveness of the two methods in allowing for extreme atomic heterogeneity in the derivation of intensity statistics. The saddlepoint method was found to be much superior. However, this was a single-channel approach, which would not yield optimal statistical tests if pursued further.

Only the multichannel results will be quantitatively correct: if a single channel is used from normalized data, the ME condition tends to prevent the build up of high electron density at the heavy-atom sites, as if it were an improbable pile up of light atoms. Thus a ME reconstruction of a strongly heterogeneous structure obtained by maximizing the entropy of the electron density map, as was done by Gull, Livesey & Sivia (1987) is of questionable quantitative validity.

Table 1. *Compared solutions of the single-channel and two-channel maximum-entropy equations*

The single-channel and two-channel maximum-entropy equations were solved by a Newton method for a hypothetical heterogeneous structure (one atom with scattering factor 80 and 160 atoms with scattering factor 6) for a range of values of $|E_{\mathbf{h}}|$. The following quantities are tabulated as a function of $|E_{\mathbf{h}}|$:

(i) single-channel case:
- κ, solution of the single-channel ME equation;
- Δ, dynamic range, as a power of 10, of the ME-updated q^{ME};
- S, total entropy;
- $|E_{2\mathbf{h}}^{\mathrm{ME}}|$, ME-extrapolated contribution to $E_{2\mathbf{h}}$;
- $|E_{3\mathbf{h}}^{\mathrm{ME}}|$, ME-extrapolated contribution to $E_{3\mathbf{h}}$;

(ii) two-channel case:
- λ, solution of the two-channel ME equation;
- $f_1\lambda$, equivalent of κ for channel no. 1;
- Δ_1, dynamic range, as a power of 10, of the ME-updated q_1^{ME};
- $f_2\lambda$, equivalent of κ for channel no. 2;
- Δ_2, dynamic range, as a power of 10, of the ME-updated q_2^{ME};
- S, total entropy;
- $|E_{2\mathbf{h}}^{\mathrm{ME}}|$, ME-extrapolated contribution to $E_{2\mathbf{h}}$;
- $|E_{3\mathbf{h}}^{\mathrm{ME}}|$, ME-extrapolated contribution to $E_{3\mathbf{h}}$.

The two approaches give different results. In particular, the dynamic range of the two channels is vastly different, which results in a decreased total entropy and a stronger extrapolation for the 'overtones' 2**h** and 3**h**. These will affect, respectively, the (*a priori*) probability and the (*a posteriori*) likelihood of any assumption concerning the presence or the localization of the heavy atom.

Single-channel calculation

$\lvert E_{\mathbf{h}}\rvert$	κ	Δ	S	$\lvert E_{2\mathbf{h}}^{\mathrm{ME}}\rvert$	$\lvert E_{3\mathbf{h}}^{\mathrm{ME}}\rvert$
0·500	0·147	0·128	−0·547	0·018	0·000
1·000	0·297	0·258	−2·196	0·074	0·004
1·500	0·452	0·393	−4·976	0·168	0·013
2·000	0·616	0·535	−8·936	0·303	0·031
2·500	0·791	0·687	−14·151	0·482	0·063
3·000	0·985	0·856	−20·735	0·711	0·114
3·500	1·206	1·408	−28·853	0·997	0·195
4·000	1·468	1·275	−38·753	1·352	0·317
4·500	1·799	1·562	−50·830	1·796	0·506
5·000	2·253	1·957	−65·769	2·362	0·807

Two-channel calculation

$\lvert E_{\mathbf{h}}\rvert$	λ	$f_1\lambda$	Δ_1	$f_2\lambda$	Δ_2	S	$\lvert E_{2\mathbf{h}}^{\mathrm{ME}}\rvert$	$\lvert E_{3\mathbf{h}}^{\mathrm{ME}}\rvert$
0·500	0·011	0·843	0·732	0·063	0·055	−0·257	0·067	0·009
1·000	0·025	1·980	1·720	0·149	0·129	−1·114	0·255	0·074
1·500	0·046	3·679	3·196	0·276	0·240	−2·854	0·487	0·215
2·000	0·072	5·789	5·029	0·434	0·377	−5·798	0·686	0·358
2·500	0·102	8·139	7·069	0·610	0·530	−10·139	0·879	0·473
3·000	0·134	10·732	9·322	0·805	0·699	−16·022	1·097	0·575
3·500	0·171	13·659	11·864	1·024	0·890	−23·623	1·358	0·681
4·000	0·214	17·086	14·841	1·281	1·113	−33·199	1·678	0·810
4·500	0·267	21·329	18·526	1·600	1·389	−45·149	2·077	0·987
5·000	0·338	27·054	23·499	2·029	1·762	−60·163	2·587	1·252

2.3. *Effect of solvent regions in macromolecular crystals*

The problem of devising a normalization procedure capable of dealing with macromolecular crystals, in that it would allow for the existence of solvent regions, has proved particularly intractable. In view of the remark made at the end of § 2.1 this is not surprising, since the assumption that solvent atoms and macromolecule atoms have the same prior distribution is manifestly absurd: it would amount to using a 'random soup' model in which the macromolecule is broken up and its atoms are uniformly mixed with those of the solvent, whereas in reality the two types of atoms are totally segregated. With the present multichannel approach, however, a satisfactory statistical model can be obtained quite straightforwardly which sheds light on the power of solvent-flattening methods in macromolecular crystallography.

Let us assume that a macromolecule is contained within a subregion U of the asymmetric unit D, and that the complementary region $D-U$ contains solvent. Let χ_U denote the indicator function of U, *i.e.* let $\chi_U(\mathbf{x})$ be 1 for $\mathbf{x}$ in U and 0 otherwise, and let

$$\mathcal{G} = (1/U)\mathcal{F}^{-1}[\chi_U] \tag{2.22}$$

be its normalized inverse Fourier transform, also called the 'interference function'. Let a set H of reflexions be chosen as before.

Let the macromolecule be specified as a heterogeneous structure according to § 2.1, and let the solvent region be filled with N_0 atoms having scattering-factor matrix $\mathbf{f}_0$. These solvent atoms have a very high temperature factor ($B \simeq 80$–200 Å^2) so that $f_0(\mathbf{h})$ falls rapidly as the resolution increases, while the macromolecule atoms have normal tem-

perature factors ($B \simeq 8$–20 Å^2). In the absence of any phase information, a sensible choice for the prior distributions of these atoms is

$$m_0(\mathbf{x}) = (D-U)^{-1}\chi_{D-U}(\mathbf{x})$$
for the solvent atoms,

and

$$m_j(\mathbf{x}) = U^{-1}\chi_U(\mathbf{x})$$
for the macromolecule atoms,

where $\chi_{D-U} = 1 - \chi_U$, the same notation being used to denote a region of space and its volume.

The basic m.g.f.'s are

$$M_0(\mathbf{t}) = (D-U)^{-1}\int_D \chi_{D-U}(\mathbf{x})\, e^{\mathbf{t}.\boldsymbol{\xi}(\mathbf{x})}\, \mathrm{d}^3\mathbf{x}$$

$$M_j(\mathbf{t}) = U^{-1}\int_D \chi_U(\mathbf{x})\, e^{\mathbf{t}.\boldsymbol{\xi}(\mathbf{x})}\, \mathrm{d}^3\mathbf{x}, \quad j = 1, \ldots, c,$$

in terms of which we may write the global m.g.f. as

$$\mathscr{Z}(\mathbf{u}) = [M_0(\mathbf{f}_0^T\mathbf{u})]^{N_0}\prod_{j=1}^{c}[M_j(\mathbf{f}_j^T\mathbf{u})]^{N_j}. \quad (2.23)$$

Degeneracies introduced by complex scattering-factor matrices can be dealt with as in equations (2.7) to (2.18). Further degeneracies may, however, arise if H contains a sizeable fraction of all reflexions to a given resolution, because of the 'geometric redundancy' between structure-factor components introduced by the molecular envelope U: as shown earlier (Crowther, 1967; Bricogne, 1974), the vectors $\{\boldsymbol{\xi}(\mathbf{x})|\mathbf{x}\in U\}$ only span a subspace $(\mathscr{H}_1)$ of relative dimension U/D of the whole structure-factor space, while the vectors $\{\boldsymbol{\xi}(\mathbf{x})|\mathbf{x}\in D-U\}$ span the complementary subspace $(\mathscr{H}_0)$ of relative dimension $(D-U)/D$, $(\mathscr{H}_0)$ and $(\mathscr{H}_1)$ being the eigenspaces for eigenvalues 0 and 1 respectively of Crowther's (1967) **H** matrix. As a result, the Hessian matrices $\nabla^2(\log M_j)$ and $\nabla^2(\log M_0)$ will have near-zero curvature in subspaces $(\mathscr{H}_0)$ and $(\mathscr{H}_1)$ respectively. The Hessian matrix whose strict positive definiteness is essential to the saddlepoint method is, by (2.18),

$$\mathbf{Q} = \mathrm{reg}\left\{N_0\mathbf{f}_0\nabla^2(\log M_0)\mathbf{f}_0^T + \sum_{j=1}^{c} N_j[\mathbf{f}_j\nabla^2(\log M_j)\mathbf{f}_j^T]\right\}. \quad (2.24)$$

Its two summands have complementary regular subspaces $(\mathscr{H}_0)$ and $(\mathscr{H}_1)$, so that **Q** is in principle non-singular. At high resolution, however, f_0 is vanishingly small so that the first summand virtually disappears, leaving **Q** almost singular in $(\mathscr{H}_0)$. This can be dealt with through a further regularization of **Q** by means of eigenvalue filtering to remove the worst parts of $(\mathscr{H}_0)$; this procedure may be cast in the same form as that of § 1.2 using generalized inverses, since the **H** matrix, being a *projector* (Bricogne, 1974), is its own generalized inverse (Graybill, 1969). The recentring data $\mathbf{F}^*$ must then be accordingly projected into the allowed subspace $(\mathscr{H}_1)$ prior to seeking to fulfil the saddlepoint condition (2.11).

Proceeding to (2.13), we obtain the updated ME distributions for the various types of atoms as

$$q_0^{\mathrm{ME}}(\mathbf{x}) = \chi_{D-U}(\mathbf{x})\, e^{\boldsymbol{\lambda}.\mathbf{f}_0\boldsymbol{\xi}(\mathbf{x})}/[(D-U)Z_j(\boldsymbol{\lambda})]$$

$$q_j^{\mathrm{ME}}(\mathbf{x}) = \chi_U(\mathbf{x})\, e^{\boldsymbol{\lambda}.\mathbf{f}_j\boldsymbol{\xi}(\mathbf{x})}/[UZ_j(\boldsymbol{\lambda})], \quad j = 1, \ldots, c.$$

Because the scattering-factor matrices intervene *multiplicatively in the exponents*, and because of the rapid fall-off of $f_0(\mathbf{h})$ with increasing resolution, q_0^{ME} will remain very smooth, and all the fine detail specified by $\mathbf{F}^*$ will appear within the region U containing the macromolecule. This is mirrored by the fact that the covariance matrix **Q** permits large fluctuations in the subspace spanned by $\mathbf{X}(\mathbf{x})$ for $\mathbf{x}\in U$, but very little fluctuation in the complementary subspace: therefore the conditional distributions $\mathscr{P}^{\mathrm{SP}}(\mathbf{F}_K|\mathbf{F}_H = \mathbf{F}_H^*)$ will discourage the build up of further detail outside U (see MEFDM, § 4.2.1), and hence will provide a statistical technique for enforcing *in advance* solvent flatness during phase extension. This is clearly preferable to enforcing solvent flatness *a posteriori* by iteratively masking the electron density map by the envelope function $\chi_U(\mathbf{x})$, hoping to find a fixed point for that iterative procedure (Bricogne, 1974; Schevitz, Podjarny, Zwick, Hughes & Sigler, 1981; Wang, 1985; Leslie, 1987).

Thus the special difficulties, due to geometric redundancies, encountered here in dealing with solvent regions have as a positive counterpart a very substantial reward: by structure-factor algebra (see Appendix) the covariances between contributions emanating from neighbouring reciprocal-lattice points can be a substantial fraction of unity, since they are obtained as sample values of the interference function $\mathscr{G}$ near its origin peak. As the number N of atoms increases, the width and strength of this origin peak remains approximately constant in reciprocal-lattice units, because biological macromolecules remain globular as they get bigger, instead of becoming fractal objects. Therefore the large covariances due to solvent regions remain close to unity as N increases, which in classical direct methods would correspond to the existence of $|E|$ values of order $N^{1/2}$! As was argued in MEFDM, § 8.3, it is mainly this circumstance which invalidates previous conclusions that probabilistic methods should be expected to be powerless on macromolecules: these strong nearest-neighbour couplings turn the set of unknown phases into an interacting system akin to an 'inhomogeneous Ising model' or 'spin glass', whose tendency to exhibit critical behaviour will greatly assist the propagation of phase information.

As a corollary to its ability to extrapolate structure factors in a manner dependent on the choice of U,

the above statistical model of solvent regions will yield likelihood functions sensitive to that choice. Likelihood-ratio tests will therefore become available to evaluate the plausibility of various choices for U, thus providing a quantitative version of the usual 'packing considerations' and R-factor calculations, or a test for guiding the connectivity-tracing method of Bhat & Blow (1982) in choosing which regions to include in its molecular envelope. These tests will be much more stringent if a contrast-variation series is available (§ 3.3).

3. Isomorphous families of heterogeneous structures

The most common phase-determination procedures in macromolecular crystallography consist of attempting to obtain, from a given unknown structure, several sets of diffraction intensity data which differ in that the scattering power of certain subsets of atoms varies from one data set to another. The methods of multiple isomorphous replacement (MIR) (Green, Ingram & Perutz, 1954; Dickerson, Kendrew & Strandberg, 1961; Blundell & Johnson, 1976) and multi-wavelength anomalous scattering or MWAS (Phillips & Hodgson, 1980) vary the scattering power of localized substituents, while the contrast variation method (Bragg & Perutz, 1952) modifies that of the solvent. The crystal structures from which such data sets can be obtained will be said to constitute an 'isomorphous family', the term isomorphous implying the preservation of a common crystal lattice and symmetry. The multichannel approach will now be extended to such families, by assigning a 'data channel' to each of its members.

3.1. *General theory*

Let us abstract the various situations mentioned above by considering simultaneously d intensity data sets originating from an isomorphous family of d heterogeneous structures made up of c distinct species of atoms. Each member $k = 1, \ldots, d$ of the family consists of N_j atoms of species j, identically and independently distributed with prior distribution $m_j(\mathbf{x})$ $(j = 1, \ldots, c)$ which is the *same* for all k. Each atom of species j contributes to the vector $\mathbf{F}_k$ of observable structure-factor values for member k through a scattering-factor matrix $\mathbf{f}_{jk}$ which may be different for different values of k; in particular, if scatterer type j is absent from structure k, then $\mathbf{f}_{jk} = \mathbf{0}$. These matrices, connecting the jth 'scatterer channel' to the kth 'data channel', are the basis devices in this formalism for describing the way in which the members of an isomorphous family share the same scatterers but with different scattering powers. We may take advantage of our ability to deal with negative scattering factors, and use equal mixtures of positive and negative 'clutter' atoms to model statistically the type of non-isomorphism caused by local distortions of the native structure.

Let $\mathbf{F}$ denote the global vector of simultaneously observable structure-factor values attached to a set H of reflexions, across the entire family of d structures. We write

$$\mathbf{F} = \bigoplus_{k=1}^{d} \mathbf{F}_k \tag{3.1}$$

where $\oplus$ indicates a columnwise direct sum operation (*i.e.* the concatenation of column vectors).

Let random vector $\boldsymbol{\xi}(\mathbf{x})$ and the m.g.f.'s M_j be defined as in § 2.1, and let $\mathbf{X}_{jk}(\mathbf{x})$ denote the random vector of contributions of an atom of species j to the observable structure-factor vector $\mathbf{F}_k$ for member k of the family:

$$\mathbf{X}_{jk}(\mathbf{x}) = \mathbf{f}_{jk}\boldsymbol{\xi}(\mathbf{x}), \quad j = 1, \ldots, c; \quad k = 1, \ldots, d. \tag{3.2}$$

The contribution of that atom to the global vector $\mathbf{F}$ may then be written

$$\mathbf{X}_j(\mathbf{x}) = \mathbf{f}_j\boldsymbol{\xi}(\mathbf{x}) \tag{3.3}$$

where

$$\mathbf{X}_j(\mathbf{x}) = \bigoplus_{k=1}^{d} \mathbf{X}_{jk}(\mathbf{x}) \tag{3.4}$$

and where

$$\mathbf{f}_j = \bigoplus_{k=1}^{d} \mathbf{f}_{jk}, \tag{3.5}$$

the global scattering matrix of type j, is the columnwise direct sum of the individual matrices $\mathbf{f}_{jk}$. The rank of these matrices cannot exceed its maximum possible value for *one* of the heterogeneous structures of the isomorphous family, and thus the vectors $\mathbf{X}_j(\mathbf{x})$ will lie in a subspace of bounded dimension whatever the size d of the family.

With this extension of the notation, (3.3) is identical to (2.2) and the derivation follows formally the same course as for a single structure. The global m.g.f. is

$$\begin{aligned}\mathscr{Z}(\mathbf{u}) &= \prod_{j=1}^{c} [M_j(\mathbf{f}_j^T\mathbf{u})]^{N_j} \\ &= \prod_{j=1}^{c} \left[M_j\left(\sum_{k=1}^{d} \mathbf{f}_{jk}^T\mathbf{u}_k \right) \right]^{N_j}\end{aligned} \tag{3.6}$$

where

$$\mathbf{u} = \bigoplus_{k=1}^{d} \mathbf{u}_k$$

is partitioned in the same way as $\mathbf{F}$.

Zero or near-zero eigenvalues in $\nabla^2(\log \mathscr{Z})$ due to the low rank of the scattering-factor matrices and/or to solvent regions can be removed by regularization, as explained in §§ 1.2 and 2.3 respectively, so that

the saddlepoint condition

$$\nabla(\log \mathscr{Z}) = \mathbf{F}^* = \bigoplus_{k=1}^{d} \mathbf{F}_k^* \tag{3.7}$$

may be fulfilled in the allowed subspace, possibly after projecting the recentring data $\mathbf{F}^*$. This leads to updated ME distributions for each type of atom:

$$q_j^{\mathrm{ME}}(\mathbf{x}) = [m_j(\mathbf{x})/Z_j(\boldsymbol{\lambda})]\, e^{\boldsymbol{\lambda}.\mathbf{f}_j\boldsymbol{\xi}(\mathbf{x})}$$

$$= [m_j(\mathbf{x})/Z_j(\boldsymbol{\lambda})] \exp\left[\sum_{k=1}^{d} \boldsymbol{\lambda}_k.\mathbf{f}_{jk}\boldsymbol{\xi}(\mathbf{x})\right] \tag{3.8}$$

with

$$Z_j(\boldsymbol{\lambda}) = M_j(\mathbf{f}_j^T\boldsymbol{\lambda}) = M_j\left(\sum_{k=1}^{d} \mathbf{f}_{jk}^T\boldsymbol{\lambda}_k\right), \tag{3.9}$$

and to the saddlepoint approximation

$$\mathscr{P}^{\mathrm{SP}}(\mathbf{F}^*) = e^{\mathscr{S}}[\det(2\pi\mathbf{Q})]^{-1/2} \tag{3.10}$$

where

$$\mathscr{S} = \log \mathscr{Z} - \boldsymbol{\lambda}.\mathbf{F}^* = \sum_{j=1}^{c} N_j \mathscr{S}_{m_j}(q_j^{\mathrm{ME}}) \tag{3.11}$$

is the total relative entropy, and where

$$\mathbf{Q} = \mathrm{reg}\,[\nabla^2_{\boldsymbol{\lambda}\boldsymbol{\lambda}}(\log \mathscr{Z})] \tag{3.12a}$$

$$= \mathrm{reg}\left\{\sum_{j=1}^{c} N_j[\mathbf{f}_j\nabla^2(\log M_j)\mathbf{f}_j^T]\right\}. \tag{3.12b}$$

Before regularization, this matrix may be partitioned into blocks as

$$\mathbf{Q} = \begin{bmatrix} \mathbf{Q}_{11} & \cdots & \mathbf{Q}_{1d} \\ \vdots & & \vdots \\ \mathbf{Q}_{d1} & \cdots & \mathbf{Q}_{dd} \end{bmatrix}, \tag{3.12c}$$

where each block

$$\mathbf{Q}_{k'k''} = \sum_{j=1}^{c} N_j[\mathbf{f}_{jk'}\nabla^2(\log M_j)\mathbf{f}_{jk''}^T] \tag{3.12d}$$

may be calculated from the trigonometric covariance matrices $\nabla^2(\log M_j)$, given by the structure-factor algebra, and from the scattering-factor matrices $\mathbf{f}_{jk}$.

3.2. *Application to MIR and MWAS methods*

If H consists of a single reflexion, then $\mathbf{F}^*$ contains all the (M)IR/(MW)AS data for that reflexion and $\mathscr{P}(\mathbf{F}^*)$ constrains all the structure-factor values to fulfil the geometric conditions of the Harker construction, the 'lack of closure' for each derivative being attributed to the 'clutter' atoms for that derivative. This is simply the model of Blow & Crick (1959) in which the lack of isomorphism is modelled in real space by a uniform distribution of clutter. The present approach would allow one to accommodate non-uniform ME distributions of clutter, which would be useful as ME residual maps for diagnostic purposes. Most importantly, however, this statistical approach can deal with many reflexions at a time, and thus use statistical relations between phases to resolve the remaining ambiguities in the phase indications produced by the Blow & Crick (1959) method.

If H consists of three reflexions forming a triplet, the joint probability distribution $\mathscr{P}^{\mathrm{SP}}(\mathbf{F}^*)$ will yield as an approximation when $\mathbf{F}^* \simeq \mathbf{0}$ all the probabilistic formulae recently derived by Hauptman and others (Hauptman, 1982; Giacovazzo, 1983*a*; Cascarano & Giacovazzo, 1985; Giacovazzo, 1987; Klop, Krabbendam & Kroon, 1987) and cast in the form of inference rules by Karle (1983, 1984, 1985, 1986), following earlier work by Kroon, Spek & Krabbendam (1977) and Heinerman, Krabbendam, Kroon & Spek (1978). These developments, however, suffer from certain limitations, which are overcome in the present multichannel approach.

(1) They use the Edgeworth series, and hence are inaccurate for large moduli; the present approach uses the saddlepoint method, which is exempt from this limitation.

(2) They use uniform distributions of atoms, which will prevent a proper representation of the progressive localization of heavy atoms and of the effect of solvent in proteins; the present approach can deal with arbitrary non-uniform distributions of atoms, which can accommodate all these phenomena.

(3) They produce *small-base* joint distributions, concerning individual or few phase invariants, which are unable (because N is large) to capture any statistical correlations between the three macromolecular contributions to the triple of structure factors. The sharp indications of mixed-phase invariants found by Fortier, Weeks & Hauptman (1984) follow from the fact that large isomorphous or anomalous differences imply near alignment in the three Harker constructions between heavy-atom and macromolecular structure factors, thus causing the latter indirectly to obey phase relations similar to those obeyed by the former. This interpretation, from which Karle's rules follow easily, was arrived at by Fortier, Fraser & Moore (1986), Fraser (1987) and Fortier (1987) on the basis of extensive numerical evidence that sharp triplet phase indications are found only for a selected subset of instances where a single isomorphous derivative suffices to produce *unimodal* phase probability densities for the three members of the triplet. Thus these recent developments leave untouched the central problem of the (M)IR/(MW)AS method, namely the bimodality of most of its phase indications. This restriction was to be expected, since these ambiguities can only be resolved by means of a criterion for preferring certain combinations of *macromolecular* structure-factor values to others; small-base j.p.d.'s are unable to do this, because N is large. By contrast, the present multichannel approach uses *large-base*

j.p.d.'s, which are capable of capturing strong statistical phase relations even for large structures (MEFDM, § 7.3), and hence has potentially the power to resolve these ambiguities.

(4) These approaches do not allow for the effects of lattice-preserving non-isomorphism, and do not provide for possible errors in the data; whereas here we may use a mixture of positive and negative 'clutter' atoms to model this type of non-isomorphism, while the experimental data will be consulted through likelihood functions which take measurement errors into account.

(5) New formulae need to be derived for different invariants, different space groups, different choices of centric and acentric reflexions, and different assumptions regarding the nature of substituents. In this work, by contrast, a single uniform procedure will yield a 'large-base' j.p.d. through a single formula (3.10) for any basis set of reflexions H, in any space group, for any combination of (M)IR and (MW)AS.

Other attempts have been made to couple isomorphous replacement to entropy maximization (Wilkins & Stuart, 1986; Bryan & Banner, 1987) using the lack-of-closure residual of the Blow & Crick (1959) error model as a constraint function. Similar work was carried out by the present author in the course of the structure determination of an Fab-lysozyme complex (Amit, Mariuzza, Phillips & Poljak, 1986) but serious problems of non-isomorphism were encountered, which emphasized the need for the present reformulation of the entire probabilistic basis of substitution methods from first principles in terms of joint distributions. In paper II, likelihood functions will be built from these joint distributions and applied to the problems of locating heavy atoms and refining their parameters.

Finally, it has become a common practice to *combine* (M)IR/(MW)AS phase information, encoded according to the method of Hendrickson & Lattman (1970), with other sources of phase information such as those given by the tangent formula (Hendrickson, Love & Karle, 1973), non-crystallographic symmetry (Bricogne, 1976) or solvent flatness (Wang, 1985). This encoding and combination of phase information is done independently for each reflexion, and hence cannot represent statistical correlations between phases in a natural way. By contrast, the present approach accommodates, without any loss, all the available sources of phase information, which are *automatically combined* into a single *joint* probability distribution.

3.3. *Application to the contrast-variation method*

In a macromolecular crystal, it is sometimes possible to modify the scattering power of the solvent (*e.g.* by varying the D_2O/H_2O ratio for neutrons, or by changing the nature or concentration of the salt for X-rays), thus causing intensity changes at low angle from which some structural information can be deduced. The method has met with several successes in the low-resolution study of several large biological assemblies (*e.g.* Bentley, Lewit-Bentley, Finch, Podjarny & Roth, 1984) but its use has not been extended to higher resolution, for need of an adequate treatment of the effects of density non-uniformities within the macromolecule. The most recent statistical treatment, due to Roth (1987), is based on Wilson's statistics, and hence does not produce any phase relations between the structure factors belonging to different reflexions.

A full probabilistic treatment of the contrast-variation method is immediately obtained by considering an isomorphous family of heterogeneous structures differing by the mean electron (or neutron scattering-length) density in the solvent region $D-U$. This model does accommodate the phenomenon of contrast matching, as is easily seen by extending the assumptions of § 2.3 to an isomorphous family. In the absence of any phase information (*i.e.* at $\boldsymbol{\lambda}=\mathbf{0}$), the expectation value of $\mathbf{F}$ under $\mathcal{P}(\mathbf{F})$ is given by

$$\nabla(\log \mathcal{L})|_{\boldsymbol{\lambda}=\mathbf{0}} = N_0\mathbf{f}_0\mathcal{F}^{-1}[m_0] + \sum_{j=1}^{c} N_j\mathbf{f}_j\mathcal{F}^{-1}[m_j]$$

where $\mathcal{F}^{-1}[m]$ denotes the vector of Fourier coefficients of function m for $\mathbf{h}\in H$. But because $\chi_{D-U}=1-\chi_U$ we have

$$m_0(\mathbf{x}) = 1/(D-U) - [U/(D-U)]m_j(\mathbf{x}) \quad \text{for all } j=1,\ldots,c,$$

and so, for $\mathbf{h}\neq\mathbf{0}$,

$$\mathcal{F}^{-1}[m_0](\mathbf{h}) = -[U/(D-U)]\mathcal{F}^{-1}[m_j](\mathbf{h}) \quad \text{for all } j=1,\ldots,c,$$

so that

$$\langle\mathbf{F}\rangle = \left[\left(\sum_{j=1}^{c} N_j\mathbf{f}_j\right) - [U/(D-U)]N_0\mathbf{f}_0\right]\mathcal{G}, \quad (3.13)$$

where $\mathcal{G}$ is the columnwise direct sum of d copies of the vector of values of the interference function at points $\mathbf{h}\in H$ of the reciprocal lattice. At low resolution,

$$\sum_{j=1}^{c} N_j f_{jk}(\mathbf{h}) = \rho^M U \quad \text{for all } k=1,\ldots,d, \quad (3.14)$$

where ρ^M is the mean electron density of the macromolecule, while

$$N_0 f_{0k}(\mathbf{h}) = \rho_k^S \quad \text{for each } k=1,\ldots,d, \quad (3.15)$$

where ρ_k^S is the mean solvent electron density for the kth contrast level. Relation (3.13) may therefore be written as

$$\langle\mathbf{F}_k\rangle = U(\rho^M - \rho_k^S)\mathcal{G}, \quad k=1,\ldots,d, \quad (3.16)$$

which is the basic relation of the contrast-variation method.

At higher resolution, the density non-uniformities within the macromolecule invalidate (3.14). It is then possible to use a series of contrast-variation measurements to determine the moduli of the protein and solvent structure factors and the absolute value of the angle between them (Roth, Lewit-Bentley & Bentley, 1984); but no method so far exists to determine the actual absolute phases.

By contrast, the saddlepoint approximation (3.10) to the joint probability distribution of structure factors belonging to a contrast-variation series provides the necessary phase relations to initiate the phasing of a macromolecular structure at low resolution by means of contrast-variation measurements, without being limited to resolutions where the inside of the macromolecule is essentially featureless. Any anomalous-scattering effect from contrast agents such as Cs, Se or lanthanide ions can be handled by this formalism, and will further help define the molecular envelope U. If non-uniform adsorption of some ions occurs at the same time, this can be modelled statistically by representing the solvent itself as a heterogeneous mixture of several species of atoms, each species being assigned a distinct channel.

4. Known structural fragments

It is often the case in the course of a crystal structure determination that the location and orientation of some atoms or molecular fragments become known, and that this knowledge is then to be 'recycled' to assist the overall phase determination, or that one wishes to detect or confirm their presence by means of a statistical test. This type of situation arises in the 'molecular replacement' method (Rossmann, 1972), and also when one is seeking to locate substituents in an isomorphous family of structures. In all cases it is desirable to describe quantitatively the effect of the presence of such fragments on the joint distribution of all observable structure factors. This will be done here by creating a 'deterministic channel' to accommodate the fragment, alongside the channels already assigned to the different types of random atoms.

4.1. *General theory*

Let us consider an isomorphous family of structures defined in § 3. Let $(\mathbf{H}_0)$ be the hypothesis that these structures consist of N_j atoms of type j $(j=1,\ldots,c)$, distributed according to prior $m_j(\mathbf{x})$. Then the m.g.f. for $\mathscr{P}(\mathbf{F})$ is, as in (3.6),

$$\mathscr{Z}(\mathbf{u}) = \prod_{j=1}^{c} [Z_j(\mathbf{u})]^{N_j} \quad \text{with} \quad Z_j(\mathbf{u}) = M_j(\mathbf{f}_j^T\mathbf{u}), \quad (4.1)$$

where M_j is defined in (2.1).

Now let $(\mathbf{H}_1)$ be the hypothesis that the structures contain a fragment consisting of n_j atoms of type j $(j=1,\ldots,c)$ in a fixed (known or parametrized) position or orientation. Then the j.p.d. $\mathscr{P}'(\mathbf{F})$ will differ from $\mathscr{P}(\mathbf{F})$ in two respects:

(1) the known fragment contributes a vector

$$\mathbf{F}^{\text{calc}} = \bigoplus_{k=1}^{d} \mathbf{F}_k^{\text{calc}}$$

to $\mathbf{F}$ which is *no longer random* but 'deterministic', so that it contributes a factor $\exp[\mathbf{u}.\mathbf{F}^{\text{calc}}]$ to the m.g.f. $\mathscr{Z}'$ of $\mathscr{P}'$;

(2) the random part of the model is itself modified: only $N_j - n_j$ atoms of each species j remain randomly distributed, and their prior distribution may no longer be $m_j(\mathbf{x})$ but becomes $m_j'(\mathbf{x})$ because the 'deterministic' atoms in the fragment will exclude the random atoms from the region they occupy.

Thus the m.g.f. $\mathscr{Z}'$ of $\mathscr{P}'$ under hypothesis $(\mathbf{H}_1)$ is

$$\mathscr{Z}'(\mathbf{u}) = \exp[\mathbf{u}.\mathbf{F}^{\text{calc}}] \prod_{j=1}^{c} [Z_j'(\mathbf{u})]^{N_j - n_j} \quad (4.2a)$$

with

$$Z_j'(\mathbf{u}) = M_j'(\mathbf{f}_j^T\mathbf{u}) \quad (4.2b)$$

where

$$M_j'(\mathbf{t}) = \int_D m_j'(\mathbf{x})\, e^{\mathbf{t}.\boldsymbol{\xi}(\mathbf{x})}\, \mathrm{d}^3\mathbf{x} \quad (4.2c)$$

is the m.g.f. corresponding to the exclusion-modified prior distribution $m_j'(\mathbf{x})$ for each $j=1,\ldots,c$.

The difference between $\mathscr{P}$ and $\mathscr{P}'$ may be illustrated as follows. Let the known fragment exclude random atoms from a region U of the asymmetric unit. Then, in the absence of phase information (*i.e.* at $\boldsymbol{\lambda}=\mathbf{0}$), the difference between the vectors of first moments $\langle\mathbf{F}\rangle$ and $\langle\mathbf{F}\rangle'$ under the two hypotheses is

$$\langle\mathbf{F}\rangle' - \langle\mathbf{F}\rangle = \boldsymbol{\nabla}\log(\mathscr{Z}'/\mathscr{Z})$$

$$= \mathbf{F}^{\text{calc}} - \sum_{j=1}^{c} n_j \mathbf{f}_j \mathscr{F}^{-1}[m_j - m_j'], \quad (4.3)$$

while the two covariance matrices differ by

$$\mathbf{Q}' - \mathbf{Q} = \boldsymbol{\nabla}^2 \log(\mathscr{Z}'/\mathscr{Z})$$

$$= -\sum_{j=1}^{c} n_j \mathbf{f}_j \boldsymbol{\nabla}^2 \log(M_j/M_j')\mathbf{f}_j^T. \quad (4.4)$$

In (4.3) the subtrahend is the expectation value of the contribution of the random atoms displaced by the fragment, which will be proportional to the vector $\mathscr{G}$ of values of the interference function (see §§ 2.3 and 3.3) corresponding to the region U occupied by the fragment; while in (4.4) it is the contribution to the covariance matrix emanating from those same displaced random atoms, which by structure-factor algebra (see § 2.3) will also involve sample values of the interference function $\mathscr{G}$. Therefore, the presence

of the fragment biases the first moments in the subspace spanned by the $\mathbf{X}_j(\mathbf{x})$ for $\mathbf{x} \in U$, and reduces the dispersion of $\mathscr{P}(\mathbf{F})$ in that subspace. These two effects will result in a sharper joint distribution, offset from its original centre.

Upon recentring around a given vector $\mathbf{F}^*$, the difference between $\mathscr{P}$ and $\mathscr{P}'$ will be accentuated by a different ME updating of the prior distribution of atoms. Under $(\mathbf{H}_0)$, $m_j(\mathbf{x})$ becomes

$$q_j^{\mathrm{ME}}(\mathbf{x}) = [m_j(\mathbf{x})/Z_j(\boldsymbol{\lambda})]\, e^{\boldsymbol{\lambda}.\mathbf{f}_j\xi(\mathbf{x})}$$

with $\boldsymbol{\lambda}$ determined by

$$\nabla(\log \mathscr{Z}) \equiv \sum_{j=1}^{c} N_j \mathbf{f}_j \nabla_\tau(\log M_j) = \mathbf{F}^*;$$

while under $(\mathbf{H}_1)$, $m_j'(\mathbf{x})$ becomes

$$q_j'^{\mathrm{ME}}(\mathbf{x}) = [m_j'(\mathbf{x})/Z_j'(\boldsymbol{\lambda}')]\, e^{\boldsymbol{\lambda}'.\mathbf{f}_j\xi(\mathbf{x})}$$

with $\boldsymbol{\lambda}'$ determined by

$$\nabla(\log \mathscr{Z}') \equiv \mathbf{F}^{\mathrm{calc}} + \sum_{j=1}^{c} (N_j - n_j)\mathbf{f}_j \nabla(\log M_j') = \mathbf{F}^*.$$

The covariance matrices $\mathbf{Q}$ and $\mathbf{Q}'$ will also be updated differently, and finally the total entropies $\mathscr{S} = \log \mathscr{Z} - \boldsymbol{\lambda}.\mathbf{F}^*$ and $\mathscr{S}' = \log \mathscr{Z}' - \boldsymbol{\lambda}'.\mathbf{F}^*$ will be different.

4.2. *Scope of the statistical model*

As has just been shown the presence (known, or assumed as a hypothesis to be tested) of a molecular fragment will lead to sharper estimates of both joint and conditional distributions of structure factors, because the random component of the structure has had some of its atoms removed and placed into a 'deterministic' fluctuation-free component. It is worth noting that this extra phase information will be *automatically combined*, in the saddlepoint estimates $\mathscr{P}$ and $\mathscr{P}'$, with any other phase information which could be extracted by the MIR, MWAS or contrast-variation (CV) methods from the available data.

Likelihood functions built from these conditional distributions will make it possible to test $(\mathbf{H}_1)$ against $(\mathbf{H}_0)$ on the basis of the observed moduli $|\mathbf{F}_K|^{\mathrm{obs}}$. Since the description of the fragment considered in $(\mathbf{H}_1)$ may contain parameters describing (say) its position, orientation and occupancy, testing $(\mathbf{H}_1)$ against $(\mathbf{H}_0)$ will provide a way of estimating these parameters. For instance, it will be possible to perform rotation and translation searches for fragments by maximum-likelihood methods rather than by Patterson superposition; or to assess the statistical significance of certain bound water molecules at the final stages of a macromolecular structure refinement by likelihood criteria rather than by tests involving R factors. The advantage of these new approaches is that, as shown by Neyman & Pearson (1933*a*, *b*), likelihood criteria are more powerful than any others. This task will be pursued in paper II.

4.3. *Comparison with previous methods*

Numerous methods exist for recycling reliably identified molecular fragments into the structure determination process, and this comparison will be limited to those connected with direct methods.

The first study of the statistical effects of a known fragment (or 'heavy atom') on the probability distribution of structure factors is that of Sim (1959). His treatment, however, uses Wilson statistics for each reflexion, and so neglects the *joint* distribution problem altogether.

Beurskens has used the presence of a known fragment to recondition the triplet phase relations holding between the contributions from the rest of the structure to produce the highly successful *DIRDIF* program (Van den Hark, Prick & Beurskens, 1976; Prick, Beurskens & Gould, 1978, 1983). This procedure has been cast in the language of joint distributions by Giacovazzo (1983*b*) and Camalli, Giacovazzo & Spagna (1985). Although there is some controversy as to the precise relation between them [see Beurskens (1987) and Camalli, Giacovazzo & Spagna (1987)], these two treatments both suffer from the usual limitations of the traditional approach to direct methods, mentioned in § 3.2. In this instance their most serious drawback is their inability to deal with the non-uniform distribution of the atoms making up the rest of the structure, which contains a great deal of phasing power (see § 2.3). Some attempts have previously been made by Wilson (1964), Nigam (1972) and Nigam & Wilson (1980) to take into account exclusion effects due to certain symmetry elements, and by Pradhan, Ghosh & Nigam (1985) to study those due to heavy atoms in special positions. However, these investigations have been concerned with intensity statistics rather than with phase relations, and they did not consider exclusion effects due to known fragments.

Recently Marvin, Bryan & Nave (1987) have suggested that the presence of a fragment should be reflected by a modification of the prior distribution of the atoms (which are all randomly distributed) by means of an 'imprint' of the fragment. The analysis (4.3) and (4.4) shows that this is correct for first-order moments, but not for second-order moments: it fails to take into account the fact that the fragment contribution is *deterministic*, but instead treats it as a modulation of the random component of the structure. This leads to overestimation of the dispersion of conditional distributions, and hence would lead to quantitatively incorrect likelihood functions.

5. Non-crystallographic symmetries

The exploitation of non-crystallographic symmetries (Rossmann & Blow, 1963) by real-space symmetry averaging (Bricogne, 1974, 1976) to assist phase determination has been a substantial advance in macromolecular crystallography. It has been used to solve the structures of a large number of proteins [see Wilson, Skehel & Wiley (1981) for a particularly difficult case], and made it possible ten years ago to tackle virus structures (Bloomer, Champness, Bricogne, Staden & Klug, 1978; Harrison, Olson, Schutt, Winkler & Bricogne, 1978). Although the method has also succeeded in providing some degree of phase extension in situations of high symmetry (Nordman, 1980; Gaykema, Volbeda & Hol, 1985; Hogle, Chow & Filman, 1985; Arnold, Vriend, Luo, Griffith, Kamer, Erickson, Johnson & Rossmann, 1987), this process is slow and unstable, and does not allow an *ab initio* phase determination.

Here we will incorporate the real-space treatment of non-crystallographic symmetries given earlier (Bricogne, 1974) into the statistical framework of the ME method, in order to couple the phase-improvement capabilities of the former to the powerful phase-extrapolation properties of the latter.

5.1. *General theory*

Let us assume that r copies of a molecule ($r>1$) are present in the asymmetric unit of a crystal. Let U be the bounded (non-periodic) region occupied by one such molecule in some reference frame and let $\mathscr{G}$ be the associated interference function. Let the r molecules in D be related to the reference molecule by a set Γ of transformations

$$\mathbf{x}\rightarrow\mathbf{T}_p\mathbf{x}=\mathbf{C}_p\mathbf{x}+\mathbf{d}_p, \qquad p=1,\ldots,r, \tag{5.1}$$

under which U becomes $U_p=\mathbf{T}_pU$; the various U_p for different values of p are disjoint, except perhaps for common boundary points. If Γ is a group, then a point $\mathbf{x}$ may have a non-trivial isotropy subgroup $\Gamma_{\mathbf{x}}$, with $|\Gamma_{\mathbf{x}}|$ elements.

The joint distribution $\mathscr{P}(\mathbf{F})$ of structure factors is profoundly altered by these assumptions: the atoms in the different regions U_p, which would otherwise be statistically independent, are now totally correlated. The random-atom model therefore consists of placing atoms randomly in U accordingly to some prior distribution $m(\mathbf{x})$ which vanishes outside U, an atom placed at random position $\mathbf{x}$ now giving rise to a random vector $\boldsymbol{\xi}^{\rm nc}(\mathbf{x})$ of trigonometric structure-factor contributions whose expression involves an expansion by the non-crystallographic (as well as by the crystallographic) symmetry operations.

The basic definitions (0.1) and (0.2) are thus replaced by

$$\alpha_{\mathbf{h}}^{\rm nc}(\mathbf{x})+i\beta_{\mathbf{h}}^{\rm nc}(\mathbf{x})=\Xi^{\rm nc}(\mathbf{h},\mathbf{x}) \tag{5.2a}$$

for $\mathbf{h}$ acentric and

$$\gamma_{\mathbf{h}}^{\rm nc}(\mathbf{x})=\exp(-i\theta_{\mathbf{h}})\Xi^{\rm nc}(\mathbf{h},\mathbf{x}) \tag{5.2b}$$

for $\mathbf{h}$ centric, where

$$\Xi^{\rm nc}(\mathbf{h},\mathbf{x})=|\Gamma_{\mathbf{x}}|^{-1}\sum_{p=1}^{r}\Xi(\mathbf{h},\mathbf{T}_p\mathbf{x}), \tag{5.3}$$

the function $\Xi(\mathbf{h},\mathbf{x})$ being defined by (0.2). Conventions (0.3) regarding the arrangement of the coordinates into the column vector $\boldsymbol{\xi}^{\rm nc}(\mathbf{x})$ are retained. We may then write

$$\boldsymbol{\xi}^{\rm nc}(\mathbf{x})=|\Gamma_{\mathbf{x}}|^{-1}\sum_{p=1}^{r}\boldsymbol{\xi}(\mathbf{T}_p\mathbf{x}), \tag{5.4}$$

and the m.g.f. for the distribution of this random vector is

$$M^{\rm nc}(\mathbf{t})=\int_D m(\mathbf{x})\exp[\mathbf{t}.\boldsymbol{\xi}^{\rm nc}(\mathbf{x})]\,{\rm d}^3\mathbf{x}. \tag{5.5}$$

The covariance structure of this distribution is much richer than that derived from ordinary m.g.f.'s (0.5). Indeed, the off-diagonal elements of its covariance matrix $\nabla^2\log(M^{\rm nc})$ are mixed second-order moments

$$\langle\xi_{\mathbf{h}}^{\rm nc}(\mathbf{x})\xi_{-\mathbf{h}'}^{\rm nc}(\mathbf{x})\rangle;$$

by generalized structure-factor algebra (see Appendix), these contain contributions of the form

$$\mathscr{G}[(\mathbf{R}_g\mathbf{C}_p)^T\mathbf{h}-(\mathbf{R}_{g'}\mathbf{C}_{p'})^T\mathbf{h}'], \qquad p,p'=1,\ldots,r; \quad g,g'\in G,$$

which sample the interference function $\mathscr{G}$ at *non-integral* points of the reciprocal lattice. Whenever $\mathbf{h}$ and $\mathbf{h}'$ are such that their orbits under the combination of local and global symmetries contain a pair of points closer than the spacing between integral lattice points, this contribution will be close to unity, thus inducing very strong correlations between the corresponding structure factors. The counterpart of these strong correlations is that the covariance matrix is singular, having a regular subspace of relative dimension U/D as explained in § 2.3. These algebraic phenomena are in accord with the intuitive idea that symmetry amounts to total correlation, and reduces the number of degrees of freedom.

With these new definitions, the formal structure of all previous developments can be retained (including the introduction of multiple scatterer channels $j=1,\ldots,c$, multiple data channels $k=1,\ldots,d$, and fragments), provided one uses m.g.f.'s like (5.5) for atoms obeying the non-crystallographic symmetry, and ordinary m.g.f.'s like (2.1) for other atoms. The latter category comprises solvent atoms, but may also include a mixture of positive and negative 'clutter' atoms within the regions U_p to represent possible deviations from exact non-crystallographic symmetry.

For instance, the joint distribution $\mathscr{P}(\mathbf{F})$ for an isomorphous family of structures, with r identical molecules in its asymmetric unit, each containing N_j atoms of species j distributed as $m_j(\mathbf{x})$, and with N_0 atoms of solvent outside these subunits, will have a m.g.f. whose form is easily adapted from (2.23) to yield

$$\mathscr{Z}(\mathbf{u}) = [M_0(\mathbf{f}_0^T\mathbf{u})]^{N_0} \prod_{j=1}^{c} [M_j^{nc}(\mathbf{f}_j^T\mathbf{u})]^{N_j}. \quad (5.6)$$

Because of the singularity of the trigonometric covariance matrices $\nabla^2 \log (M_j^{nc})$, a careful regularization will have to be applied by selecting (as in § 2.3) the subspace allowed by the geometric redundancy of the situation, which is again the image space of Crowther's **H** matrix or equivalently of the real-space averaging operator (Bricogne, 1974). After projecting the recentring data $\mathbf{F}^*$ onto this allowed subspace, the saddlepoint condition

$$\nabla (\log \mathscr{Z}) \equiv N_0\mathbf{f}_0\nabla (\log M_0) + \sum_{j=1}^{c} N_j\mathbf{f}_j\nabla(\log M_j^{nc}) = \mathbf{F}^* \quad (5.7)$$

can be fulfilled for a unique $\boldsymbol{\lambda}$ belonging to that subspace, leading to the updated ME distributions

$$q_0^{ME}(\mathbf{x}) = [m_0(\mathbf{x})/Z_0(\boldsymbol{\lambda})] \exp [\boldsymbol{\lambda}.\mathbf{f}_0\boldsymbol{\xi}(\mathbf{x})] \quad (5.8)$$

$$q_j^{ME}(\mathbf{x}) = [m_j(\mathbf{x})/Z_j^{nc}(\boldsymbol{\lambda})] \exp [\boldsymbol{\lambda}.\mathbf{f}_j\boldsymbol{\xi}^{nc}(\mathbf{x})]. \quad (5.9)$$

Since the q_j^{ME} are built from symmetrized expressions $\boldsymbol{\xi}^{nc}(\mathbf{x})$, they automatically obey the local symmetry Γ.

The saddlepoint approximation $\mathscr{P}^{SP}(\mathbf{F}^*)$ has the same formal expression (3.10)–(3.11) as for an isomorphous family, except that the covariance matrix **Q** involves the Hessians of symmetrized m.g.f.'s M_j^{nc} for the macromolecule atoms:

$$\mathbf{Q} = \text{reg} \left\{ N_0\mathbf{f}_0\nabla^2 (\log M_0)\mathbf{f}_0^T + \sum_{j=1}^{c} N_j[\mathbf{f}_j\nabla^2 (\log M_j^{nc})\mathbf{f}_j^T] \right\}. \quad (5.10)$$

It is a straightforward exercise in notation to extend this treatment to the case where there are several types of molecules, each with its own local symmetry, in the asymmetric unit of the crystal.

5.2. *Scope of the statistical theory of non-crystallographic symmetry*

The above treatment of the impact of local symmetries on the phase-determination process is a powerful extension of the existing theory. Indeed, the latter only specifies that any vector of structure factors **F** should obey the condition $\mathbf{F} = \mathbf{HF}$ (Crowther, 1967), or equivalently that $\boldsymbol{\rho} = \mathbf{B}\boldsymbol{\rho}$ [*i.e.* that the electron density $\boldsymbol{\rho}$ should be invariant by real-space averaging and solvent flattening (Bricogne, 1974)], but no further indications are given as to *where* in the allowed eigenspace of **H** or **B** the actual solution is more likely to lie. By contrast, the saddlepoint approximation $\mathscr{P}^{SP}$ to the joint distributions of structure factors was designed to provide these very indications: it can attribute different probabilities to two vectors $\mathbf{F}^*$ having the same moduli and both lying in the allowed subspace ($\mathscr{H}_1$).

The current methodology (Bricogne, 1976) uses a phase combination procedure based on Sim's formula to merge the phase information generated by real-space averaging with initial MIR information, thus treating the averaged solvent-flattened map as the map of a 'known fragment'. As pointed out in § 4.3, this uses Wilson statistics for each reflexion, thus ignoring all phase correlations across reflexions. By contrast, the present approach directly leads to a joint distribution $\mathscr{P}^{SP}(\mathbf{F})$ where **F** contains data for a whole isomorphous family; therefore $\mathscr{P}^{SP}$ automatically combines *all* phase information available from isomorphous derivatives, solvent flatness and local symmetry into a *joint* distribution which also represents faithfully the probabilistic relations between different phases.

As far as phase extension is concerned, the ME updating formulae (5.8) and (5.9) show that the ME-extrapolated values $\mathbf{F}_K^{ME}$ of any set K of non-basis reflexions will obey the local symmetry and solvent-flatness constraints exactly, while the structure (5.10) of the covariance matrix **Q** shows that the build up of further detail by the terms in $\mathbf{F}_k$ will be encouraged if it falls within the molecular envelopes U_p and obeys the correct local symmetry, but discouraged otherwise. Conditional distributions $\mathscr{P}^{SP}(\mathbf{F}_K|\mathbf{F}_H = \mathbf{F}_H^*)$ will therefore have great power of phase extension under non-crystallographic symmetry constraints, enforcing the latter in advance of rather than after the fact as in the iterative procedures used by Nordman (1980), Gaykema *et al.* (1985) and Arnold *et al.* (1987). As a result, likelihood functions derived from these conditional distributions will afford stringent tests of the validity of the ingredients of the statistical model, such as its geometry Γ or the initial phase assumptions $\mathbf{F}_H^*$. These tests may be regarded as hybrids between the use of Patterson superposition methods, such as the rotation function of Rossmann & Blow (1962), and that of intensity statistics in the 'hypercentric' case (Lipson & Woolfson, 1952; Rogers & Wilson, 1953; Wilson, 1952, 1987). Likelihood optimization with respect to parameters describing the geometry Γ, the choice of the region U, of local symmetries will yield new rotation/translation search procedures and statistical tests for the detection and identification of non-crystallographic symmetry; while its optimization with respect to the phases in $\mathbf{F}_H^*$ will provide a method of phase refinement once the geometry has

been characterized, which may make it possible to solve structures having non-crystallographic symmetry *ab initio*. These developments will be presented in paper II.

5.3. *A remark on the statistical interpolation of structure factors*

In their first investigation of non-crystallographic symmetry and solvent flatness, Rossmann & Blow (1963) referred to a short paper by Sayre (1952) which led to the attribution of the phase information latent in such situations to the possibility of somehow *interpolating* structure factors from integral to non-integral points of the reciprocal lattice. This interpretation was made plain when it was shown (Bricogne, 1974) that Crowther's **H** matrix, in terms of which these phase relations could be cast, was indeed closely related to Shannon's (1949) interpolation operator. From a practical point of view, it turned out to be computationally more efficient to operate in direct space by density masking and averaging (Bricogne, 1974, 1976, 1982*a*), but the role played by structure-factor interpolation retains its theoretical importance.

The statistical model presented above does incorporate this interpolation phenomenon, albeit in the less-familiar form of a *probabilistic interpolation* whose result is not a set of numbers but a joint probability distribution for these numbers (Wiener, 1949). In the usual formalism, interpolation in structure-factor space is a kind of 'dispersion relation' which reflects the localization within a *bounded* (*i.e.* non-periodic) envelope U of all the unique densities from which the crystal is built; given the complete set of structure factors $\mathbf{F}_H^*$, where H is now the complete set of unique integral points of the reciprocal lattice, Shannon's interpolation formula allows one to calculate any set $\mathbf{F}_K$ of values of the molecular transform at non-integral points. In the present formalism, knowledge of the molecular envelope U allows one to compute, by generalized structure-factor algebra, all the moments necessary for the calculation of the joint distribution $\mathcal{P}^{\rm SP}(\mathbf{F}_H, \mathbf{F}_K)$, even if $\mathbf{F}_K$ refers to non-integral Miller indices; the result of the probabilistic interpolation from $\mathbf{F}_H^*$ to $\mathbf{F}_K$ is then embodied in the conditional distribution $\mathcal{P}^{\rm SP}(\mathbf{F}_K|\mathbf{F}_H = \mathbf{F}_H^*)$ which constitutes the result of the probabilistic interpolation in question. As the amount of data in $\mathbf{F}_H^*$ increases, this conditional distribution becomes very sharp, its limiting form being a (deterministic) Shannon interpolation.

It is remarkably fortunate that the same device which allows one to obtain better approximations of joint distributions (namely the use of ME non-uniform prior distributions of atoms and of the saddlepoint method) should also shed light on the origin of non-crystallographic symmetry phase relations, which are of an exact rather than probabilistic nature.

6. Multiple crystal forms

The rationalization of the phasing power of geometric redundancies in terms of molecular-transform interpolation makes it obvious that the availability of several crystal forms of a given molecule is essentially the same phenomenon as the presence of non-crystallographic symmetry in one crystal form. The relation between phases thus implied was examined by Main & Rossmann (1966), then cast in matrix form by Crowther (1969). A real-space presentation of the general situation of ν crystal forms, each possibly having its own non-crystallographic symmetry and/or solvent boundary constraints, was given in an earlier paper (Bricogne, 1974) and its implementation has recently helped solve a difficult protein structure of high biological significance (Bjorkman, Saper, Samraoui, Bennet, Strominger & Wiley, 1987). This latter formulation will now be incorporated into the statistical framework of the ME method. As each crystal form may here comprise an isomorphous family, and as known fragments are also accommodated, this final level of generalization may be viewed as an attempt at a 'grand unification' of all phase-determination procedures.

6.1. *General theory*

Let crystal form ℓ ($\ell = 1, \ldots, \nu$) be characterized by its lattice $\mathcal{R}^{(\ell)}$, defined by the coordinates in some common reference frame of its basis vectors $\mathbf{a}^{(\ell)}, \mathbf{b}^{(\ell)}, \mathbf{c}^{(\ell)}$; by its space group $G^{(\ell)}$ of symmetry operations $(\mathbf{R}_{g\ell}^{(\ell)}, \mathbf{t}_{g\ell}^{(\ell)})$; and possibly by a set $\Gamma^{(\ell)} = \{\mathbf{T}_{p\ell}^{(\ell)} | p_\ell = 1, \ldots, r_\ell\}$ of local non-crystallographic symmetry operations $(\mathbf{C}_{p\ell}^{(\ell)}, \mathbf{d}_{p\ell}^{(\ell)})$. Let $\mathbf{x}^{(\ell)}$ denote the real-space position vector

$$x_1^{(\ell)}\mathbf{a}^{(\ell)} + x_2^{(\ell)}\mathbf{b}^{(\ell)} + x_3^{(\ell)}\mathbf{c}^{(\ell)};$$

let $\mathbf{h}^{(\ell)}$ denote the reciprocal-space lattice vector

$$h_1^{(\ell)}\mathbf{a}^{*(\ell)} + h_2^{(\ell)}\mathbf{b}^{*(\ell)} + h_3^{(\ell)}\mathbf{c}^{*(\ell)};$$

and let $H^{(\ell)}$ be a set of unique non-origin reflexions for form ℓ; H will then denote the set-theoretic union of all the $H^{(\ell)}$ for $\ell = 1, \ldots, \nu$. Finally let $\boldsymbol{\xi}^{(\ell)}(\mathbf{x}^{(\ell)})$ and $\boldsymbol{\xi}^{(\ell)\rm nc}(\mathbf{x}^{(\ell)})$ respectively denote the vectors of trigonometric structure-factor expressions in the absence (§ 0.3) and the presence (§ 5.2) of non-crystallographic symmetries $\Gamma^{(\ell)}$ for reflexions $\mathbf{h}^{(\ell)} \in H^{(\ell)}$.

Let the unknown macromolecule be described in some other reference frame, in known or parametrized relation to the first, as a heterogeneous assembly of N_j atoms of species j ($j = 1, \ldots, c$) with prior distribution $m_j(\mathbf{x})$ confined to a bounded region U. An atom placed at random position $\mathbf{x} \in U$ will induce the placement of an atom at

$$\mathbf{x}^{(\ell)} = \mathbf{L}^{(\ell)}\mathbf{x} = \boldsymbol{\Delta}^{(\ell)}\mathbf{x} + \boldsymbol{\delta}^{(\ell)} \qquad (6.1)$$

in the basic subunit of each crystal form ℓ, where the 'clutching' transformations $\mathbf{L}^{(\ell)}$ can be calculated

from the global geometry of all the lattices and frames; these atoms will in turn be expanded by the local and crystallographic symmetries of form ℓ. The vector $\hat{\boldsymbol{\xi}}(\mathbf{x})$ of 'clutched' trigonometric structure-factor expressions for the ν crystal forms simultaneously is therefore

$$\hat{\boldsymbol{\xi}}(\mathbf{x}) = \bigoplus_{\ell=1}^{\nu} \boldsymbol{\xi}^{(\ell)\mathrm{nc}}(\mathbf{L}^{(\ell)}\mathbf{x}) \tag{6.2}$$

where $\bigoplus$ denotes a columnwise direct sum, and where $\mathbf{x}$ is the *same* for all ℓ (hence the term 'clutched').

The m.g.f. for its distribution $\hat{p}$ is

$$\hat{M}(\mathbf{t}) = \int_D m(\mathbf{x}) \exp[\hat{\mathbf{t}}.\hat{\boldsymbol{\xi}}(\mathbf{x})]\, \mathrm{d}^3\mathbf{x}. \tag{6.3}$$

Because of the clutching relations (6.1), the covariance structure of this distribution is very rich. By using generalized structure-factor algebra *across crystal forms*, it is easily seen that mixed second-order moments in the covariance matrix $\nabla^2 \log(\hat{M})$ contain contributions of the form

$$\mathscr{G}[(\mathbf{R}_{g\ell}^{(\ell)}\mathbf{C}_{p\ell}^{(\ell)}\boldsymbol{\Delta}^{(\ell)})^T\mathbf{h}^{(\ell)} - (\mathbf{R}_{g\ell'}^{(\ell')}\mathbf{C}_{p\ell'}^{(\ell')}\boldsymbol{\Delta}^{(\ell')})^T\mathbf{h}^{(\ell')}]$$

which sample the interference function $\mathscr{G}$ at generic (*i.e.* non-integral) points of reciprocal space since the different reciprocal lattices $\mathscr{R}^{*(\ell)}$ are non-congruent. The situation is thus the same as that created by non-crystallographic symmetry with a single form: the covariance matrix is singular, its regular subspace being the image subspace ($\mathscr{H}_1$) of the multiple crystal form $\mathbf{H}$ matrix or of the equivalent averaging operator.

We may now introduce the scattering-factor matrices $\mathbf{f}_{jk_\ell}^{(\ell)}$ describing the scattering power with which atoms of species j ($j = 1, \ldots, c$) contribute to member k_ℓ ($k_\ell = 1, \ldots, d_\ell$) of the isomorphous family attached to crystal form ℓ. These matrices may be aggregated by columnwise direct sum into the global matrices describing the interconnections between all the channels labelled by j, k, ℓ:

$$\hat{\mathbf{f}}_j = \bigoplus_{\ell=1}^{\nu} \left(\bigoplus_{k_\ell=1}^{d_\ell} \mathbf{f}_{jk_\ell}^{(\ell)} \right). \tag{6.4}$$

The contribution $\hat{\mathbf{X}}_j(\mathbf{x})$ of a macromolecule atom of species j placed at $\mathbf{x} \in U$ to the global structure-factor vector

$$\hat{\mathbf{F}} = \bigoplus_{\ell=1}^{\nu} \mathbf{F}^{(\ell)} = \bigoplus_{\ell=1}^{\nu} \left(\bigoplus_{k_\ell=1}^{d_\ell} \mathbf{F}_{k_\ell}^{(\ell)} \right)$$

is then

$$\hat{\mathbf{X}}_j(\mathbf{x}) = \hat{\mathbf{f}}_j \hat{\boldsymbol{\xi}}(\mathbf{x}), \tag{6.5}$$

and the m.g.f. of its distribution is

$$\hat{Z}_j(\hat{\mathbf{u}}) = \hat{M}_j[(\hat{\mathbf{f}}_j)^T\hat{\mathbf{u}}] \tag{6.6}$$

where $\hat{\mathbf{u}}$ is segmented in the same way as $\hat{\mathbf{F}}$:

$$\hat{\mathbf{u}} = \bigoplus_{\ell=1}^{\nu} \mathbf{u}^{(\ell)} = \bigoplus_{\ell=1}^{\nu} \left(\bigoplus_{k_\ell=1}^{d_\ell} \mathbf{u}_{k_\ell}^{(\ell)} \right).$$

Therefore, the m.g.f. for the distribution of the macromolecular component of $\hat{\mathbf{F}}$ is

$$\mathscr{Z}_{\mathrm{mac}}(\hat{\mathbf{u}}) = \prod_{j=1}^{c} [\hat{Z}_j(\hat{\mathbf{u}})]^{N_j}. \tag{6.7}$$

If now each crystal form ℓ contains $N_0^{(\ell)}$ atoms of solvent, with prior distribution $m_0^{(\ell)}(\mathbf{x}^{(\ell)})$ and scattering-factor matrix

$$\mathbf{f}_0^{(\ell)} = \left(\bigoplus_{k_\ell=1}^{d_\ell} \mathbf{f}_{0k_\ell}^{(\ell)} \right),$$

the m.g.f. of its solvent contribution to $\mathbf{F}^{(\ell)}$ is

$$\mathscr{Z}_{\mathrm{solv}}^{(\ell)}(\mathbf{u}^{(\ell)}) = [Z_0^{(\ell)}(\mathbf{u}^{(\ell)})]^{N_0^{(\ell)}} \tag{6.8}$$

where $Z_0^{(\ell)}$ is calculated in the usual way (2.3) from $m_0^{(\ell)}$.

Finally, the m.g.f. for the joint distribution $\mathscr{P}(\hat{\mathbf{F}})$ is

$$\hat{\mathscr{Z}}(\hat{\mathbf{u}}) = \mathscr{Z}_{\mathrm{mac}}(\hat{\mathbf{u}})[\bigotimes_{\ell=1}^{\nu} \mathscr{Z}_{\mathrm{solv}}^{(\ell)}(\mathbf{u}^{(\ell)})] \tag{6.9}$$

where $\bigotimes$ denotes the tensor product of functions of the distinct segments $\mathbf{u}^{(\ell)}$ of the full vector $\hat{\mathbf{u}}$ of carrying variables. It would be a simple matter, involving extra notation but no new principles, to incorporate into this model the presence of known fragments (common or not to all the forms, and obeying or not their non-crystallographic symmetry) or to have several different kinds of subunits shared in various ways by the different crystal forms.

We may apply the regularization procedures of §§ 1.2, 2.3 and 5.1 to the Hessian matrix $\nabla^2(\log \hat{\mathscr{Z}})$ and project accordingly the recentring data $\hat{\mathbf{F}}^*$ to ensure that the saddlepoint condition

$$\nabla(\log \hat{\mathscr{Z}}) = \hat{\mathbf{F}}^* \tag{6.10}$$

can be fulfilled for a unique $\hat{\boldsymbol{\lambda}}$ (segmented as $\hat{\mathbf{u}}$ above) in the regular subspace. The updated ME distributions of the various types of atoms will then be

$$q_0^{(\ell)\mathrm{ME}}(\mathbf{x}^{(\ell)}) = [m_0^{(\ell)}(\mathbf{x}^{(\ell)})/Z_0^{(\ell)}(\boldsymbol{\lambda})] \times \exp[\boldsymbol{\lambda}^{(\ell)}.\mathbf{f}_0^{(\ell)}\boldsymbol{\xi}^{(\ell)}(\mathbf{x}^{(\ell)})] \tag{6.11}$$

for the solvent atoms in crystal form ℓ, and

$$q_j^{\mathrm{ME}}(\mathbf{x}) = [m_j(\mathbf{x})/Z_j(\hat{\boldsymbol{\lambda}})] \exp[\hat{\boldsymbol{\lambda}}.\hat{\mathbf{f}}_j\hat{\boldsymbol{\xi}}(\mathbf{x})] \quad (j = 1, \ldots, c) \tag{6.12}$$

for the atoms of the macromolecule common to all the crystal forms.

The saddlepoint approximation of $\mathscr{P}$ at $\hat{\mathbf{F}}^*$ is then

$$\mathscr{P}^{\mathrm{SP}}(\hat{\mathbf{F}}^*) = \exp(\hat{\mathscr{S}})[\det(2\pi\hat{\mathbf{Q}})]^{-1/2} \tag{6.13}$$

where

$$\hat{\mathscr{S}} = \log \hat{\mathscr{Z}} - \hat{\boldsymbol{\lambda}}.\hat{\mathbf{F}}^* = \sum_{j=1}^{c} N_j \mathscr{S}_{m_j}(q_j^{\mathrm{ME}}) + \sum_{\ell=1}^{\nu} N_0^{(\ell)} \mathscr{S}_{m_0^{(\ell)}}(q_0^{(\ell)\mathrm{ME}}) \quad (6.14)$$

is the weighted sum of the relative entropies associated respectively with the macromolecule atoms and with the solvent atoms of the various crystal forms; and where

$$\mathbf{Q} = \mathrm{reg}\,[\nabla^2_{\hat{\boldsymbol{\lambda}}\hat{\boldsymbol{\lambda}}}(\log \hat{\mathscr{Z}})] \quad (6.15)$$

is constructed in the usual way [by a straightforward adaptation of (3.12)] from the scattering-factor matrices and from the trigonometric covariance matrices $\nabla^2(\log \hat{M})$ and $\nabla^2(\log M_0^{(\ell)})$ for $\ell = 1, \ldots, \nu$ whose elements are calculable by generalized structure-factor algebra.

6.2. *Scope of the statistical theory of multiple crystal forms*

The same remarks apply to the above treatment as those made in § 5.2. The previous non-statistical theory only provided an invariance condition $\mathbf{F} = \mathbf{HF}$ or $\boldsymbol{\rho} = \mathbf{B}\boldsymbol{\rho}$; whereas we now have the means of assigning different probabilities to different sets of structure factors obeying this condition. Phase combination occurs automatically in the construction of $\mathscr{P}^{\mathrm{SP}}$, and its result is presented as a *joint* distribution; whereas the current procedure works on each reflexion independently, and so ignores phase correlations. Finally, the equivalence of the various copies of the unknown molecule(s) is automatically preserved in the ME updating process, and thus in the extrapolated structure factors, so that conditional distributions $\mathscr{P}^{\mathrm{SP}}(\hat{\mathbf{F}}_K | \hat{\mathbf{F}}_H = \hat{\mathbf{F}}_H^*)$ enforce it *in advance* during phase extension. Therefore, the corresponding likelihood functions will again allow the identification of the geometrical relations between the crystal forms, and the refinement of the phase values in $\hat{\mathbf{F}}_H^*$ once the geometry has been characterized. This will be done in paper II.

6.3. *Treatment of non-isomorphism due to lattice distortions*

The lack of isomorphism between heavy-atom derivatives and native crystals of macromolecules is a ubiquitous problem in the use of the MIR method. If the crystal lattice is left undistorted by the substitution, it was shown in § 3 that an equal mixture of positive and negative 'clutter' atoms could be used to represent the statistical effects of local distortions of the native structure, which are usually dealt with *via* the Blow & Crick (1959) lack-of-isomorphism parameter. When heavy-atom substitution distorts the crystal lattice, however, no treatment has so far been available.

Clearly, what has been missing so far is the 'statistical interpolation' device (§ 5.3) which is the basis for the results presented in §§ 5 and 6. In fact, it suffices to treat the native structure and its lattice-distorted derivatives as different crystal forms (representing whatever substitution has occurred by means of the scattering-factor matrices $\mathbf{f}_{jk\ell}^{(\ell)}$) to obtain a *completely general treatment of non-isomorphous derivatives.* The knowledge of the molecular boundary is a necessary prerequisite to the use of this method, since its continuous transform (the interference function) is the active agent in the statistical interpolation of structure factors between non-isomorphous reciprocal lattices.

Since the preparation of heavy-atom derivatives of protein crystals, or the changes of mother-liquor composition used for the purpose of solvent-contrast variation, frequently cause unwanted changes of unit-cell parameters, this novel possibility of dealing with lattice distortions should be able to rescue many difficult macromolecular structure determinations.

7. Summary

This section will summarize briefly some of the more important aspects of the derivations presented above.

7.1. *Effective computability*

All the mathematical entities handled in this work are effectively computable. In particular, the universal procedure for constructing the saddlepoint approximation $\mathscr{P}^{\mathrm{SP}}(\mathbf{F}^*)$ runs as follows:

(1) update the initial distributions of all classes of random atoms so as to fulfil the multichannel saddlepoint (or maximum-entropy) condition;

(2) compute the total relative entropy $\mathscr{S}$ as the N-weighted sum of the entropies of these updated distributions relative to the initial ones;

(3) compute the trigonometric covariance matrices from the Fourier coefficients of these ME distributions by generalized structure-factor algebra;

(4) compute the global covariance matrix from the trigonometric matrices and the scattering-factor matrices, and regularize it to get $\mathbf{Q}$;

(5) compute

$$\mathscr{P}^{\mathrm{SP}}(\mathbf{F}^*) = e^{\mathscr{S}}[\det(2\pi\mathbf{Q})]^{-1/2}.$$

All these operations can be carried out numerically on a large scale by existing methods.

7.2. *Universality, basis-set size, and non-uniformity*

The procedure given is applicable to any collection of reflexions, in any space group(s), in any situation encountered in practice. It is no longer necessary to derive large numbers of lengthy algebraic formulae

for small sets of reflexions under narrowly defined assumptions, then to write as many computer programs to implement them: numerical results can be obtained directly through a unique computational scheme.

Large numbers of reflexions may be considered simultaneously: up to a few hundred with ordinary matrix-inversion methods, and several tens of thousands if the inversion of the covariance matrix and the evaluation of its determinant are carried out by Fourier-transform methods. Such 'large-base' j.p.d.'s can capture strong phase relations even for macromolecules, as was demonstrated in MEFDM, § 7.3.

Non-uniform distributions of atoms play an essential role in this approach. The possibility of handling them correctly gives access to sources of phase relations for which no statistical theory had yet been formulated: solvent flattening, excluded volume effects of fragments, non-crystallographic symmetry and multiple crystal forms. In the latter two cases, non-uniform prior distributions of atoms *must* be handled at the outset, even in the absence of explicit phase information, since it is necessary to break the crystal periodicity in order to *localize* the basic subunit to which a particular set of local symmetries is to be applied.

Problems of rational dependence may be dealt with in a variety of ways, depending on their precise nature. Heavy atoms in special positions may be treated as known fragments; heavy atoms in general but rational positions may be first detected then represented through a separate channel; exact geometric correlations between atoms may be treated as non-crystallographic symmetries; while stacking effects may be modelled through non-uniform prior distributions. Only the case of probabilistic correlations between atoms eludes the present treatment, because it violates the assumption of the statistical independence of the atoms.

7.3. *Multichannel structure*

The great diversity of known sources of phase information has been accommodated by creating a hierarchical structure of 'channels': for different atom types, for different molecular fragments, for different members of an isomorphous family, and for different crystal forms. The interconnections between the random atom channels and the various data channels are specified through scattering-factor matrices, which play a role similar to that of the 'design matrix' in multivariate statistical analysis. This multichannel approach lends itself well to structured programming.

7.4. *Normalization*

Well known difficulties in the theory of normalization for heterogeneous structures have been examined from a new standpoint. It has been shown that no satisfactory answer can be hoped for within the classical theory: the solution resides in the use of the multichannel approach, where each atom type is assigned its own non-uniform distribution, with unnormalized data.

7.5. *Phase combination*

The scheme of Hendrickson & Lattman (1970) for combining various forms of phase information uses *independent* phase probability densities for each reflexion and thus cannot accommodate statistical relations between these phases. Here, all available sources of phase information are incorporated, at a consistent level of approximation, into a *joint* distribution of all structure factors, which can accommodate these relations without any loss of information. The detailed formal structure of this distribution will be examined in paper II.

7.6. *Determinantal inequalities*

The possibility of preserving the strict positive definiteness of the Hessian matrix $\nabla^2(\log \mathscr{Z})$ by suitable regularization has repeatedly pointed out the existence of families of determinantal inequalities, even in circumstances (*e.g.* complex scattering factors) where none would have been expected to hold. They follow from the positivity of the various prior distributions of random atoms, and not from the positivity of the electron (or scattering-length) density.

7.7. *Relation to maximum-entropy theory*

The present approach views entropy maximization as an *accessory calculation* to the use of the saddlepoint method for constructing joint and conditional distributions of structure factors and likelihood functions, which are the basic devices of Bayesian inference methods. It treats the updated distributions q^{ME} as versatile computational intermediates rather than as 'preferred maps' in the terminology of Livesey & Skilling (1985). The orthodox ME viewpoint, as applied to the treatment of heterogeneous structures (Gull, Livesey & Sivia, 1987) and of partial structures (Marvin, Bryan & Nave, 1987), has been shown here to fall short of giving quantitatively correct results, indicating that the ME theory on its own may not always have the unambiguous heuristic power which has been claimed on its behalf.

The study of vanishing principal curvatures in the Hessian matrix $\nabla^2(\log \mathscr{Z})$, and their removal by regularization in order to recover the uniqueness of the saddlepoint, has shed light on the almost magical robustness properties of the ME method stated by Jaynes (1968), showing them to be based on the implicit use of generalized inverses of matrices.

Finally, the ME method only 'looks back' at the cost of accommodating the explicit trial phase assumptions contained in $\mathbf{F}^*$; while the likelihood functions we will construct from $\mathscr{P}^{SP}$ are able to 'look ahead' in a global way at the information contained in the yet unphased moduli, thus providing the basic feedback mechanism of Bayesian methodology. Since the saddlepoint approximation $\mathscr{P}^{SP}$ involves not only the Shannon entropy $\mathscr{S}$, but also a normalization term $[\det(2\pi\mathbf{Q})]^{-1/2}$ (related to the Burg entropy) which plays an important role in likelihood calculations, it should be clear that the present Bayesian approach goes well beyond a naive reliance on entropy maximization.

Concluding remarks

The goal defined at the outset has been reached: a universal procedure has been described which allows the explicit numerical computation of the saddlepoint approximation to the joint probability distribution of an arbitrary collection of structure-factor values. This has been accomplished in a setting of sufficient generality to encompass all currently used sources of phase information.

The resulting expression for the joint distribution has been shown to extend and to improve upon all existing statistical approaches to the phase problem, and to provide an optimal procedure for combining all sources of phase information available in any given instance.

This merging of all phasing methods within a unique formalism offers the guarantee that the forthcoming derivations of conditional distributions and likelihood functions from the universal j.p.d. expression, to be presented in paper II, will automatically apply to all combinations of phasing techniques without requiring special developments for each of them.

I am indebted to Trinity College, Cambridge, England for a Visiting Fellowship which made possible the writing of this paper.

APPENDIX

Generalized structure-factor algebra for the calculation of trigonometric moments

Given the distribution $m(\mathbf{x})$ of random atomic positions, it is possible to calculate all moments such as (0.6) from the Fourier coefficients of $m(\mathbf{x})$ thanks to the fact that the functions $\Xi(\mathbf{h}, \mathbf{x})$ defined by (0.2) form an *algebra, i.e.* that products of such functions may be expressed as linear combinations of functions of the same type. This 'structure-factor algebra' was first investigated by Bertaut (1955*b*, 1956, 1959) and Bertaut & Waser (1957) and has recently been re-examined by Giacovazzo (1980). Shmueli & Kaldor (1981, 1983) computed even moments of the right-hand side of (0.2) with $|G_{\mathbf{x}}| = 1$. All these authors compiled tables of linearization coefficients for the various classes of reflexions in different families of space groups, but this classification carries implicitly the assumption that $m(\mathbf{x})$ is *uniform*, so that the only non-vanishing moments are those for which a null triplet of Miller indices is generated during linearization.

In the present work, $m(\mathbf{x})$ is non-uniform, so that *non-vanishing moments become the rule rather than the exception.* However, the original calculation of Bertaut may still be used to express the general moment in terms of the Fourier coefficients of $m(\mathbf{x})$. This calculation is then generalized so as to deal with non-crystallographic symmetries and multiple crystal forms.

A.1. *Preliminaries*

Using the definitions and conventions of §§ 0.0 and 0.1, we may rewrite (0.3) as

$$\xi_j(\mathbf{x}) = \mathscr{R}e[\exp(-i\omega_j)\Xi(\mathbf{h}_j, \mathbf{x})] \qquad (A.1.1)$$

where $\mathbf{h}_j$ denotes the Miller indices pertaining to the jth component of $\boldsymbol{\xi}$. If $\mathbf{h}(r)$ designates the reflexion numbered r in H, then (*A*.1.1) is equivalent to (0.3) if

$$\mathbf{h}_{2r-1} = \mathbf{h}(r), \qquad \omega_{2r-1} = 0, \qquad r = 1, \ldots, n_a,$$

$$\mathbf{h}_{2r} = \mathbf{h}(r), \qquad \omega_{2r} = \pi/2, \qquad r = 1, \ldots, n_a,$$

$$\mathbf{h}_{2n_a+s} = \mathbf{h}(n_a+s), \quad \omega_{2n_a+s} = \theta[\mathbf{h}(n_a+s)], \qquad s = 1, \ldots, n_c.$$

In other words, the α's and β's in (0.1*a*) may be treated as particular cases of γ's in (0.1*b*). This simplification has the advantage of allowing more flexibility in the choice of coordinates in $\mathbb{R}^n$; for example, it is possible to choose $\omega_{2r-1} = \varphi_{\mathbf{h}(r)}$ and $\omega_{2r} = \varphi_{\mathbf{h}(r)} + \pi/2$ so as to define a radial and an azimuthal direction with respect to some trial phase φ; it can also be used to handle origin-fixing choices in the form of restrictions to linear subspaces of the full structure-factor space.

It will be convenient to introduce the notation

$$e_{\mathbf{h}}(\mathbf{x}) \equiv e(\mathbf{h}, \mathbf{x}) \equiv \exp(2\pi i \mathbf{h}.\mathbf{x}). \qquad (A.1.2)$$

This function has the fundamental properties

$$\text{conjg}(e_{\mathbf{h}}) = e_{-\mathbf{h}} \qquad (A.1.3a)$$

$$e_{\mathbf{h}} e_{\mathbf{k}} = e_{\mathbf{h}+\mathbf{k}} \qquad (A.1.3b)$$

$$e_{\mathbf{h}}(S_g\mathbf{x}) = e_{\mathbf{h}}(\mathbf{t}_g)\, e(\mathbf{R}_g^T\mathbf{h}, \mathbf{x}). \qquad (A.1.3c)$$

The trigonometric structure-factor expression (0.2) may be written

$$\Xi(\mathbf{h}, \mathbf{x}) = |G_{\mathbf{x}}|^{-1} \sum_{g \in G} e(\mathbf{h}, S_g\mathbf{x}) \qquad (A.1.4)$$

and it enjoys the property of being invariant under G:

$$\Xi(\mathbf{h}, S_g\mathbf{x}) = \Xi(\mathbf{h}, \mathbf{x}) \quad \forall g \in G. \qquad (A.1.5)$$

Ξ is the kernel of the discrete inverse Fourier transform $\mathscr{F}_G^{-1}$ for space group G:

$$F_\mathbf{h} = \sum_{\mathbf{x}\in D} \Xi(\mathbf{h}, \mathbf{x})\rho_\mathbf{x} = \mathscr{F}_G^{-1}[\rho](\mathbf{h}) \qquad (A.1.6)$$

where the summation is over a crystallographic grid.

A.2. *Trigonometric moments of non-uniform distributions of atoms*

Let $u, v, w, \ldots$ be integers between 1 and $n = 2n_a + n_c$.

The first-order moments of the distribution p of ξ are easily obtained since by (A.1.1) and (A.1.6)

$$\langle \xi_j \rangle = \mathscr{R}e\{\exp(-i\omega_j)\mathscr{F}_G^{-1}[m](\mathbf{h}_j)\} \qquad (A.2.1)$$

so that $\langle \boldsymbol{\xi} \rangle$ is essentially the vector of Fourier coefficients of $m(\mathbf{x})$ for reflexions $\mathbf{h}$ in H. In traditional treatments of structure-factor algebra $m(\mathbf{x})$ is uniform, and so $\langle \boldsymbol{\xi} \rangle = \mathbf{0}$; but here $\langle \boldsymbol{\xi} \rangle$ may take any value compatible with the positivity of $m(\mathbf{x})$.

The second-order moments require Bertaut's linearization method. Let us first write

$$\begin{aligned}\xi_u(\mathbf{x})\xi_v(\mathbf{x}) &= \mathscr{R}e[\exp(-i\omega_u)\Xi(\mathbf{h}_u, \mathbf{x})] \\ &\quad \times \mathscr{R}e[\exp(-i\omega_v)\Xi(\mathbf{h}_v, \mathbf{x})] \\ &= \tfrac{1}{2}\{\mathscr{R}e[\exp\{-i(\omega_u + \omega_v)\} \\ &\quad \times \Xi(\mathbf{h}_u, \mathbf{x})\Xi(\mathbf{h}_v, \mathbf{x})] \\ &\quad + \mathscr{R}e[\exp\{-i(\omega_u - \omega_v)\} \\ &\quad \times \Xi(\mathbf{h}_u, \mathbf{x})\Xi(-\mathbf{h}_v, \mathbf{x})]\}, \qquad (A.2.2)\end{aligned}$$

so that the calculation of $\langle \xi_u\xi_v \rangle$ is reduced to that of quantities of the form $\langle \Xi_{\mathbf{h}_u}\Xi_{\varepsilon\mathbf{h}_v} \rangle$, where $\varepsilon = \pm 1$. The latter proceeds as follows:

$$\begin{aligned}&\Xi_{\mathbf{h}_u}(\mathbf{x})\Xi_{\varepsilon\mathbf{h}_v}(\mathbf{x}) \\ &= \left[|G_\mathbf{x}|^{-1} \sum_{g_u\in G} e(\mathbf{h}_u, S_{g_u}\mathbf{x})\right] \\ &\quad \times \left[|G_\mathbf{x}|^{-1} \sum_{g_v\in G} e(\varepsilon\mathbf{h}_v, S_{g_v}\mathbf{x})\right] \\ &= |G_\mathbf{x}|^{-2} \sum_{g_u\in G} \left\{e(\mathbf{h}_u, S_{g_u}\mathbf{x}) \sum_{g\in G} e[\varepsilon\mathbf{h}_v, S_g(S_{g_u}\mathbf{x})]\right\} \\ &\qquad\qquad \text{with } g_v = gg_u \\ &= |G_\mathbf{x}|^{-2} \sum_{g_u\in G} \left\{e(\mathbf{h}_u, S_{g_u}\mathbf{x}) \right. \\ &\quad \left. \times \sum_{g\in G} e(\varepsilon\mathbf{h}_v, \mathbf{t}_g)e(\varepsilon\mathbf{R}_g^T\mathbf{h}_v, S_{g_u}\mathbf{x})\right\} \qquad \text{by (A.1.3}c\text{)} \\ &= |G_\mathbf{x}|^{-1} \sum_{g\in G} e(\varepsilon\mathbf{h}_v, \mathbf{t}_g) \\ &\quad \times \left[|G_\mathbf{x}|^{-1} \sum_{g_u\in G} e(\mathbf{h}_u + \varepsilon\mathbf{R}_g^T\mathbf{h}_v, S_{g_u}\mathbf{x})\right] \\ &\qquad\qquad \text{by (A.1.3}b\text{)} \\ &= |G_\mathbf{x}|^{-1} \sum_{g\in G} e(\varepsilon\mathbf{h}_v, \mathbf{t}_g)\Xi(\mathbf{h}_u + \varepsilon\mathbf{R}_g^T\mathbf{h}_v, \mathbf{x}). \qquad (A.2.3)\end{aligned}$$

This is Bertaut's fundamental linearization formula. It immediately yields, by (A.1.6),

$$\begin{aligned}\langle \Xi_{\mathbf{h}_u}\Xi_{\varepsilon\mathbf{h}_v} \rangle &= |G_\mathbf{x}|^{-1} \sum_{g\in G} e(\varepsilon\mathbf{h}_v, \mathbf{t}_g)\mathscr{F}_G^{-1}[m] \\ &\quad \times (\mathbf{h}_u + \varepsilon\mathbf{R}_g^T\mathbf{h}_v, \mathbf{x}) \qquad (A.2.4)\end{aligned}$$

so that this quantity, denoted (u, v, ε_{uv}) below, is calculable from the Fourier coefficients of the prior distribution of atoms $m(\mathbf{x})$. Relation (A.2.2) then yields the moments originally sought:

$$\langle \xi_u\xi_v \rangle = \tfrac{1}{2} \sum_{\varepsilon_{uv}=\pm 1} \mathscr{R}e\{\exp[-i(\omega_u + \varepsilon_{uv}\omega_v)](u, v, \varepsilon_{uv})\}. \qquad (A.2.5)$$

This expression is entirely general and holds for any admixture of centric and acentric reflexions in any space group.

By repeated application of this procedure, products of any number of ξ's may be linearized, and the corresponding moments (0.6) may be evaluated numerically from the Fourier coefficients of $m(\mathbf{x})$. Cumulants (0.7) may be obtained by using their standard expressions in terms of moments (*e.g.* Klug, 1958).

A.3. *Non-crystallographic symmetries*

In the presence of a set Γ of non-crystallographic symmetries, Ξ and ξ become respectively Ξ^{nc} (5.3) and ξ^{nc} (5.4), between which relations (A.1.1) and (A.2.2) contine to hold. The counterpart of (A.1.4) is now

$$\begin{aligned}\Xi^{\text{nc}}(\mathbf{h}, \mathbf{x}) &= |\Gamma_\mathbf{x}|^{-1} \sum_{p=1}^{r} \left\{|G_\mathbf{x}|^{-1} \sum_{g\in G} e[\mathbf{h}, S_g(T_p\mathbf{x})]\right\} \\ &= |G_\mathbf{x}|^{-1} \sum_{g\in G} e(\mathbf{h}, \mathbf{t}_g) \\ &\quad \times \left\{|\Gamma_\mathbf{x}|^{-1} \sum_{p=1}^{r} e(\mathbf{h}, \mathbf{R}_g\mathbf{d}_p)e[(\mathbf{R}_g\mathbf{C}_p)^T\mathbf{h}, \mathbf{x}]\right\}. \\ &\qquad\qquad (A.3.1)\end{aligned}$$

However, because of the non-crystallographic character of Γ, the amalgamation of the transformations in G and Γ fails to produce a closed group. As a result the rearrangement on the second line of (A.2.3), which is a finite group permutation, cannot take place, so that the product $\Xi^{\text{nc}}_{\mathbf{h}_u}\Xi^{\text{nc}}_{\varepsilon\mathbf{h}_v}$ does not admit a linearization in terms of Ξ^{nc}'s but only in

terms of e's:

$$\Xi^{\rm nc}(\mathbf{h}_u,\mathbf{x})\Xi^{\rm nc}(\varepsilon\mathbf{h}_v,\mathbf{x})$$
$$=|G_{\mathbf{x}}|^{-2}|\Gamma_{\mathbf{x}}|^{-2}\sum_{g_u\in G}\sum_{g_v\in G}\sum_{p_u=1}^{r}\sum_{p_v=1}^{r}e[\mathbf{h}_u,\mathbf{R}_{g_u}\mathbf{d}_{p_u}+\mathbf{t}_{g_u}]$$
$$\times e[\varepsilon\mathbf{h}_v,\mathbf{R}_{g_v}\mathbf{d}_{p_v}+\mathbf{t}_{g_v}]$$
$$\times e[(\mathbf{R}_{g_u}\mathbf{C}_{p_u})^T\mathbf{h}_u+\varepsilon(\mathbf{R}_{g_v}\mathbf{C}_{p_v})^T\mathbf{h}_v,\mathbf{x}] \qquad (A.3.2)$$

where the vectors $(\mathbf{R}_g\mathbf{C}_p)^T\mathbf{h}$ are in general non-integral. These circumstances cause no difficulty, because the prior distribution $m(\mathbf{x})$ of the atoms obeying the local symmetry Γ is now *localized* (non-periodic), so that its *continuous transform* $\mathscr{F}_c^{-1}[m]$ can be defined unambiguously and calculated numerically. We may thus complete the calculation by writing

$$\langle\Xi^{\rm nc}(\mathbf{h}_u,\cdot)\Xi^{\rm nc}(\varepsilon_{uv}\mathbf{h}_v,\cdot)\rangle$$
$$=\sum_{g_u\in G}\sum_{g_v\in G}\sum_{p_u=1}^{r}\sum_{p_v=1}^{r}e[\mathbf{h}_u,\mathbf{R}_{g_u}\mathbf{d}_{p_u}+\mathbf{t}_{g_u}]$$
$$\times e[\varepsilon_{uv}\mathbf{h}_v,\mathbf{R}_{g_v}\mathbf{d}_{p_v}+\mathbf{t}_{g_v}]$$
$$\times\mathscr{F}_c^{-1}[|G_{\mathbf{x}}|^{-2}|\Gamma_{\mathbf{x}}|^{-2}m][(\mathbf{R}_{g_u}\mathbf{C}_{p_u})^T\mathbf{h}_u$$
$$+\varepsilon_{uv}(\mathbf{R}_{g_v}\mathbf{C}_{p_v})^T\mathbf{h}_v] \qquad (A.3.3)$$

and substituting this quantity (u, v, ε_{uv}) into (*A*.2.5) to get the desired moment $\langle\xi_u^{\rm nc}\xi_v^{\rm nc}\rangle$. Recursion yields moments (hence cumulants) of any order.

Once again, this result is completely general in that it accommodates any space group G, any local symmetry Γ, any non-uniform distribution $m(\mathbf{x})$, and any admixture of centric and acentric reflexions H.

A.4. Multiple crystal forms

We now wish to evaluate moments of the components of the 'clutched' structure-factor contributions (6.2), *i.e.* quantities of the form

$$\langle\xi_u^{(\ell){\rm nc}}\xi_v^{(\ell'){\rm nc}}\rangle \qquad (A.4.1)$$

where

$$\xi_u^{(\ell){\rm nc}}(\mathbf{x})=\mathscr{R}e[\exp(-i\omega_u^{(\ell)})\Xi^{(\ell){\rm nc}}(\mathbf{h}_u^{(\ell)},L^{(\ell)}\mathbf{x})] \qquad (A.4.2)$$

and

$$\Xi^{(\ell){\rm nc}}(\mathbf{h}^{(\ell)},\mathbf{x})=|\Gamma_{\mathbf{x}}^{(\ell)}|^{-1}\sum_{p_\ell=1}^{r_\ell}\left(|G_{\mathbf{x}}^{(\ell)}|^{-1}\times\sum_{g_\ell\in G^{(\ell)}}e\{\mathbf{h}^{(\ell)},S_{g_\ell}^{(\ell)}[T_{p_\ell}^{(\ell)}(\mathbf{L}^{(\ell)}\mathbf{x})]\}\right), \qquad (A.4.3)$$

the $\mathbf{L}^{(\ell)}$ being the clutching transformations (6.1), and where u lies in the index range for form ℓ while v lies on that for form ℓ'.

If $\ell=\ell'$, we may use (*A*.3.3) after incorporating $\mathbf{L}^{(\ell)}$ into the local symmetry $\Gamma^{(\ell)}$ of form ℓ. Otherwise a brute-force expansion generalizing (*A*.3.2), followed by evaluations of the continuous transform $\mathscr{F}_c^{-1}[m]$, yields

$$\langle\Xi^{(\ell){\rm nc}}(\mathbf{h}_u^{(\ell)},\cdot)\Xi^{(\ell'){\rm nc}}(\varepsilon_{uv}\mathbf{h}_v^{(\ell')},\cdot)\rangle$$
$$=\sum_{g_u\in G}\sum_{g_v\in G}\sum_{p_u=1}^{r_u}\sum_{p_v=1}^{r_v}e[\mathbf{h}_u^{(\ell)},\mathbf{R}_{g_u}^{(\ell)}(\mathbf{C}_{p_u}^{(\ell)}\boldsymbol{\delta}^{(\ell)}+\mathbf{d}_{p_u}^{(\ell)})+\mathbf{t}_{g_u}^{(\ell)}]$$
$$\times e[\varepsilon_{uv}\mathbf{h}_v^{(\ell')},\mathbf{R}_{g_v}^{(\ell')}(\mathbf{C}_{p_v}^{(\ell')}\boldsymbol{\delta}^{(\ell')}+\mathbf{d}_{p_v}^{(\ell')})+\mathbf{t}_{g_v}^{(\ell')}]$$
$$\times\mathscr{F}_c^{-1}[|G_{\mathbf{x}}^{(\ell)}|^{-1}|G_{\mathbf{x}}^{(\ell')}|^{-1}|\Gamma_{\mathbf{x}}^{(\ell)}|^{-1}|\Gamma_{\mathbf{x}}^{(\ell')}|^{-1}m]$$
$$\times[(\mathbf{R}_{g_u}^{(\ell)}\mathbf{C}_{p_u}^{(\ell)}\boldsymbol{\Delta}_{p_u}^{(\ell)})^T\mathbf{h}_u^{(\ell)}$$
$$+\varepsilon_{uv}(\mathbf{R}_{g_v}^{(\ell')}\mathbf{C}_{p_v}^{(\ell')}\boldsymbol{\Delta}_{p_v}^{(\ell')})^T\mathbf{h}_v^{(\ell')}]. \qquad (A.4.4)$$

The moment (*A*.4.1) is then obtained by means of formula (*A*.2.2).

By recursion, moments (and so cumulants) of any order may be evaluated numerically from the Fourier coefficients of $m(\mathbf{x})$.

This completes the task of extending Bertaut's structure-factor algebra in order to calculate explicitly the moments and cumulants required for the effective computation of the saddlepoint approximations to the most general joint distribution considered in this paper.

References

AMIT, A. G., MARIUZZA, R. A., PHILLIPS, S. E. V. & POLJAK, R. J. (1986). *Science*, **233**, 747-753.

ARNOLD, E., VRIEND, G., LUO, M., GRIFFITH, J. P., KAMER, G., ERICKSON, J. W., JOHNSON, J. E. & ROSSMANN, M. G. (1987). *Acta Cryst.* A**43**, 346-361.

BARAKAT, R. (1974). *Opt. Acta*, **21**, 903-921.

BENTLEY, G. A., LEWIT-BENTLEY, A., FINCH, J. T., PODJARNY, A. D. & ROTH, M. (1984). *J. Mol. Biol.* **176**, 55-75.

BERTAUT, E. F. (1955*a*). *Acta Cryst.* **8**, 537-543, 554-548.

BERTAUT, E. F. (1955*b*). *Acta Cryst.* **8**, 823-832.

BERTAUT, E. F. (1956). *Acta Cryst.* **9**, 322.

BERTAUT, E. F. (1959). *Acta Cryst.* **12**, 541-549, 570-574.

BERTAUT, E. F. & WASER, J. (1957). *Acta Cryst.* **10**, 606-607.

BEURSKENS, P. T. (1987). *Acta Cryst.* A**43**, 283-284.

BHAT, T. N. & BLOW, D. M. (1982). *Acta Cryst.* A**38**, 21-29.

BJORKMAN, P. J., SAPER, M. A., SAMRAOUI, B., BENNET, W. S., STROMINGER, J. L. & WILEY, D. C. (1987). *Nature (London)*, **329**, 506-512.

BLACKWELL, D. & GIRSHICK, M. A. (1954). *Theory of Games and Statistical Decisions.* New York: John Wiley.

BLOOMER, A. C., CHAMPNESS, J. N., BRICOGNE, G., STADEN, R. & KLUG, A. (1978). *Nature (London)*, **276**, 362-368.

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

BLUNDELL, T. L. & JOHNSON, L. N. (1976). *Protein Crystallography.* London: Academic Press.

BOX, G. E. P. & TIAO, G. C (1973). *Bayesian Inference in Statistical Analysis.* Reading, MA: Addison-Wesley.

BRAGG, W. L. & PERUTZ, M. F. (1952). *Acta Cryst.* **5**, 277-283.

BRICOGNE, G. (1974). *Acta Cryst.* A**30**, 395-405.

BRICOGNE, G. (1976). *Acta Cryst.* A**32**, 832-847.

BRICOGNE, G. (1982*a*). In *Computational Crystallography*, edited by D. SAYRE, pp. 223-230. Oxford Univ. Press.

BRICOGNE, G. (1982*b*). In *Computational Crystallography*, edited by D. SAYRE, pp. 258-264. Oxford Univ. Press.

BRICOGNE, G. (1984). *Acta Cryst.* A**40**, 410-445.

BRICOGNE, G. (1987). *Acta Cryst.* **A43**, C278.
BRICOGNE, G. (1988). In *Crystallographic Computing* 4, edited by N. ISAACS & M. R. TAYLOR. IUCr/Oxford Univ. Press.
BRITTEN, P. L. & COLLINS, D. M. (1982). *Acta Cryst.* **A38**, 129-132.
BRYAN, R. K. & BANNER, D. W. (1987). *Acta Cryst.* **A43**, 556-564.
CAMALLI, M., GIACOVAZZO, C. & SPAGNA, R. (1985). *Acta Cryst.* **A41**, 605-613.
CAMALLI, M., GIACOVAZZO, C. & SPAGNA, R. (1987). *Acta Cryst.* **A43**, 285-286.
CARATHÉODORY, C. (1911). *Rend. Circ. Mat. Palermo*, **32**, 193-217.
CASCARANO, G. & GIACOVAZZO, C. (1985). *Acta Cryst.* **A41**, 408-413.
CASTLEDEN, I. R. (1987). *Acta Cryst.* **A43**, 384-393.
CROWTHER, R. A. (1967). *Acta Cryst.* **22**, 758-764.
CROWTHER, R. A. (1969). *Acta Cryst.* **B25**, 2572-2580.
DANIELS, H. E. (1954). *Ann. Math. Stat.* **25**, 631-650.
DE GROOT, M. H. (1970). *Optimal Statistical Decisions.* New York: McGraw-Hill.
DICKERSON, R. E., KENDREW, J. C. & STRANDBERG, B. E. (1961). *Acta Cryst.* **14**, 1188-1195.
FORTIER, S. (1987). Personal communication.
FORTIER, S., FRASER, M. E. & MOORE, N. J. (1986). *Acta Cryst.* **A42**, 149-156.
FORTIER, S., WEEKS, C. M. & HAUPTMAN, H. (1984). *Acta Cryst.* **A40**, 544-548, 646-651.
FOSTER, F. & HARGREAVES, A. (1963). *Acta Cryst.* **16**, 1124-1133, 1133-1139.
FOWLER, R. H. (1936). *Statistical Mechanics*, 2nd ed. Cambridge Univ. Press.
FRASER, M. E. (1987). PhD thesis. Queens Univ., Kingston, Ontario, Canada.
FRENCH, S. (1978). *Acta Cryst.* **A34**, 728-738.
FRENCH, S. & WILSON, K. (1978). *Acta Cryst.* **A34**, 517-525.
GAYKEMA, W. P. J., VOLBEDA, A. & HOL, W. G. J. (1985). *J. Mol. Biol.* **187**, 225-275.
GIACOVAZZO, C. (1980). *Direct Methods in Crystallography.* London: Academic Press.
GIACOVAZZO, C. (1983*a*). *Acta Cryst.* **A39**, 585-592.
GIACOVAZZO, C. (1983*b*). *Acta Cryst.* **A39**, 685-692.
GIACOVAZZO, C. (1987). *Acta Cryst.* **A43**, 73-75.
GRAMLICH, V. (1984). *Acta Cryst.* **A40**, 610-616.
GRAYBILL, F. A. (1969). *Introduction to Matrices with Applications in Statistics.* Belmont: Wadsworth Publishing Company.
GREEN, D. W., INGRAM, V. M. & PERUTZ, M. F. (1954) *Proc. R. Soc. London*, **225**, 287-307.
GULL, S. F., LIVESEY, A. K. & SIVIA, D. S. (1987). *Acta Cryst.* **A43**, 112-117.
HARKER, D. & KASPER, J. S. (1948) *Acta Cryst.* **1**, 70-75.
HARRISON, S. C., OLSON, A. J., SCHUTT, C. E., WINKLER, F. K. & BRICOGNE, G. (1978). *Nature (London)*, **276**, 368-373.
HAUPTMAN, H. (1975). *Acta Cryst.* **A30**, 472-476.
HAUPTMAN, H. (1982). *Acta Cryst.* **A38**, 289-294, 632-641.
HEINERMAN, J. J. L., KRABBENDAM, H., KROON, J. & SPEK, A. L. (1978). *Acta Cryst.* **A34**, 447-450.
HENDRICKSON, W. A. & LATTMAN, E. E. (1970). *Acta Cryst.* **B26**, 136-143.
HENDRICKSON, W. A., LOVE, W. E. & KARLE, J. (1973). *J. Mol. Biol.* **74**, 331-361.
HOGLE, J. M., CHOW, M. & FILMAN, D. J. (1985). *Science*, **229**, 1358-1365.
HÖRMANDER, L. (1973). *An Introduction to Complex Analysis in Several Variables*, 2nd ed. Amsterdam: North-Holland.
JAYNES, E. T. (1957). *Phys. Rev.* **106**, 620-630.
JAYNES, E. T. (1968). *IEEE Trans. Syst. Sci. Cybern.* **4**, 227-241.
KARLE, J. (1983). *Acta Cryst.* **A39**, 800-805.
KARLE, J. (1984). *Acta Cryst.* **A40**, 1-4, 4-11, 366-373, 374-379, 526-531.
KARLE, J. (1985). *Acta Cryst.* **A41**, 182-189, 387-394.
KARLE, J. (1986). *Acta Cryst.* **A42**, 246-253.
KARLE, J. & HAUPTMAN, H. (1950). *Acta Cryst.* **3**, 181-187.
KHINCHIN, A. I. (1949). *Mathematical Foundations of Statistical Mechanics.* New York: Dover.
KLOP, E. A., KRABBENDAM, H. & KROON, J. (1987). *Acta Cryst.* **A43**, 810-820.
KLUG, A. (1958). *Acta Cryst.* **11**, 515-543.
KROON, J., SPEK, A. L. & KRABBENDAM, H. (1977). *Acta Cryst.* **A33**, 382-385.
LEBEDEV, N. N. (1972). *Special Functions and their Applications.* New York: Dover.
LESLIE, A. G. W. (1987). *Acta Cryst.* **A43**, 134-136.
LIPSON, H. & WOOLFSON, M. M. (1952). *Acta Cryst.* **5**, 680-682.
LIVESEY, A. K. & SKILLING, J. (1985). *Acta Cryst.* **A41**, 113-122.
MAIN, P. & ROSSMANN, M. G. (1966). *Acta Cryst.* **21**, 67-72.
MARVIN, D. A., BRYAN, R. K. & NAVE, C. (1987). *J. Mol. Biol.* **193**, 315-343.
NARAYAN, R. & NITYANANDA, R. (1982). *Acta Cryst.* **A38**, 122-128.
NEYMAN, J. & PEARSON, E. (1933*a*). *Proc. Cambridge Philos. Soc.* **29**, 492-510.
NEYMAN, J. & PEARSON, E. (1933*b*). *Philos. Trans. R. Soc. London Ser. A*, **231**, 289-337.
NIGAM, G. D. (1972). *Indian J. Pure Appl. Phys.* **10**, 655-656.
NIGAM, G. D. & WILSON, A. J. C. (1980). *Acta Cryst.* **A36**, 832-833.
NORDMAN, C. E. (1980). *Acta Cryst.* **A36**, 747-754.
OATLEY, S. & FRENCH, S. (1982). *Acta Cryst.* **A38**, 537-549.
PALEY, R. E. A. C. & WIENER, N. (1934). *Fourier Transforms in the Complex Domain. Am. Math. Soc. Colloquium Publications*, Vol. 19. Providence: American Mathematical Society.
PESCHAR, R. & SCHENK, H. (1987). *Acta Cryst.* **A43**, 513-522.
PHILLIPS, J. C. & HODGSON, K. O. (1980). *Acta Cryst.* **A36**, 856-864.
PRADHAN, D., GHOSH, S. & NIGAM, G. D. (1985). *Structure and Statistics in Crystallography*, edited by A. J. C. WILSON, pp. 43-51. Guilderland: Adenine Press.
PRICK, P. A. J., BEURSKENS, P. T. & GOULD, R. O. (1978). *Acta Cryst.* **A34**, S42.
PRICK, P. A. J., BEURSKENS, P. T. & GOULD, R. O. (1983). *Acta Cryst.* **A39**, 570-576.
ROGERS, D. & WILSON, A. J. C. (1953). *Acta Cryst.* **6**, 439-449.
ROSSMANN, M. G. (1972). *The Molecular Replacement Method.* New York: Gordon & Breach.
ROSSMANN, M. G. & BLOW, D. M. (1962). *Acta Cryst.* **15**, 24-31.
ROSSMANN, M. G. & BLOW, D. M. (1963). *Acta Cryst.* **16**, 39-45.
ROTH, M. (1987). *Acta Cryst.* **A43**, 780-787.
ROTH, M., LEWIT-BENTLEY, A. & BENTLEY, G. A. (1984). *J. Appl. Cryst.* **17**, 77-84.
SAYRE, D. (1952). *Acta Cryst.* **5**, 843.
SCHEVITZ, R. W., PODJARNY, A. D., ZWICK, M., HUGHES, J. J. & SIGLER, P. B. (1981). *Acta Cryst.* **A37**, 669-677.
SCHWARTZ, L. (1966). *Théorie des Distributions.* Paris: Hermann.
SHANNON, C. E. (1949). *Proc. Inst. Radio Eng. NY*, **37**, 10-21.
SHMUELI, U. (1979). *Acta Cryst.* **A35**, 282-286.
SHMUELI, U. (1982). *Acta Cryst.* **A38**, 362-371.
SHMUELI, U. & KALDOR, U. (1981). *Acta Cryst.* **A37**, 76-80.
SHMUELI, U. & KALDOR, U. (1983). *Acta Cryst.* **A39**, 615-621.
SHMUELI, U. & WEISS, G. H. (1985). *Acta Cryst.* **A41**, 401-408.
SHMUELI, U. & WEISS, G. H. (1986). *Acta Cryst.* **A42**, 240-246.
SHMUELI, U. & WEISS, G. H. (1987). *Acta Cryst.* **A43**, 93-98.
SHMUELI, U., WEISS, G. H. & KIEFER, J. E. (1985). *Acta Cryst.* **A41**, 55-59.
SHMUELI, U., WEISS, G. H., KIEFER, J. E. & WILSON, A. J. C. (1984). *Acta Cryst.* **A40**, 651-660.
SHMUELI, U. & WILSON, A. J. C. (1981). *Acta Cryst.* **A37**, 342-353.
SHMUELI, U. & WILSON, A. J. C. (1983). *Acta Cryst.* **A39**, 225-233.
SIM, G. A. (1959). *Acta Cryst.* **12**, 813-815.
SRINIVASAN, R. & PARTHASARATHY, S. (1976). *Some Statistical Applications in X-ray Crystallography.* Oxford: Pergamon Press.
SZEGÖ, G. (1920). *Math. Z.* **6**, 167-202.
TOEPLITZ, O. (1911). *Rend. Circ. Mat. Palermo*, **32**, 191-192.

TSOUCARIS, G. (1970). *Acta Cryst.* A**26**, 492-499.
VAN DEN HARK, TH. E. M., PRICK, P. A. J. & BEURSKENS, P. T. (1976). *Acta Cryst.* A**32**, 816-821.
WANG, B. C. (1985). In *Methods of Enzymology*, Vol. 115: *Diffraction Methods for Biological Macromolecules*, edited by H. WYCKOFF, C. H. W. HIRS & S. N. TIMASHEFF. New York: Academic Press.
WEISS, G. H., SHMUELI, U., KIEFER, J. E. & WILSON, A. J. C. (1985). *Structure and Statistics in Crystallography*, edited by A. J. C. WILSON, pp. 23-42. Guilderland: Adenine Press.
WIENER, N. (1949). *Extrapolation, Interpolation, and Smoothing of Stationary Time Series.* Cambridge, MA: MIT Press.
WILKINS, S. W. & STUART, D. (1986). *Acta Cryst.* A**42**, 197-202.
WILSON, A. J. C. (1949). *Acta Cryst.* **2**, 318-321.
WILSON, A. J. C. (1952). *Research*, **5**, 589-590.
WILSON, A. J. C. (1964). *Acta Cryst.* **17**, 1591-1592.
WILSON, A. J. C. (1987). *Acta Cryst.* A**43**, 554-556.
WILSON, I. A., SKEHEL, J. J. & WILEY, D. C. (1981). *Nature (London)*, **289**, 366-373.

Acta Cryst. (1988). A**44**, 545-551

Low-Resolution Phase Extension and Refinement by Maximum Entropy

BY A. D. PODJARNY AND D. MORAS

Laboratoire de Cristallographie Biologique, IBMC, 15 Rue R. Descartes, 67084 Strasbourg, France

J. NAVAZA

Laboratoire de Physique, Centre Universitaire Pharmaceutique, Tour B, 92290 Châtenay-Malabry, France

AND P. M. ALZARI

Immunologie Structurale, Institut Pasteur, 25 Rue du Dr Roux, 75724 Paris, France

(*Received* 27 *October* 1987; *accepted* 11 *March* 1988)

Abstract

The problem of phase refinement and extension at very low resolution (30-25 Å) is treated with an algorithm that combines a maximum-entropy approach, a binary modelling of the electron density, refinement of the proposed map against the observed amplitudes and solvent flattening outside a molecular envelope. The algorithm is applied to data for the complex of aspartyl-tRNA and aspartyl-tRNA synthetase in three different cases: (1) X-ray amplitudes and phases calculated from a partial model; (2) mixed observed and calculated X-ray amplitudes and phases from a partial model; and (3) observed neutron amplitudes and phases from a very approximate model. The change of correlation with the correct map at 30 Å resolution is used as a measure of correctness. Upon application of the algorithm, this correlation changes from 59 to 97% in case 1, from 59 to 77% in case 2 and from 72 to 90% in case 3. In all cases, the method is successful in correcting large phase errors, deleting noise regions and producing the correct low-resolution molecular image.

Introduction

Macromolecular crystallography is a unique tool for imaging the structures of protein and nucleic acids. Images are obtained from the Fourier transform of the diffraction pattern of the crystal. In the classical picture the scattered radiation consists of X-rays at a resolution where individual atoms are close to being resolved, and the phases with their error estimation are obtained by the multiple-isomorphous-replacement (MIR) method (Blow & Crick, 1959).

When all the necessary conditions are fulfilled, the classical approach is extremely powerful and a very detailed image of the macromolecule is obtained. However, it is not always possible to obtain high-quality crystals that diffract to high resolution and the necessary heavy-atom derivatives. In this case, alternative phasing techniques can be used. These are varied, and the following examples can be cited (the list is clearly not exhaustive).

(1) Electron microscopy of an ordered specimen can give a 3D image of 7 Å resolution (Henderson & Unwin, 1975), though most of the image reconstructions have been limited to about 25 Å resolution.

(2) A low-resolution translation search with a crude model can generate phases between 30 and 15 Å resolution (Podjarny *et al.*, 1987).

(3) Neutron diffraction with different D_2O/H_2O levels can be used instead of heavy atoms, if one component is known. Phases are generally good to 30 Å resolution, and can extend as far as 15 Å resolution (Bentley, Lewitt-Bentley, Finch, Podjarny & Roth, 1984).

0108-7673/88/040545-07$03.00

© 1988 International Union of Crystallography