Structure factor Statistics and density

Garib N Murshudov

MRC-LMB, Cambridge, UK

Contents

Introduction: atoms, structure factors and central limit theorem

Distribution of structure factors and their amplitudes

Effect of twinning

Moments of intensities and systematic errors

Density correlation: effect of phases and amplitudes

Mean and variance algebra

If we have two random variables X1 and X2 then for mean and variance of the sum of these random variables the following is true:

$<X_1 + X_2> = <X_1> + <X_2>$

$var(X_1+X_2) = var(X_1) + 2\ cov(X_1,X_2) + var(X_2)$

$<c\ X_1> = c <X_1>$

$var(c\ X_1) = c^2\ var(X_1)$

If random variables are independent then $cov(X_1,X_2) = 0$. I.e. for independent variables mean values and variances are summed

Central limit theorem

If there are n independent random variables $\{X_i\}$ with mean m_i and standard deviations σ_i and if mean values and standard deviation are finite then as n increases

$$Y = X_1+X_2+....X_n$$

Tends to become random variable with Gaussian distribution with the mean

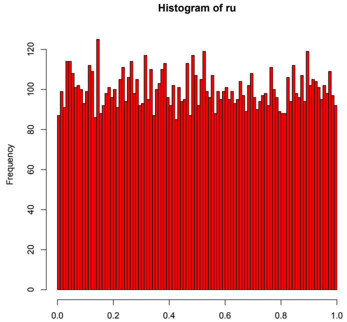
$$m_Y=m_1+m_2+....m_n$$

And with the variance

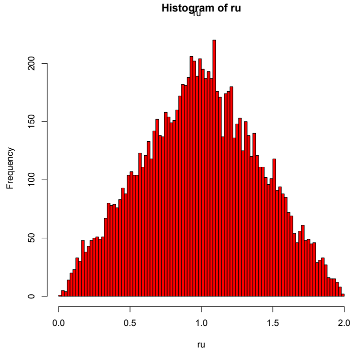
$$\sigma_Y^2 = \sigma_1^2 + \sigma_2^2 + \ldots + \sigma_n^2$$

An example

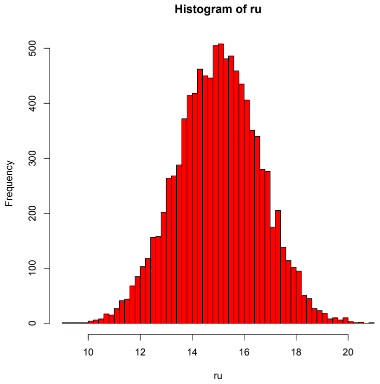
A random variable r_i has uniform distribution between 0 and 1. Its histogram (we generate 10000 random variables using uniform random number generator). Mean value = 0.5, and variance =1/12



When we add two uniform random variables. The sum already becomes unimodal, symmetric and starts to resemble Gaussian distribution



After adding 30 uniform random variables together. It is already can be approximated with Gaussian distribution very well. Mean value of sum of 30 uniform random variables is 30*0.5 = 15 and variance is 30/12 = 2.5



Distribution of structure factors

In a crystal we have set of atoms with parameters $\{c_i, x_i, B_i\}$ – occupancies, positions and B values. For simplicity we consider only positions as random variables. We also assume that occupancies are equal to 1. Structure factor equations (written for real and imaginary parts):

$$A = \sum c_i f_i(s) e^{-B|s|^4/4} \cos(2\pi s x_i), \quad B = \sum c_i f_i(s) e^{-B|s|^4/4} \sin(2\pi s x_i)$$

$$A = \sum a_i \qquad B = \sum b_i$$

Distribution of real and imaginary parts of structure factors approach to normal distribution with mean

$<A> = \Sigma <a_i>, \quad <B>=\Sigma<b_i>$

And variances:

$Var(A) = \Sigma var(a_i), \quad var(B) = \Sigma var(b_i)$

It is safe to assume that real and imaginary parts are independent

Distribution of structure factors

Assume that we have an atomic model with coordinates $\{x_{i,c}\}$. Model has some errors. Then

$<a_i> = D <a_{i,c}>$, $var(a_i) = ½ (1-D^2) f^2{}_i(s)$

Where D reflects errors in the position. If we have no information about coordinates then D=0. The distribution of the real and imaginary parts of structure factors is normal distribution with

$<A> = 0$, $var(A) = ½ \Sigma f^2{}_i(s)$

Confusingly $\Sigma f^2{}_i(s)$ is denoted as Σ. So the distributions of A and B are $N_1(0, \Sigma/2)$ and therefore joint probability distribution of A and B is $N_2(0, \Sigma/2\ \mathbf{I})$. Where **I** is an identity matrix. We usually say that the distribution of structure factors – **F** is two-dimensional normal distribution.

When we have some atomic models then the distribution of F becomes two dimensional normal distribution with mean DF_c and variance $(1-D^2)\Sigma_k/2 + \Sigma_u)\mathbf{I}$

Distribution of intensities

In crystallography we observe intensities of structure factors – $I = A^2+B^2$.

Now we use the following fact:

If $\{X_i\}$ are standard normal random variable (with mean 0 and variance 1) then sum (X_i) is χ^2_n distribution with degrees of freedom n, where n is the number of random variables.

So $2I/\Sigma = 2(A^2+B^2)/\Sigma$ will have χ^2_2 with degrees of freedom 2.

Some properties of intensities:

$$<I> = \Sigma, <I^2> = 2\Sigma, var(I) = \Sigma$$

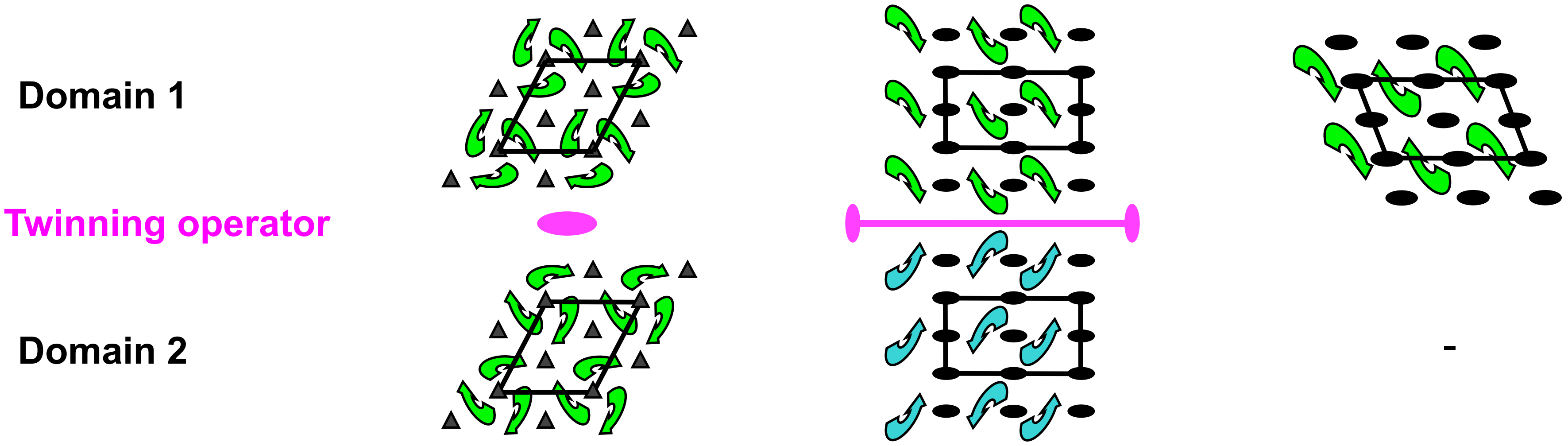
It means that

$<I^2>/<I>^2 = 2, var(I)/<I>^2 = 1$

Application: perfect twin

merohedral and pseudo-merohedral twinning

Crystal symmetry:	P3	P2	P2
Constrain:	-	β = 90º	-
Lattice symmetry *: (rotations only)	P622	P222	P2
Possible twinning:	merohedral	pseudo-merohedral	-



Crystal lattice is invariant with respect to twinning operator.

The crystal is NOT invariant with respect to twinning operator.

Effect on intensity statistics: intuitive approach

Take a simple case. We have two intensities: weak and strong. When we sum them we will have four options w+w, w+s, s+w, s+s. So we will have one weak, two medium and one strong reflection.

As a results of twinning, proportion of weak and strong reflections becomes small and the number of medium reflections increases. It has effect on intensity statistics

In probabilistic terms: without twinning, the distribution of intensities is χ^2 with degree of freedom 2 and after perfect twinning degree of freedom increases and becomes 4. χ^2 distributions with higher degree of freedom behave like normal distribution

Intensity statistics: twin

If there are n copies of crystal with equal size domains then the observed intensities are average of intensities from all these domains:

$$I_T = \frac{1}{n}\sum I_i$$

The distribution of $2n\, I_T/\Sigma$ *is* χ^2_{2n}, i.e. chi-squared distribution with degrees of freedom 2n.

Properties

$$\langle I\rangle = \Sigma,\ \langle I^2\rangle = (n+1)/n\ \Sigma,\ var(I) = 1/n\ \Sigma$$

$$\langle I^2\rangle/\langle I\rangle^2 = (n+1)/n,\ var(I)/\langle I\rangle^2 = 1/n$$

I.e. fourth moment converges to 1 and variance of intensities converges to 0 as n increases.

X^2 distributions

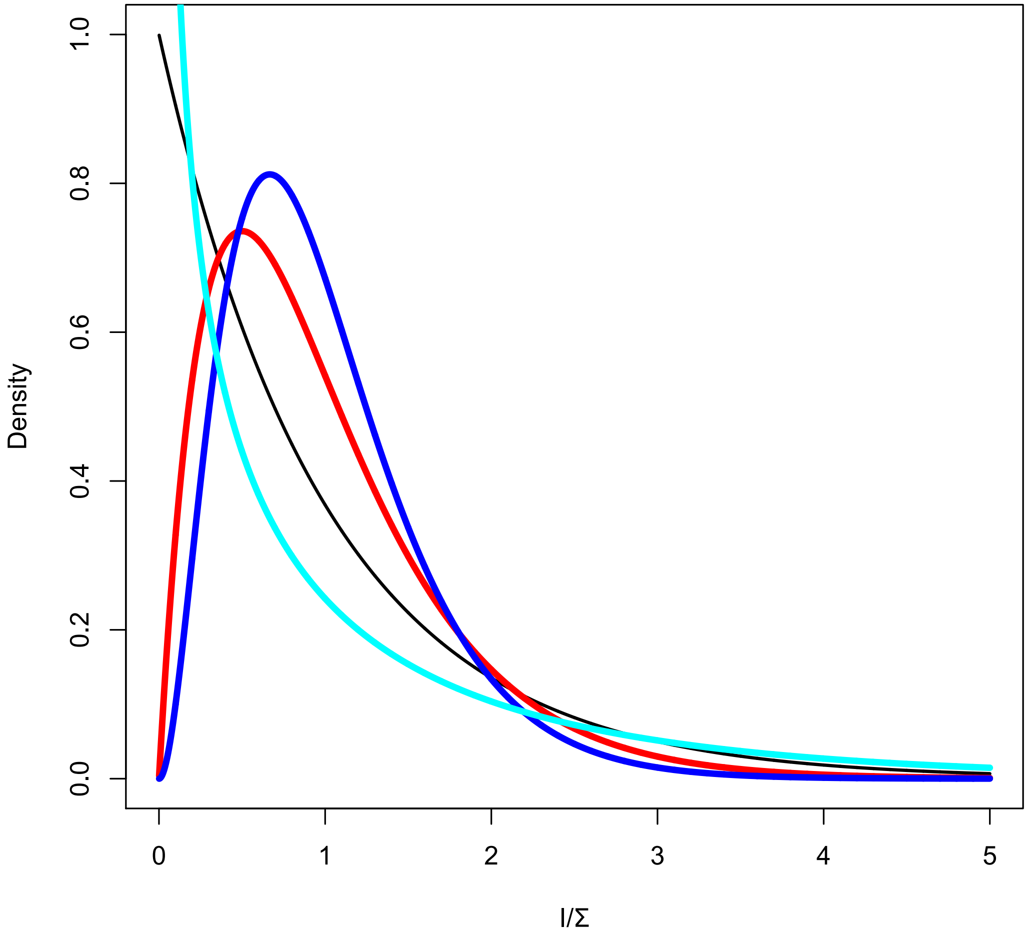
Densities of distributions for intensities: χ^2 distributions (with suitable change of variables).

Cyan – centric

Black – acentric

Red - acentric two twin domains

Blue – acentric 3 twin domains



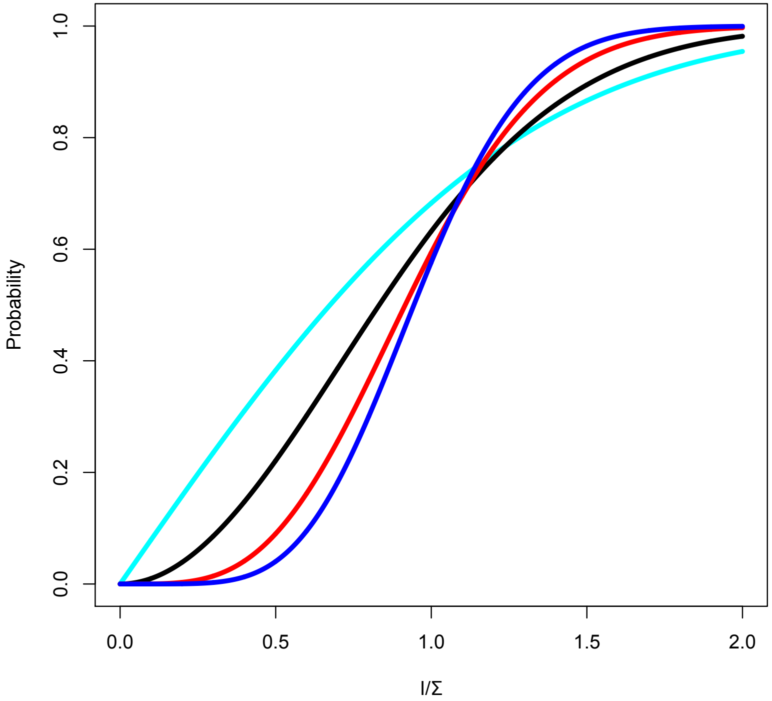
Cumulative distribution for $I^{1/2}$ with suitable change of variables (N(z) plots)

Cyan – centric

Black – acentric

Read – acentric with two twin domains

Blue – acentric with 3 twin domains



Structure factor statistics and twin

When we move from centric to acentric reflection variance of intensities is reduced. When there are twin then variance again is reduced. It means that Rmerge will be smaller if we are dealing with twins.

$$R_{centric} > R_{acentric} > R_{twin2} > \ldots$$

Such variance reduction also means that R factors for data from twinned crystals will be smaller than those from untwinned. Again same order as in R merge is obeyed.

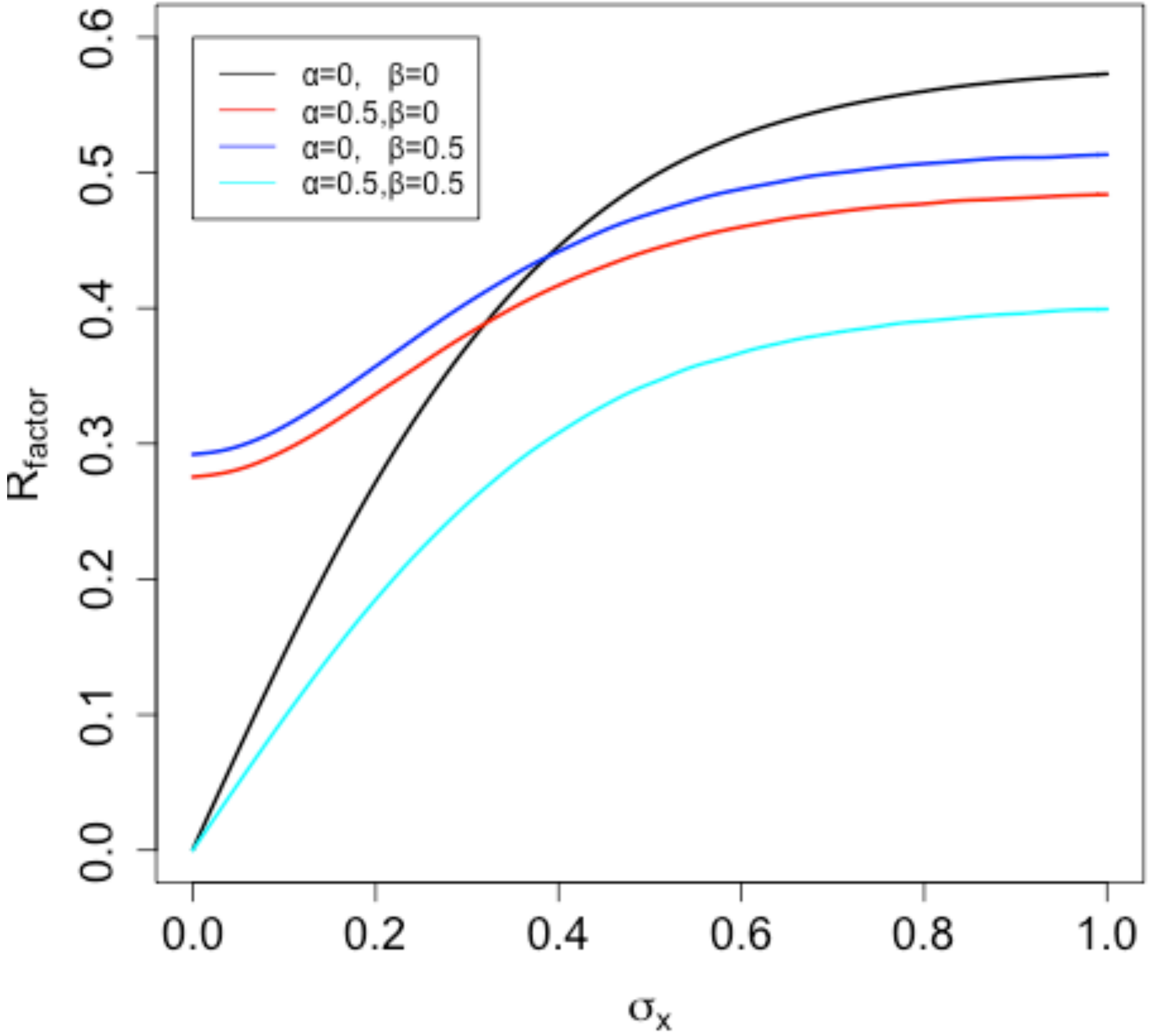
Twin: R values

Rvalues for random structures (no other peculiarities)

Twin	Modeled	Not modeled
Yes	0.41	0.49
No	0.52	0.58

Murshudov GN “Some properties of Crystallographic Reliability index – Rfactor: Effect of Twinning” Applied and Computational Mathematics”, 2011:10;250-261

Rvalue for structures with different model errors: Combination of real and modeled perfect twin fractions



Conclusions I

Distribution of intensities is related to chi-squared distribution

Distribution of signal in the data affect R-factors and other statistics

R factors for twinned crystals tend to be smaller however information content is also smaller

Structure factor statistics: effect of errors in the data

Let us assume that we observe data and noise in the data is additive

$$I_o = I + n$$

Usually one of the diagnostic techniques used is so-called fourth moment plots.

$<I_o^2>/<I_o>^2$ is calculated and plotted vs resolution. Let us see how this statistics would behave under different circumstances. If we consider perfect twin with k domains and acentric case (recall that: var(I) = <I>/k))

$$\frac{<I_o^2>}{<I_o>^2} = 1 + \frac{\frac{1}{k} + +2\frac{\mathrm{cov}(I,n)}{<I>^2} + \frac{\mathrm{var}(n)}{<I>^2}}{(1+\frac{<n>}{<I>})^2}$$

If covariance between noise and signal is ignored then:

$$\frac{<I_o^2>}{<I_o>^2} = 1 + \frac{\frac{1}{k} + \frac{\mathrm{var}(n)}{<I>^2}}{(1+\frac{<n>}{<I>})^2}$$

<n> is average noise. Usually we assume that average noise is 0.

Structure factor statistics: effect of errors in the data

Average noise is one of the components of systematics noises. There may be various reasons for this: 1) backgrounds may not be estimated accurately, for example when ice rings are present; 2) diffuse scattering may have some effect; 3) there may be more than one crystal, for example split crystals.

If crystal is not twinned and the variance of the noise is very small then:

$$\frac{<I_o^2>}{<I_o>^2} = 1 + \frac{1}{(1+\frac{<n>}{<I>})^2}$$

If <n>/<I> is sufficiently large then this ratio would come close to 1. The same phenomenon happens if the number of twin domains is large. When there are k twin domains, average noise is 0 and variance of noise is very small:

$$\frac{<I_o^2>}{<I_o>^2} = \frac{k+1}{k}$$

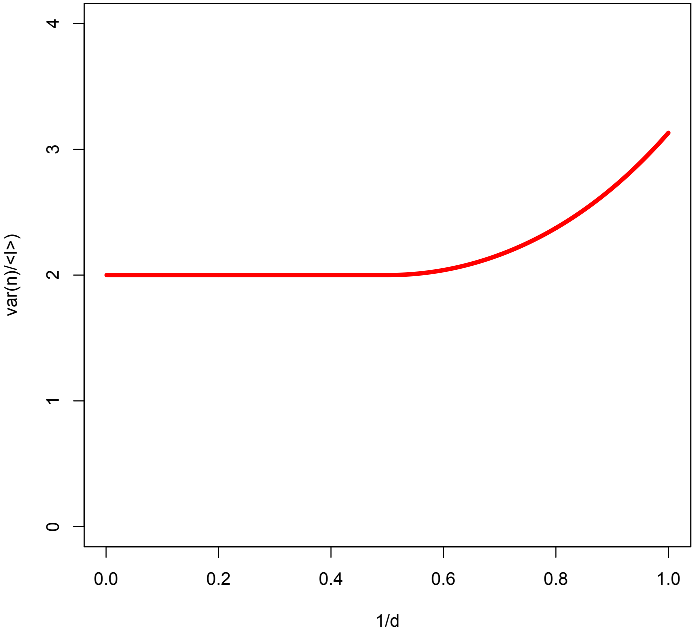
When k is large then this ratio again becomes close to 1. I.e. if average noise is not 0 then crystals may be interpreted to be twinned and it may affect wrong strategy to be used in structure solution and refinement

Structure factor statistics: effect of errors in the data

Finally if average noise is 0 then

$$\frac{< I_o^2 >}{< I_o >^2} = \frac{k+1}{k} + \frac{\mathrm{var}(n)}{< I >^2}$$

i.e. fourth moment would start from a constant value and ten as noise level increases it would go up. This plot should look like as in this figure. In very good data this ratio stays more or less constant.



Conclusions II

Noise structure of the data should be analyzed carefully

Systematic noise can obscure interpretation of data

If systematic noise is detected then data may need to be reintegrated

Importance of phases and amplitudes

The distribution of intensities of structure factors is related to chi-squared distribution.

In single crystal acentric case degrees of freedom is 2.

Properties of the distribution:

$$< I >= \Sigma, < I^2 >= 2\Sigma^2, \mathrm{var}(I) = \Sigma^2$$

And for amplitudes:

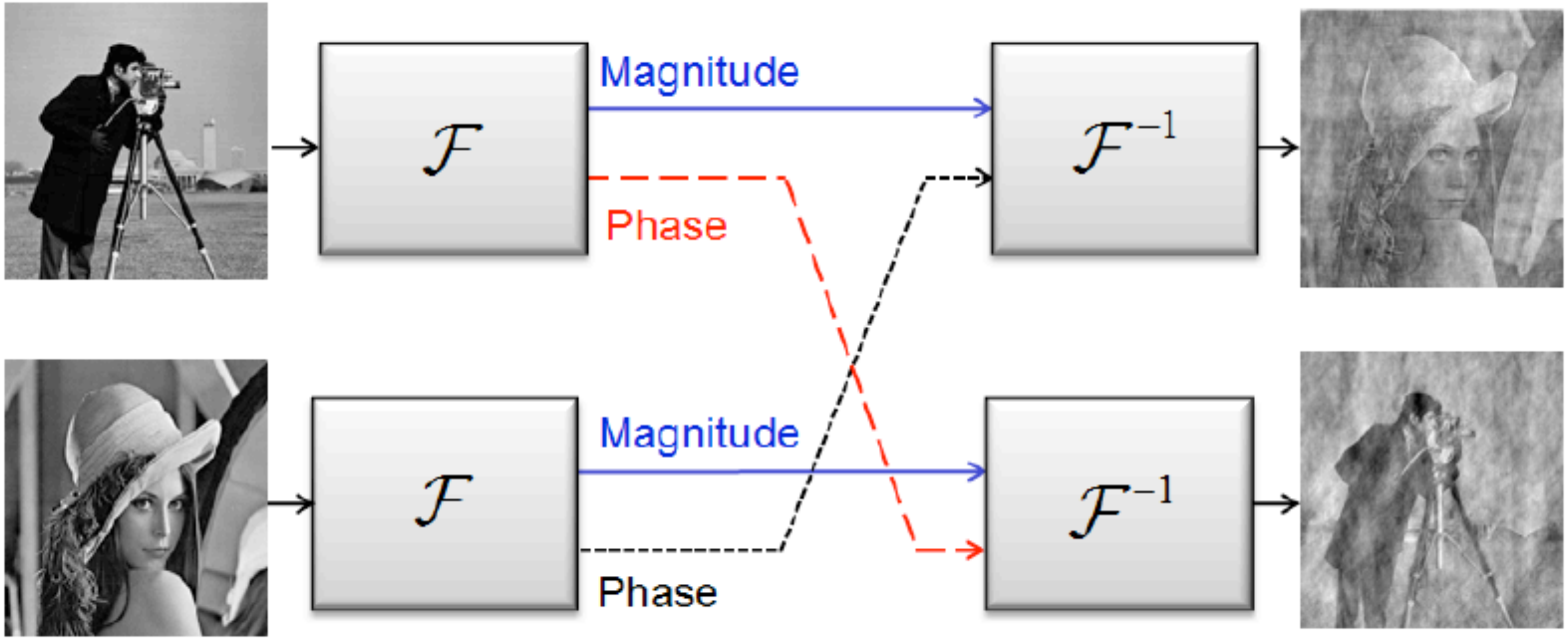
$$<| F |>= \frac{\sqrt{\pi}}{2}\sqrt{\Sigma} \approx 0.886\Sigma,\ <|\mathrm{F}|^2 >= \Sigma,\ \mathrm{var}(|\mathrm{F}|)= \frac{4-\pi}{4}\Sigma \approx 0.215\Sigma$$

We can use these properties to calculate various statistics. For example what would be R factor if instead of observed amplitudes we use expected values of amplitudes? It turns out to be around 0.42.

Importance of phases

It is well known that phases are very important. It is one of the examples demonstrating it. It may be not very good example but nevertheless …

Can we show in terms of numbers the meaning of this statement?



Importance of phases

Correlation between two electron density:

$$cor(\rho_1,\rho_2)=\frac{<\rho_1\rho_2>-<\rho_1><\rho_2>}{\sqrt{\mathrm{var}(\rho_1)\,\mathrm{var}(\rho_2)}}=\frac{<F_1F_2>-<F_1><F_2>}{\sqrt{\mathrm{var}(F_1)\,\mathrm{var}(F_2)}}$$

Correlation can be calculated in real or Fourier space. Advantage of Fourier space is that we can calculate it in Fourier shells or resolution bins. <F> = 0 in all resolution bins apart from that containing the origin. We almost never need the origin.

Using in Fourier space means that we can analyse the quality of maps depending on resolution (or frequency). We also will use this fact:

$$var(F) = <|F|^2>$$

So correlation in resolution bin for Fourier coefficients (complex coefficients) is

$$cor(F_1,F_2)=\frac{<F_1F_2>}{\sqrt{\mathrm{var}(F_1)\,\mathrm{var}(F_2)}}=\frac{<|F_1||F_2|\cos(\varphi_1-\varphi_2)>}{<|F_1|^2><|F_2|^2>}$$

Importance of phases

We can write correlation in terms of E values also. E values are defined as:

$$E = \frac{F}{\sqrt{<|F|^2>}}$$

So correlation in Fourier shells is

$$cor(F_1, F_2) =< E_1 E_2 >=<| E_1 || E_2 | \cos(\varphi_1 - \varphi_2) >$$

Now let us consider several cases:

1) Phases are perfect, amplitudes are perfect then cor=1.0
2) Phases are random, amplitudes are perfect then cor=0

Importance of phases

3) Phases are perfect, amplitudes are random (i.e. not related to the structure we are solving)

$$cor(F_1,F_2) = <|E_1||E_2|> = <|E_1|><|E_2|>$$

Now if we use the properties of the distribution of |F| then we get:

$$<|E_1|> = <|E_1|> = \frac{\sqrt{\pi}}{2}$$

i.e. correlation becomes:

$$cor(F_1,F_2) = <|E_1|><|E_2|> = \frac{\pi}{4} \approx 0.785$$

If phases are perfect then correlation between true and observed electron densities with random amplitudes can be as high as 0.785.

If we do know that amplitudes are random then we can replace them with constant value equal to the the expected value of the amplitudes on this resolution bin. E value of a constant equal to 1. So in this case correlation will be:

$$cor(F_1,F_2) = <|E_1|> = \frac{\sqrt{\pi}}{2} \approx 0.886$$

One obvious conclusion is that if we have phase information then unobserved E values can be replaced with the expected value. It is how “free-lunch” algorithm works

Importance of phases

We can ask another question: what should be the quality of the data so that adding them would give higher correlation than the use of constant values for them. (Under assumption that phases are perfect or there is some phase information and that is independent from amplitudes (????????))

$$<|E_1||E_2|> \quad \geq \frac{\sqrt{\pi}}{2} \quad \Rightarrow \quad \mathrm{cor}(|F_1|,|F_2|) \quad \geq \frac{2\sqrt{\pi}-\pi}{4-\pi} \approx 0.47$$

Correlation between "true" and "observed" amplitudes must be at least 0.47 so that the data gives some information about the model. Otherwise using expected values of amplitudes would give better correlation than using "observed" data.

One should be careful in interpretation of this statement: 1) we observe intensities; 2) errors in intensities have very different distribution that errors in amplitudes; 3) if many assumptions would be obeyed then we can go as low as half data correlation between amplitudes around 0.124

Conclusion III

Phases are important (obviously). Map correlation with perfect phase and random amplitude can go up to 0.78.

Replacing random amplitudes by the expected value may improve map (correlation can go up to 0.88)

Low quality data might be useful (half data correlation for amplitudes can go down to 0.124)

However bias problem needs to be addressed.

Literature

Wilson (1949) The probability distribution of X-ray intensities, Acta Cryst, 2, 318-321

Luzzatti Traitement statistique des erreurs dans la determination des structures cristallines, Acta Cryst, 5, 802-810

Srinavasan and Parthasarathy (1976) Some Statistical Applications in X-ray Crystallography