Acta Crystallographica Section A
Foundations and Advances
ISSN 2053-2733

Received 17 June 2014
Accepted 27 October 2014

An alternative method for the calculation of joint probability distributions. Application to the expectation of the triplet invariant

J. Brosius

Kerkweg 7A, Rotselaar 3110, Belgium. Correspondence e-mail: brosius.jan@gmail.com

This paper presents a completely new method for the calculation of expectations (and thus joint probability distributions) of structure factors or phase invariants. As an example, a first approximation of the expectation of the triplet invariant (up to a constant) is given and a *complex* number is obtained. Instead of considering the atomic vector positions or reciprocal vectors as the fundamental random variables, the method samples over all functions (distributions) with a given number of atoms and given Patterson function. The aim of this paper was to explore the feasibility of the method, so the easiest problem was chosen: the calculation of the expectation value of the triplet invariant in $P1$. Calculation of the joint *probability* distribution of the triplet is not performed here but will be done in the future.

1. Introduction

Let us consider the definition of the structure factor

$$E_{\mathbf{h}} = \int \mathrm{d}\mathbf{x}\, \exp(2\pi i \mathbf{h}\cdot\mathbf{x})\rho(\mathbf{x}). \quad (1)$$

In crystallography only the absolute values $|E_{\mathbf{h}}|$ are given from measurements, the phase $\varphi_{\mathbf{h}}$ $[E_{\mathbf{h}} = |E_{\mathbf{h}}|\exp(i\varphi_{\mathbf{h}})]$ is undetermined. One tries to calculate the $\varphi_{\mathbf{h}}$ by statistical methods. In the past (and up to now) there were only two statistical approaches:

(i) One considers the function

$$\mathbf{h} \longrightarrow E_{\mathbf{h}}$$

and one uses a uniform measure on reciprocal space (the space of all $\mathbf{h}$). For this we do not need to know the exact distribution $\rho(\mathbf{x})$ [as is clear from the definition of $E_{\mathbf{h}}$ (ρ does not depend on $\mathbf{h}$)] but use only the additional information that ρ is a sum of N 'peak' functions,

$$\rho = \sum_{i=1}^{N} \frac{1}{N^{1/2}} \delta_{\mathbf{r}_i},$$

where $\delta_{\mathbf{r}_i}(\mathbf{x}) \equiv \delta(\mathbf{x}-\mathbf{r}_i)$ and the $\mathbf{r}_i$ are the actual but unknown atomic positions. The 'peak strength' $1/N^{1/2}$ is determined by the requirement that the mean of $|E_{\mathbf{h}}|^2$ equals 1 when sampling uniformly over reciprocal space,

$$\left\langle |E_{\mathbf{h}}|^2 \right\rangle_{\mathbf{h}} = 1.$$

With this setup the random variable $E_{\mathbf{h}}$ becomes a function of N random variables $\mathbf{h} \rightarrow \exp(2\pi i \mathbf{h}\cdot\mathbf{r}_k)$,

$$E_{\mathbf{h}} = \frac{1}{N^{1/2}} \sum_{k=1}^{N} \exp(2\pi i \mathbf{h}\cdot\mathbf{r}_k).$$

There is, however, one problem. In order for the random variables $\mathbf{h} \rightarrow \exp(2\pi i \mathbf{h}\cdot\mathbf{r}_k)$ to be independent the following condition must be satisfied.

There are *no* relations among the $\mathbf{r}_k$ of the form

$$\sum_{k=1}^{N} n_k \mathbf{r}_k = 0, \quad (2)$$

where the n_k are integer numbers. Theoretically, we may circumvent this problem by displacing the $\mathbf{r}_k$ by a very small amount ε_k $(\mathbf{r}_k \rightarrow \mathbf{r}_k + \varepsilon_k)$. But experimentally there is a problem: we then need a *very* large number of experimental data $|E_{\mathbf{h}}|$ (without experimental error) to be able to use this theoretical assumption and we know that this is not the case. Because of this condition (2) becomes in reality:

There are *no* relations among the $\mathbf{r}_k$ of the form

$$\sum_{k=1}^{N} n_k \mathbf{r}_k \simeq 0.$$

One then calculates the joint probability for a set of structure factors and then one calculates the conditional probability of the phases given the magnitudes of the structure factors. This method was proposed by Hauptman and Karle (*e.g.* Karle & Hauptman, 1953) and led to some theoretically interesting algebraic relations, but they did not work in practice.

(ii) One considers the atomic position vectors as the fundamental random variables, in which case the structure factors become random variables of the $\mathbf{x}_k$:

$$(\mathbf{x}_k)_k \rightarrow E_{\mathbf{h}} = \frac{1}{N^{1/2}} \sum_{k=1}^{N} \exp(2\pi i \mathbf{h}\cdot\mathbf{x}_k),$$

where every $\mathbf{x}_k$ ranges *uniformly* and *independently* over the unit cell (see *e.g.* Giacovazzo, 1975; Heinerman, 1975; Klug,

1958). The probability distribution of a set of structure factors is then calculated and eventually the conditional distributions of the phases given the magnitudes of the structure factors are used as the *a posteriori* distribution of the phases. In his classical paper Klug (1958) showed that the two approaches had a drawback: the strength of the derived formulas depended on inverse powers of $N^{1/2}$ (N being the number of atoms). So structures with many atoms give less reliable formulas than structures with less atoms. Klug also showed that the algebraic equations that one obtains with the first statistical approach are less reliable for larger N. Most direct methods (*e.g.* Xu & Hauptman, 2004) one uses today are based on these two statistical approaches. Our approach in the past was to consider suitably chosen prior distributions for the atomic position vectors (see Brosius, 2008*a*,*b*,*c*, 2012). Unfortunately when one gives up the *independence* of the random position vectors the calculations are more tedious and up to the approximation used we did not get good results for very large N. In the second part of Brosius (2012) we used special constraints on the atomic position vectors and still kept the independence of the atomic vectors: the results were promising and the calculation of the joint probability density (j.p.d.) of structure factors was easy.

2. Sampling over all atomic distributions ρ

If one looks at the definition of a structure factor (1) there is a third possibility. Indeed, one may sample over ρ (uniformly if possible) and then one can consider the random variables:

$$\rho \longrightarrow \int \mathrm{d}\mathbf{x}\, \exp(2\pi i \mathbf{h}\cdot\mathbf{x})\rho(\mathbf{x}) = E_{\mathbf{h}}.$$

What are now the conditions that ρ should obey? Clearly the first and most important one is: ρ must be compatible with the given Patterson function P; thus

$$\int \mathrm{d}\mathbf{y}\, \rho(\mathbf{y})\rho(\mathbf{x}+\mathbf{y}) = P(\mathbf{x}) \text{ for all } \mathbf{x}.$$

There are also two additional conditions that ρ can satisfy:

(i) If we can suppose that we have a point particle structure with strengths $1/N^{1/2}$ then

$$\rho^2 = \frac{\delta(\mathbf{0})}{N^{1/2}}\,\rho.$$

The $\delta(\mathbf{0})$ factor is necessary if one supposes that the peaks are delta functions. This gives Sayre's equation

$$U_{\mathbf{h}} = \langle E_{\mathbf{k}} E_{\mathbf{h}-\mathbf{k}} \rangle_{\mathbf{k}}$$

if we take $\delta(\mathbf{0})$ = number of observed structure factors.

(ii) We can apply the condition

$$\int \mathrm{d}\mathbf{x}\, \rho(\mathbf{x}) = N^{1/2}$$

if one knows the total number of atoms in the unit cell.

3. The expectation value of the triplet up to first order

A sampling over all ρ's seems, however, an impossible task to do. Fortunately, a similar problem arises in quantum field theory. There one uses such a sampling for 'second quantization'. Good introductions to this method can be found in Weinberg (2005*a*,*b*), Chaichian & Demichev (2001), Masujima (2009) and Siegel (2005). We obtain for the expectation value of the triplet a *complex* number: to first order the expectation of the triplet invariant is

$$\langle E_{\mathbf{h}} E_{\mathbf{k}} E_{-\mathbf{h}-\mathbf{k}} \rangle = \mathrm{Cte}\left[\frac{1}{(\hat{Q}_{\mathbf{h}}+i)(\hat{Q}_{\mathbf{h}+\mathbf{k}}+i)} + \frac{1}{(\hat{Q}_{\mathbf{k}}+i)(\hat{Q}_{\mathbf{h}+\mathbf{k}}+i)} + \frac{1}{(\hat{Q}_{\mathbf{h}}+i)(\hat{Q}_{\mathbf{k}}+i)}\right], \tag{3}$$

where

$$\hat{Q}_{\mathbf{h}} \equiv R_{\mathbf{h}}^2 - 1.$$

The classical methods [see equations (1) and (2) above] give (only)

$$\langle E_{\mathbf{h}} E_{\mathbf{k}} E_{-\mathbf{h}-\mathbf{k}} \rangle_{\mathrm{cl}} = \frac{1}{N^{1/2}}. \tag{4}$$

However, we cannot compare blindly formula (4) with formula (3) since the constant could be equal to $1/N^{1/2}$ or perhaps worse $1/N$. But we can do something else: we can compare ratios. Indeed, let us define $\mathbb{E}(\mathbf{h},\mathbf{k})$ by

$$\mathbb{E}(\mathbf{h},\mathbf{k}) = \left[\frac{1}{(\hat{Q}_{\mathbf{h}}+i)(\hat{Q}_{\mathbf{h}+\mathbf{k}}+i)} + \frac{1}{(\hat{Q}_{\mathbf{k}}+i)(\hat{Q}_{\mathbf{h}+\mathbf{k}}+i)} + \frac{1}{(\hat{Q}_{\mathbf{h}}+i)(\hat{Q}_{\mathbf{k}}+i)}\right]. \tag{5}$$

Then we can compare the ratio

$$\frac{\langle E_{\mathbf{h}} E_{\mathbf{k}} E_{-\mathbf{h}-\mathbf{k}} \rangle}{\langle E_{\mathbf{h}_0} E_{\mathbf{k}_0} E_{-\mathbf{h}_0-\mathbf{k}_0} \rangle} = \frac{\mathrm{Cte}\,\mathbb{E}(\mathbf{h},\mathbf{k})}{\mathrm{Cte}\,\mathbb{E}(\mathbf{h}_0,\mathbf{k}_0)} = \frac{\mathbb{E}(\mathbf{h},\mathbf{k})}{\mathbb{E}(\mathbf{h}_0,\mathbf{k}_0)} \tag{6}$$

with the ratio

$$\frac{\langle E_{\mathbf{h}} E_{\mathbf{k}} E_{-\mathbf{h}-\mathbf{k}} \rangle_{\mathrm{cl}}}{\langle E_{\mathbf{h}_0} E_{\mathbf{k}_0} E_{-\mathbf{h}_0-\mathbf{k}_0} \rangle_{\mathrm{cl}}} = \frac{1/N^{1/2}}{1/N^{1/2}} = 1, \tag{7}$$

where $\mathbf{h}_0$ and $\mathbf{k}_0$ are some fixed vectors. Now both formulas (6) and (7) do not depend anymore on N. We shall discuss equation (6) numerically in the next section.

4. Using methods from quantum field theory

4.1. The setting

Let us now explain how we obtained formula (3). As said above, we must (at least?) use the constraint $\int \mathrm{d}\mathbf{y}\rho(\mathbf{y})\rho(\mathbf{x}+\mathbf{y}) = P(\mathbf{x})$ for every $\mathbf{x}$. This can be expressed by

$$\prod_{\mathbf{x}} \delta\big[\textstyle\int \mathrm{d}\mathbf{y}\, \rho(\mathbf{y})\rho(\mathbf{x}+\mathbf{y}) - P(\mathbf{x})\big].$$

That is, the electronic distribution ρ must satisfy the constraint $[\int \mathrm{d}\mathbf{y}\, \rho(\mathbf{y})\rho(\mathbf{x}+\mathbf{y}) - P(\mathbf{x})]$ at every $\mathbf{x}$. If for some $\mathbf{x}$ this is not satisfied, then the corresponding delta function is zero and then $\prod_{\mathbf{x}}$ gives zero for the total product. Next we must impose

(that is, if we can be sure that we have an almost equal atom point distribution) the condition

$$\rho^2(\mathbf{x}) = \frac{\delta(\mathbf{0})}{N^{1/2}}\,\rho(\mathbf{x})$$

at every point $\mathbf{x}$. The coefficient $\delta(\mathbf{0})$ is a constant that is (annoyingly) ∞; this is a consequence of the point-atom structure

$$\rho(\mathbf{x}) = \sum_{i=1}^{N} \delta(\mathbf{x} - \mathbf{r}_i)$$

and the relations

$$\delta(\mathbf{x} - \mathbf{r}_i)\delta\big(\mathbf{x} - \mathbf{r}_j\big) \equiv 0 \text{ if } i \neq j$$
$$\delta(\mathbf{x} - \mathbf{r}_i)^2 \equiv \delta(\mathbf{0})\,\delta(\mathbf{x} - \mathbf{r}_i).$$

We cannot do meaningful calculations with ∞. Fortunately NSA [non-standard analysis; see Diener & Reeb (1989), Nelson (1977, 1987)] comes to our rescue. We shall consider $\delta(\mathbf{0})$ as an infinite *number* and not as ∞. Now we can do calculations! For example, $[\delta(\mathbf{0})]^{1/2}$ now has a meaning: it is an infinite number but we can compare it with $\delta(\mathbf{0})$. Indeed $[\delta(\mathbf{0})]^{1/2}/\delta(\mathbf{0}) = 1/[\delta(\mathbf{0})]^{1/2}$ and is an infinitesimal number. Hopefully, these infinite numbers disappear in the final formula. We will see that this is indeed the case. We regard this as one indication that we have calculated correctly. [Another indication is the symmetry of $\mathbb{E}(\mathbf{h}, \mathbf{k})$ in $\mathbf{h}$, $\mathbf{k}$ and $\mathbf{h} + \mathbf{k}$.] We now impose the condition $\rho^2(\mathbf{x}) = [\delta(\mathbf{0})/N^{1/2}]\rho(\mathbf{x})$ at every $\mathbf{x}$ by considering an enhanced product

$$\prod_{\mathbf{x}} \delta\left[\rho^2(\mathbf{x}) - \frac{\delta(\mathbf{0})}{N^{1/2}}\,\rho(\mathbf{x})\right]\delta\left[\int \mathrm{d}\mathbf{y}\,\rho(\mathbf{y})\rho(\mathbf{x}+\mathbf{y}) - P(\mathbf{x})\right].$$

Again this product is equal to zero whenever one (or more) constraints (delta's) is zero. Finally, the last constraint does not depend on $\mathbf{x}$ but is the *constant* delta $\delta(\int \mathrm{d}\mathbf{x}\,\rho - N^{1/2})$. This constraint imposes the last general information about ρ. The total function that we have to consider is then

$$\prod_{\mathbf{x}}\left\{\delta\left[\rho^2(\mathbf{x}) - \frac{\delta(\mathbf{0})}{N^{1/2}}\,\rho(\mathbf{x})\right]\delta\left[\int \mathrm{d}\mathbf{y}\,\rho(\mathbf{y})\rho(\mathbf{x}+\mathbf{y}) - P(\mathbf{x})\right]\right\} \times \delta\left[\int \mathrm{d}\mathbf{x}\,\rho(\mathbf{x}) - N^{1/2}\right].$$

It remains for us now to calculate a probability.

4.2. The form of the probability distribution over the sampling space of all ρ

To do this we want to *sample* uniformly over the electronic distributions ρ. Since we have used all our constraints, we cannot *a priori* prefer one disribution ρ' over another ρ. Let us write then the infinitesimal probability that ρ (given the constraint function from above) lies between $\rho(\mathbf{x})$ and $\rho(\mathbf{x}) + \mathrm{d}\rho(\mathbf{x})$ for every $\mathbf{x}$. We express this by

$$\begin{aligned}\text{Prob}\,(\rho)\mathrm{d}\rho = \prod_{\mathbf{x}} \mathrm{d}\rho(\mathbf{x})\,\delta\left[\rho^2(\mathbf{x}) - \frac{\delta(\mathbf{0})}{N^{1/2}}\,\rho(\mathbf{x})\right] \\ \times \delta\left[\int \mathrm{d}\mathbf{y}\,\rho(\mathbf{y})\rho(\mathbf{x}+\mathbf{y}) - P(\mathbf{x})\right] \\ \times \delta\left[\int \mathrm{d}\mathbf{x}\,\rho(\mathbf{x}) - N^{1/2}\right].\end{aligned} \tag{8}$$

Finally, we remark that $\rho(\mathbf{x})$ takes on values between 0 and ∞ [it is a sum of delta functions $\sum_{i=1}^{N}\delta(\mathbf{x} - \mathbf{r}_i)$!]. So we have to integrate as follows: $\prod_{\mathbf{x}}\int_0^\infty \mathrm{d}\rho(\mathbf{x})$ to get the probability over all ρ. Thus we have for the total probability

$$\begin{aligned}\int \text{Prob}\,(\rho)\,\mathrm{d}\rho = \text{Cte}\prod_{\mathbf{x}}\int_0^\infty \mathrm{d}\rho(\mathbf{x})\,\delta\left[\rho^2(\mathbf{x}) - \frac{\delta(\mathbf{0})}{N^{1/2}}\,\rho(\mathbf{x})\right] \\ \times \delta\left[\int \mathrm{d}\mathbf{y}\,\rho(\mathbf{y})\rho(\mathbf{x}+\mathbf{y}) - P(\mathbf{x})\right] \\ \times \delta\left[\int \mathrm{d}\mathbf{x}\,\rho(\mathbf{x}) - N^{1/2}\right]\end{aligned} \tag{9}$$

[where Cte is a normalization constant. This constant can be calculated by imposing the condition $\int \text{Prob}\,(\rho)\,\mathrm{d}\rho = 1$]. This looks fine but there is a problem: our constraints are *quadratic* or *linear* in ρ and we know that we can easily calculate expressions of the form

$$\int_{-\infty}^{\infty} \mathrm{d}x\,\exp(ax^2 + bx + c)f(x),$$

whereas expressions of the form

$$\int_{0}^{\infty} \mathrm{d}x\,\exp(ax^2 + bx + c)f(x)$$

pose a problem. Happily, the constraint $\delta\{\rho^2(\mathbf{x}) - [\delta(\mathbf{0})/N^{1/2}]\rho(\mathbf{x})\}$ also eliminates negative $\rho(\mathbf{x})$; it also becomes zero if $\rho(\mathbf{x})$ becomes negative! So we can say

$$\begin{aligned}\int \text{Prob}\,(\rho)\,\mathrm{d}\rho = \text{Cte}\prod_{\mathbf{x}}\int_{-\infty}^{\infty} \mathrm{d}\rho(\mathbf{x})\,\delta\left[\rho^2(\mathbf{x}) - \frac{\delta(\mathbf{0})}{N^{1/2}}\,\rho(\mathbf{x})\right] \\ \times \delta\left[\int \mathrm{d}\mathbf{y}\,\rho(\mathbf{y})\rho(\mathbf{x}+\mathbf{y}) - P(\mathbf{x})\right] \\ \times \delta\left[\int \mathrm{d}\mathbf{x}\,\rho(\mathbf{x}) - N^{1/2}\right].\end{aligned}$$

If we now want to calculate an expectation value of a function(al) $F[\rho]$ we have to calculate

$$\begin{aligned}\langle F[\rho]\rangle_\rho &= \int \mathrm{d}\rho\, F[\rho]\mathrm{Prob}\,(\rho)\\ &= \mathrm{Cte}\prod_{\mathbf{x}}\int_{-\infty}^{\infty}\mathrm{d}\rho(\mathbf{x})\,\delta\Big[\rho^2(\mathbf{x})-\frac{\delta(\mathbf{0})}{N^{1/2}}\rho(\mathbf{x})\Big]\\ &\quad\times\delta\Big[\int\mathrm{d}\mathbf{y}\,\rho(\mathbf{y})\rho(\mathbf{x}+\mathbf{y})-P(\mathbf{x})\Big]\\ &\quad\times\delta\Big[\int\mathrm{d}\mathbf{x}\,\rho(\mathbf{x})-N^{1/2}\Big]F[\rho[\mathbf{x}]].\end{aligned}$$

And this will be more pleasingly written as

$$\begin{aligned}\langle F[\rho]\rangle_\rho &= \int \mathcal{D}\rho F[\rho]\,\mathrm{Prob}\,(\rho)\\ &= \mathrm{Cte}\int\mathcal{D}\rho(\mathbf{x})\delta\Big[\rho^2(\mathbf{x})-\frac{\delta(\mathbf{0})}{N^{1/2}}\rho(\mathbf{x})\Big]\\ &\quad\times\delta\Big[\int\mathrm{d}\mathbf{y}\,\rho(\mathbf{y})\rho(\mathbf{x}+\mathbf{y})-P(\mathbf{x})\Big]\\ &\quad\times\delta\Big[\int\mathrm{d}\mathbf{x}\,\rho(\mathbf{x})-N^{1/2}\Big]F[\rho[\mathbf{x}]],\end{aligned}$$

and we shall speak of a *functional* integral over ρ.

4.3. The essence of the method: Gaussian integration and functional derivation

As is usual (and here even necessary) we rewrite a delta as a Fourier integral [thereby introducing new (quadratic) functional integrals as is explained in the calculations (we refer to the supporting information[1])]; and a functional delta $\delta[H[\rho]]$ will give rise to another functional integral:

$$\begin{aligned}\delta[H[\rho]] &\propto \prod_{\mathbf{x}}\int_{-\infty}^{\infty}\mathrm{d}\lambda(\mathbf{x})\,\exp[i\lambda(\mathbf{x})H[\rho(\mathbf{x})]]\\ &\propto \int\mathcal{D}\lambda\exp(i\lambda H[\rho]).\end{aligned}$$

We shall see in the calculations (see the supporting information) that the result after integration over ρ (when $H[\rho]$ is quadratic in ρ) gives us again a quadratic functional, that is, for the above example,

$$\int\mathcal{D}\rho\delta[H[\rho]]\propto\int\mathcal{D}\lambda\underbrace{\int\mathcal{D}\rho\exp(i\lambda H[\rho])}_{=H_1[\lambda]},$$

where we will obtain a quadratic functional $H_1[\lambda]$. So it seems almost obvious now that we shall have to do with Gaussian integrals of the form

$$\int\mathcal{D}\rho\exp\Big[\int\mathrm{d}\mathbf{x}\,\mathrm{d}\mathbf{y}A(\mathbf{x},\mathbf{y})\rho(\mathbf{x})\rho(\mathbf{y})+\int\mathrm{d}\mathbf{x}B(\mathbf{x})\rho(\mathbf{x})\Big]F[\rho]$$

where $A(\mathbf{x},\mathbf{y})$ and $B(\mathbf{x})$ are independent of ρ but may depend functionally on other functions. As we shall explain in the calculations, the above functional integral is nothing else but the *continuous* version of a (*discrete*) Gaussian integral:

$$\prod_{i=1}^{M}\int_{-\infty}^{\infty}\mathrm{d}u_i\exp\Bigg(\sum_{i=1,j=1}^{M}A_{i,j}u_iu_j+\sum_{i=1}^{N}B_iu_i\Bigg)F\big(\{u_i\}\big),$$

and that can be solved as we shall see in the calculations [notice that ρ replaces u and that $\mathbf{x}$ replaces i (and $\mathbf{y}$ replaces j)]. What is so interesting about Gaussian integrals is that they can be expressed as derivations, and derivations $[\partial/\partial u_jF(\{u_i\})]$ are much easier than integrals! And as we can foresee now the *continuous* version of a discrete derivation is a *functional* derivation $\{[\delta/\delta\rho(\mathbf{x})]F[\rho]\}$. All this is amply explained in the supporting information.

A lot of this is borrowed from quantum field theory but the statement of the problem in direct methods (DM) and the use of infinite numbers like $\delta(\mathbf{0})$ is our invention (as is also the iterative use of functional integrals). Such calculations have not been done before.

As stated before, in the final expression we have unfortunately a normalization constant [our Cte (equation (3))]. We can calculate this constant in principle. But we are not really interested in this constant; what we are really interested in is the *joint conditional probability distribution* $P(\varphi_{\mathbf{h}},\varphi_{\mathbf{k}},\varphi_{\mathbf{h}+\mathbf{k}}|R_{\mathbf{h}},R_{\mathbf{k}},R_{\mathbf{h}+\mathbf{k}})$. We calculated the expectation $\langle E_{\mathbf{h}}E_{\mathbf{k}}E_{-\mathbf{h}-\mathbf{k}}\rangle$ because it was mathematically the easiest case (as said before, this paper is a feasibility study of our method). The j.p.d. $P(\varphi_{\mathbf{h}},\varphi_{\mathbf{k}},\varphi_{\mathbf{h}+\mathbf{k}}|R_{\mathbf{h}},R_{\mathbf{k}},R_{\mathbf{h}+\mathbf{k}})$ will (likely) give a more complex formula than the ('classical Von Mises') expression

$$\begin{aligned}&P\big(\varphi_{\mathbf{h}},\varphi_{\mathbf{k}},\varphi_{\mathbf{h}+\mathbf{k}}\big|R_{\mathbf{h}},R_{\mathbf{k}},R_{\mathbf{h}+\mathbf{k}}\big)_{\mathrm{cl}}\\ &\propto\exp\{R_{\mathbf{h}}R_{\mathbf{k}}R_{\mathbf{h}+\mathbf{k}}[\cos(\varphi)\Re\langle E_{\mathbf{h}}E_{\mathbf{k}}E_{-\mathbf{h}-\mathbf{k}}\rangle\\ &\quad+\sin(\varphi)\Im\langle E_{\mathbf{h}}E_{\mathbf{k}}E_{-\mathbf{h}-\mathbf{k}}\rangle]\}\end{aligned}$$

(in which case we would indeed be better off with calculating the normalization constant). Let us then close this section by stating the problem we really should solve. We must calculate the expression

$$P\big(\varphi_{\mathbf{h}},\varphi_{\mathbf{k}},\varphi_{\mathbf{h}+\mathbf{k}}\big|R_{\mathbf{h}},R_{\mathbf{k}},R_{\mathbf{h}+\mathbf{k}}\big)\propto\int\mathcal{D}\rho\,\mathrm{Prob}\,(\rho)F[\rho],$$

where

$$\begin{aligned}F[\rho] &= \delta\Big[\int\mathrm{d}\mathbf{x}\,\rho(\mathbf{x})\cos(2\pi\mathbf{h}\cdot\mathbf{x})-R_{\mathbf{h}}\cos\varphi_{\mathbf{h}}\Big]\\ &\quad\times\delta\Big[\int\mathrm{d}\mathbf{x}\,\rho(\mathbf{x})\sin(2\pi\mathbf{h}\cdot\mathbf{x})-R_{\mathbf{h}}\sin\varphi_{\mathbf{h}}\Big]\\ &\quad\times(\text{similar terms})\end{aligned}$$

Prob (ρ) = the right-hand side of equation (8). Guided by the method developed in this paper we shall do this in the future.

5. A comparison of $\langle E_{\mathbf{h}}E_{\mathbf{k}}E_{-\mathbf{h}-\mathbf{k}}\rangle$ with $E_{\mathbf{h}}E_{\mathbf{k}}E_{-\mathbf{h}-\mathbf{k}}$

Let us again rewrite the expression for $\langle E_{\mathbf{h}}E_{\mathbf{k}}E_{-\mathbf{h}-\mathbf{k}}\rangle/\langle E_{\mathbf{h}_0}E_{\mathbf{k}_0}E_{-\mathbf{h}_0-\mathbf{k}_0}\rangle$. We obtained [see equation (6)]

$$\frac{\langle E_{\mathbf{h}}E_{\mathbf{k}}E_{-\mathbf{h}-\mathbf{k}}\rangle}{\langle E_{\mathbf{h}_0}E_{\mathbf{k}_0}E_{-\mathbf{h}_0-\mathbf{k}_0}\rangle}=\frac{\mathbb{E}(\mathbf{h},\mathbf{k})}{\mathbb{E}(\mathbf{h}_0,\mathbf{k}_0)}$$

and

$$\begin{aligned}\mathbb{E}(\mathbf{h},\mathbf{k})\equiv\Bigg[&\frac{1}{(\hat{Q}_{\mathbf{h}}+i)(\hat{Q}_{\mathbf{h}+\mathbf{k}}+i)}+\frac{1}{(\hat{Q}_{\mathbf{k}}+i)(\hat{Q}_{\mathbf{h}+\mathbf{k}}+i)}\\ &+\frac{1}{(\hat{Q}_{\mathbf{h}}+i)(\hat{Q}_{\mathbf{k}}+i)}\Bigg],\end{aligned}$$

[1] Supporting information for this paper is available from the IUCr electronic archives (Reference: SC5080).

where $\hat{Q}_{\mathbf{h}} \equiv R_{\mathbf{h}}^2 - 1$. Now

$$\frac{1}{(\hat{Q}_{\mathbf{h}} + i)} = \frac{(\hat{Q}_{\mathbf{h}} - i)}{(\hat{Q}_{\mathbf{h}}^2 + 1)} = \frac{\hat{Q}_{\mathbf{h}}}{(\hat{Q}_{\mathbf{h}}^2 + 1)} - i\frac{1}{(\hat{Q}_{\mathbf{h}}^2 + 1)}.$$

We shall now test the *hypothesis* that

$$\frac{E_{\mathbf{h}}E_{\mathbf{k}}E_{-\mathbf{h}-\mathbf{k}}}{E_{\mathbf{h}_0}E_{\mathbf{k}_0}E_{-\mathbf{h}_0-\mathbf{k}_0}} \propto \frac{\langle E_{\mathbf{h}}E_{\mathbf{k}}E_{-\mathbf{h}-\mathbf{k}}\rangle}{\langle E_{\mathbf{h}_0}E_{\mathbf{k}_0}E_{-\mathbf{h}_0-\mathbf{k}_0}\rangle} = \frac{\mathbb{E}(\mathbf{h},\mathbf{k})}{\mathbb{E}(\mathbf{h}_0,\mathbf{k}_0)}; \quad (10)$$

thus

$$E_{\mathbf{h}}E_{\mathbf{k}}E_{-\mathbf{h}-\mathbf{k}} \propto \frac{\mathbb{E}(\mathbf{h},\mathbf{k})}{\mathbb{E}(\mathbf{h}_0,\mathbf{k}_0)} E_{\mathbf{h}_0}E_{\mathbf{k}_0}E_{-\mathbf{h}_0-\mathbf{k}_0}.$$

We shall take $\mathbf{h}_0$ and $\mathbf{k}_0$ to be fixed vectors, such that $|E_{\mathbf{h}_0}E_{\mathbf{k}_0}E_{-\mathbf{h}_0-\mathbf{k}_0}|$ is maximal in the domain $|h_1| \leq M'$, $|h_2| \leq M'$, $|h_3| \leq M'$ ($M' = 7$ being the number of reciprocal vectors along one axis). We will see (from numerical tests) that

$$E_{\mathbf{h}}E_{\mathbf{k}}E_{-\mathbf{h}-\mathbf{k}} = (X + iY)\frac{\mathbb{E}(\mathbf{h},\mathbf{k})}{\mathbb{E}(\mathbf{h}_0,\mathbf{k}_0)} E_{\mathbf{h}_0}E_{\mathbf{k}_0}E_{-\mathbf{h}_0-\mathbf{k}_0}, \quad (11)$$

where X and Y are real random variables, with low variance, over the space of all couples of reciprocal vectors $(\mathbf{h}, \mathbf{k})$ such that

$$R_{\mathbf{h}_0}R_{\mathbf{k}_0}R_{\mathbf{h}_0+\mathbf{k}_0} > R_{\mathbf{h}}R_{\mathbf{k}}R_{\mathbf{h}+\mathbf{k}} \geq 4$$

for an equal point atom structure of N atoms. We shall test this for N = 10 000, 1000 and 100. The model structures are numerical simulations with randomly placed atomic vectors in the unit cell ($P1$).

(1) N = 10 000.

$\mathbf{h}_0 = (1, -2, -4)$, $\mathbf{k}_0 = (-2, 6, 4)$, $\mathbb{E}(\mathbf{h}_0, \mathbf{k}_0) = 0.1831 - 0.1559i$.

$E_{\mathbf{h}_0}E_{\mathbf{k}_0}E_{-\mathbf{h}_0-\mathbf{k}_0} = -1.18 + 9.749i$ and $R_{\mathbf{h}_0}R_{\mathbf{k}_0}R_{\mathbf{h}_0+\mathbf{k}_0} = 9.82$.

$E(X) = -0.005873$, $\sigma^2(X) = 0.02685$; $E(Y) = 0.005557$, $\sigma^2(Y) = 0.02507$.

(2) N = 1000

$\mathbf{h}_0 = (1, 0, 0)$, $\mathbf{k}_0 = (-7, -6, 5)$, $\mathbb{E}(\mathbf{h}_0, \mathbf{k}_0) = 0.2741 - 0.6392i$.

$E_{\mathbf{h}_0}E_{\mathbf{k}_0}E_{-\mathbf{h}_0-\mathbf{k}_0} = 3.644 + 3.128i$ and $R_{\mathbf{h}_0}R_{\mathbf{k}_0}R_{\mathbf{h}_0+\mathbf{k}_0} = 4.802$.

$E(X) = -0.06513$, $\sigma^2(X) = 0.1052$; $E(Y) = 0.3244$, $\sigma^2(Y) = 0.2134$.

(3) N = 1000

$\mathbf{h}_0 = (0, -6, 5)$, $\mathbf{k}_0 = (-7, 6, 2)$, $\mathbb{E}(\mathbf{h}_0, \mathbf{k}_0) = 0.2783 - 0.2539i$.

$E_{\mathbf{h}_0}E_{\mathbf{k}_0}E_{-\mathbf{h}_0-\mathbf{k}_0} = 5.002 + 4.887i$ and $R_{\mathbf{h}_0}R_{\mathbf{k}_0}R_{\mathbf{h}_0+\mathbf{k}_0} = 6.994$.

$E(X) = 0.00315$, $\sigma^2(X) = 0.06493$; $E(Y) = -0.03868$, $\sigma^2(Y) = 0.0581$.

(4) N = 100

$\mathbf{h}_0 = (1, -3, -3)$, $\mathbf{k}_0 = (6, 3, 6)$, $\mathbb{E}(\mathbf{h}_0, \mathbf{k}_0) = 0.1433 - 0.07377i$.

$E_{\mathbf{h}_0}E_{\mathbf{k}_0}E_{-\mathbf{h}_0-\mathbf{k}_0} = 6.024 - 10.29i$ and $R_{\mathbf{h}_0}R_{\mathbf{k}_0}R_{\mathbf{h}_0+\mathbf{k}_0} = 11.93$.

$E(X) = -0.00392$, $\sigma^2(X) = 0.006904$; $E(Y) = 0.04863$, $\sigma^2(Y) = 0.006777$.

Let us now do the same thing for the classical formula, *e.g.* for the case of 1000 atoms. Proceeding in the same way as above we have

$$\frac{\langle E_{\mathbf{h}}E_{\mathbf{k}}E_{-\mathbf{h}-\mathbf{k}}\rangle_{\mathrm{cl}}}{\langle E_{\mathbf{h}_0}E_{\mathbf{k}_0}E_{-\mathbf{h}_0-\mathbf{k}_0}\rangle_{\mathrm{cl}}} \propto \frac{E_{\mathbf{h}}E_{\mathbf{k}}E_{-\mathbf{h}-\mathbf{k}}}{E_{\mathbf{h}_0}E_{\mathbf{k}_0}E_{-\mathbf{h}_0-\mathbf{k}_0}} = \frac{1/N^{1/2}}{1/N^{1/2}} = 1.$$

Then (N = 1000):

$\mathbf{h}_0 = (1, -4, 3)$, $\mathbf{k}_0 = (-7, 6, 2)$, $\mathbb{E}(\mathbf{h}_0, \mathbf{k}_0) = 0.2576 - 0.4712i$.

$E_{\mathbf{h}_0}E_{\mathbf{k}_0}E_{-\mathbf{h}_0-\mathbf{k}_0} = -1.378 + 5.562i$ and $R_{\mathbf{h}_0}R_{\mathbf{k}_0}R_{\mathbf{h}_0+\mathbf{k}_0} = 5.73$.

$E(X) = -0.0116$, $\sigma^2(X) = 0.1648$; $E(Y) = -0.167$, $\sigma^2(Y) = 0.1074$.

$E_{\mathrm{cl}}(X) = -0.009683$, $\sigma^2_{\mathrm{cl}}(X) = 0.307$; $E_{\mathrm{cl}}(Y) = -0.2353$, $\sigma^2_{\mathrm{cl}}(Y) = 0.219$.

As we can see the variances for the classical case are higher (and in many cases much higher) than in our case.

6. Conclusion, future work and a brief comparison with current methods

The numerical tests seem to indicate that

$$E_{\mathbf{h}}E_{\mathbf{k}}E_{-\mathbf{h}-\mathbf{k}} \simeq (x + iy)\frac{\mathbb{E}(\mathbf{h},\mathbf{k})}{\mathbb{E}(\mathbf{h}_0,\mathbf{k}_0)} E_{\mathbf{h}_0}E_{\mathbf{k}_0}E_{-\mathbf{h}_0-\mathbf{k}_0}.$$

However, this relation cannot be used for phase determination since the values of x, y and $E_{\mathbf{h}_0}E_{\mathbf{k}_0}E_{-\mathbf{h}_0-\mathbf{k}_0}$ are unknown: x and y do depend on the structure and on $\mathbf{h}_0$ and $\mathbf{k}_0$ [the variances $\sigma^2(X)$ and $\sigma^2(Y)$ should also (for phase determination) be a factor of 10 lower]. Future work should be the calculation of the joint *probability* density $P(\varphi_{\mathbf{h}}, \varphi_{\mathbf{k}}, \varphi_{\mathbf{h}+\mathbf{k}} | R_{\mathbf{h}}, R_{\mathbf{k}}, R_{\mathbf{h}+\mathbf{k}})$. This will reveal a better relation for phase determination. So we must wait for this density. We expect a more complicated probability distribution than the classical one,

$$P_{\mathrm{cl}}(\varphi_{\mathbf{h}}, \varphi_{\mathbf{k}}, \varphi_{\mathbf{h}+\mathbf{k}} | R_{\mathbf{h}}, R_{\mathbf{k}}, R_{\mathbf{h}+\mathbf{k}}) \propto \exp\left[\frac{2R_{\mathbf{h}}R_{\mathbf{k}}R_{\mathbf{h}+\mathbf{k}}}{N^{1/2}}\cos(\varphi_{\mathbf{h}} + \varphi_{\mathbf{k}} - \varphi_{\mathbf{h}+\mathbf{k}})\right],$$

and perhaps independent of N. (This will give a formula that *can* be used for phase determination.) Then three additional integrations (besides the ones in this paper) have to be performed.

What can also be done is to use *only* the Patterson constraint

$$\int \mathrm{d}\mathbf{y}\, \rho(\mathbf{y})\rho(\mathbf{x} + \mathbf{y}) = P(\mathbf{x})$$

and not the two other additional constraints $\rho^2(\mathbf{x}) = [\delta(\mathbf{0})/N^{1/2}]\rho(\mathbf{x})$ and $\int \mathrm{d}\mathbf{x}\, \rho(\mathbf{x}) = N^{1/2}$. This has the advantage that we will obtain a formula independent of N and valid for all structures [*e.g.* structures with (a lot of) heavy atoms]: one will only have to feed the diffraction data [without additional chemical information (knowing *e.g.* N)] directly into the probability formula. This case is much easier to calculate (only two *functional* integrations must be done). We shall do this easier case in the near future.

Another interesting case is the use of Patterson vectors $\mathbf{u}$ [that is those for which $P(\mathbf{u})$ is a peak]. Such information can be obtained easily. We must then include additional constraints of the form

$$\int \mathrm{d}\mathbf{y}\, \rho(\mathbf{y})\rho(\mathbf{u}+\mathbf{y}) = P(\mathbf{u}).$$

The advantage of these constraints is that there is but a *single* integral ($\int \mathrm{d}\mathbf{y}$).

Another possibility is the use of more extended chemical information: suppose a large part of the structure is known. This information might be translated in a known function $f(\mathbf{x})$, that will be a sum of peaks of the known positions of the atoms and then one has to use a constraint of the form

$$f(\mathbf{x})\rho(\mathbf{x}) \propto f(\mathbf{x}).$$

And we observe that this constraint is only *linear* in ρ. For a classical approach of using model structures in DM we refer to recent work (Burla *et al.*, 2012).

6.1. Brief overview of current methods

(i) All DM formulas use the Patterson indirectly through the use of *conditional* probability structures (probability of the phase invariant *given* the *actual* values of the moduli of structure factors). Well known is the triplet formula $P(\varphi|R_{\mathbf{h}}, R_{\mathbf{k}}, R_{\mathbf{h}+\mathbf{k}}) \propto \exp(2R_{\mathbf{h}}R_{\mathbf{k}}R_{\mathbf{h}+\mathbf{k}}\cos\varphi/N^{1/2})$. The most probable value for φ is here 0. This formula is, however, not able to predict a negative cosine. There is no neighbourhood (Hauptman) [representation (Giacovazzo)] in the $1/N$ 'range'. In the quintet range ($1/NN^{1/2}$) we like to mention the stronger *P*13 formula (Burla *et al.*, 1994) that is able to predict negative triplets.

Another well known formula is the quartet invariant with first and second neighbourhoods (representations) in the $1/N$ range. Using this second neighbourhood, one can predict *negative* quartets although with small probabilities [for high N see Peschar & Schenk (1987)] and more recently using higher representations a stronger formula (Altomare *et al.*, 1995). However, all these formulas will eventually fail for very high N.

(ii) Direct space methods (with possibly the help of DM).

(*a*) Shake and bake [using repeatedly the tangent formula (= derived *algebraically* not probabilistically), DM (triplet and quartet formulas) and direct space (using fast Fourier transform and peak selection)] (Chang *et al.*, 1997; Langs & Hauptman, 2011), which is able to solve *ab initio* structures for $N \simeq 1000$.

A *probabilistic* formula for a correct tangent formula will be submitted by us very soon to *Acta Crystallographica Section A*; it will use the probabilistic approach explained in Brosius (1979) and it will not be adversely dependent on N. It is based on the use of the *a priori* distribution $p(\mathbf{x}, \{\varphi_{\mathbf{q}}\}_{\mathbf{q}}) \propto [\sum_{\mathbf{q}} R_{\mathbf{q}} \exp(i\varphi_{\mathbf{q}} - 2\pi i\mathbf{q}\cdot\mathbf{x})]^2$ and is mathematically rigorous.

(*b*) *SIR2011* (using more a direct space approach than a DM one) is able to solve *ab initio* structures with N around 1000 (Burla *et al.*, 2013).

(*c*) Patterson deconvolution (superposition) (Caliandro *et al.*, 2013) (direct space).

A rigorous *probabilistic* treatment based on an *a priori* Patterson superposition distribution (Brosius, 1979) can be given that does not depend adversely on N.

I would especially like to thank the referees for their careful reading of the manuscript and their comments. I am also grateful to Professor H. Schenk for all his work and patience.

References

Altomare, A., Burla, M. C., Cascarano, G., Giacovazzo, C., Guagliardi, A., Moliterni, A. G. G. & Polidori, G. (1995). *Acta Cryst.* A**51**, 305–309.
Brosius, J. (1979). PhD thesis, KULeuven, Leuven, Belgium.
Brosius, J. (2008*a*). *Acta Cryst.* A**64**, 564–570.
Brosius, J. (2008*b*). *Acta Cryst.* A**64**, 571–586.
Brosius, J. (2008*c*). *Acta Cryst.* A**64**, 560–563.
Brosius, J. (2012). *Z. Kristallogr.* **4**, 190–198.
Burla, M. C., Carrozzini, B., Cascarano, G. L., Comunale, G., Giacovazzo, C., Mazzone, A. & Polidori, G. (2012). *Acta Cryst.* A**68**, 513–520.
Burla, M. C., Giacovazzo, C., Moliterni, A. G. G. & Gonzalez Platas, J. (1994). *Acta Cryst.* A**50**, 771–778.
Burla, M. C., Giacovazzo, C. & Polidori, G. (2013). *J. Appl. Cryst.* **46**, 1592–1602.
Caliandro, R., Carrozzini, B., Cascarano, G. L., Comunale, G. & Giacovazzo, C. (2013). *Acta Cryst.* A**69**, 98–107.
Chaichian, M. & Demichev, A. (2001). *Mathematical and Computational Physics. Path Integrals in Physics*, Vol. 2. Bristol, Philadelphia: Institute of Physics Publishing.
Chang, C.-S., Weeks, C. M., Miller, R. & Hauptman, H. A. (1997). *Acta Cryst.* A**53**, 436–444.
Diener, F. & Reeb, G. (1989). *Analyse Non Standard.* Enseignement des Sciences. Paris: Hermann.
Giacovazzo, C. (1975). *Acta Cryst.* A**31**, 602–609.
Heinerman, J. J. L. (1975). *Acta Cryst.* A**31**, 727–730.
Karle, J. & Hauptman, H. (1953). *Acta Cryst.* **6**, 131–135.
Klug, A. (1958). *Acta Cryst.* **11**, 515–543.
Langs, D. A. & Hauptman, H. A. (2011). *Acta Cryst.* A**67**, 396–401.
Masujima, M. (2009). *Path Integral Quantization and Stochastic Quantization*. *Springer Tracts in Modern Physics*, Vol. 165. Berlin, Heidelberg: Springer.
Nelson, E. (1977). *Bull. Am. Math. Soc.* **83**, 1165–1198.
Nelson, E. (1987). *Radically Elementary Probability Theory.* Princeton University Press.
Peschar, R. & Schenk, H. (1987). *Acta Cryst.* A**43**, 84–92.
Siegel, W. (2005). *Fields.* arXiv:hep-th/9912205v3.
Weinberg, S. (2005*a*). *The Quantum Theory of Fields*, Vol. 1. Cambridge University Press.
Weinberg, S. (2005*b*). *The Quantum Theory of Fields*, Vol. 2. Cambridge University Press.
Xu, H. & Hauptman, H. A. (2004). *Acta Cryst.* A**60**, 153–157.