Tutorial on Structure Factor Statistics and Likelihood

Randy J Read

Background Information

Expected value = probability-weighted average: $\langle x\rangle = \int_{-\infty}^{\infty} p(x)\,x\,dx$

This is an integral, so properties of expected values come from properties of integrals.

Expected value of a sum of random variables is the sum of the expected values: $\langle x+y\rangle = \langle x\rangle + \langle y\rangle$

Expected value of a product *of* ***independent*** *random variables* is the product of the expected values: $\langle xy\rangle = \langle x\rangle\langle y\rangle$

Optional exercise: demonstrate this, starting from the knowledge that $p(x,y) = p(x)\,p(y)$ when x and y are independent.

Central Limit Theorem and the normal (Gaussian) distribution

Expected value of sum of independent random variables is the sum of their expected values. Variance of the sum is the sum of the variances. If there is a sufficient number of variables in the sum and none of them dominate the distribution (*i.e.* much bigger variance than everything else), then the sum tends towards a normal (Gaussian) probability distribution, *regardless of the form of the probability distributions of the individual variables*.

$$S = \sum_j x_j$$

$$\mu = \langle S\rangle = \sum_j \langle x_j\rangle$$

$$\sigma^2 = \left\langle \left(S - \langle S\rangle\right)^2 \right\rangle = \sum_j \left\langle \left(x_j - \langle x_j\rangle\right)^2 \right\rangle$$

$$p(S) \approx \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(S-\mu)^2}{2\sigma^2}\right) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2}(S-\mu)\left(\sigma^2\right)^{-1}(S-\mu)\right)$$

Optional exercise: show that $\left\langle \left(x_j - \langle x_j\rangle\right)^2 \right\rangle = \langle x_j^2\rangle - \langle x_j\rangle^2$

Complex normal distribution

This is a joint distribution of independent real and imaginary parts of a complex number, each with the same variance, so it looks like the product of two similar normal distributions, each with half of the total variance in the complex normal distribution.

$$\mathbf{z} = x + i\,y$$

$$\sigma^2(\mathbf{z}) = \left\langle \left(\mathbf{z} - \langle\mathbf{z}\rangle\right)\left(\mathbf{z} - \langle\mathbf{z}\rangle\right)^* \right\rangle = \sigma^2(x) + \sigma^2(y),$$ *(prove this!)*

$$p(\mathbf{z}) = \frac{1}{\pi\sigma^2} \exp\left(-\frac{\left(\left|\mathbf{z} - \langle\mathbf{z}\rangle\right|\right)^2}{\sigma^2}\right) = \frac{1}{\pi\sigma^2} \exp\left(-\left(\mathbf{z} - \langle\mathbf{z}\rangle\right)^* \left(\sigma^2\right)^{-1} \left(\mathbf{z} - \langle\mathbf{z}\rangle\right)\right)$$

Multivariate normal distribution

Variance is replaced by a variance-covariance matrix. Inverse variance-weighted deviation squared is replaced by triple product. Normalisation factor is replaced by a determinant.

$$\mathrm{p}(\mathbf{x}) = \frac{1}{|2\pi\Sigma|^{1/2}} \exp\left[-\tfrac{1}{2}(\mathbf{x}-\langle\mathbf{x}\rangle)^T \Sigma^{-1}(\mathbf{x}-\langle\mathbf{x}\rangle)\right]$$, where

elements of Σ given by $\sigma_{ij} = \left\langle (x_i - \langle x_i \rangle)(x_j - \langle x_j \rangle) \right\rangle$

Multivariate complex normal distribution

Like multivariate normal distribution for real numbers, but covariance matrix is Hermitian (it is equal to the complex conjugate of its own transpose, *i.e.* the *ij* term is the complex conjugate of the *ji* term) and transpose in triple product becomes a Hermitian transpose.

$$\mathrm{p}(\mathbf{z}) = \frac{1}{|\pi\Sigma|} \exp\left[-(\mathbf{z}-\langle\mathbf{z}\rangle)^H \Sigma^{-1}(\mathbf{z}-\langle\mathbf{z}\rangle)\right]$$, where

elements of Σ given by $\sigma_{ij} = \left\langle (\mathbf{z}_i - \langle \mathbf{z}_i \rangle)(\mathbf{z}_j - \langle \mathbf{z}_j \rangle)^* \right\rangle$

Variance of sum of correlated variables

Variance of sum of correlated random variables is the sum of the elements of a covariance matrix relating those variables, demonstrated as follows with variables that, for simplicity, are assumed to have an expected value of zero:

$$\begin{aligned}
\left\langle \left(\sum_j x_j \right)^2 \right\rangle &= \left\langle (x_1 + x_2 + \cdots + x_n)(x_1 + x_2 + \cdots + x_n) \right\rangle \\
&= \left\langle x_1^2 + 2x_1x_2 + x_2^2 + \cdots \right\rangle = \left\langle x_1^2 \right\rangle + 2\left\langle x_1x_2 \right\rangle + \left\langle x_2^2 \right\rangle + \cdots \\
&= \sum_j \left(\left\langle x_j^2 \right\rangle + 2\sum_{k \neq j} \left\langle x_j x_k \right\rangle \right)
\end{aligned}$$

Structure factor distributions

Wilson distribution

Assume that atoms are placed independently (*e.g.* one copy of non-symmetrical molecule in space group P1), then variance of structure factor is sum of variances of individual atomic contributions.

Exercise 1:

Consider a pair of atoms related by a crystallographic symmetry operation. Specifically, consider the pair related by a 2_1 screw axis along the *y* direction: $(x_2, y_2, z_2) = (-x_1, \tfrac{1}{2} + y_1, -z_1)$. The contributions of these atoms to the total structure factor are given by:

$f_1 \exp(2\pi i (hx_1 + ky_1 + lz_1))$ and

$$f_1 \exp\left(2\pi i \left(-hx_1 + k\left(\frac{1}{2} + y_1\right) - lz_1\right)\right)$$

For a general *hkl*, is there a predictable relationship between these two contributions for an atom at a random *xyz* position and its symmetry mate? What about for Miller indices *0k0*?

What are the relative expected intensities of reflections with: a) general indices *hkl*; b) indices *0k0* where *k* is odd; c) indices *0k0* where *k* is even?

Note that the reasoning about these expected intensity factors is similar to the reasoning about the effect of translational non-crystallographic symmetry.

Effect of coordinate errors on structure factor distribution

Contributions to the calculated structure factor from atoms with coordinate errors are correlated to the contributions from the atoms in the correct positions, but because of averaging of the complex contribution around the phase circle, the expected value of the contribution is reduced in magnitude. (This is analogous to the figure of merit for experimental phasing.)

Exercise 2

When we say that d_j is a real number in $\left\langle \mathbf{f}_j \exp\left(2\pi i \mathbf{h}\cdot\delta_j\right)\right\rangle = d_j\,\mathbf{f}_j$, what are we assuming about the probability distribution of the coordinate error δ_j?

Exercise 3

Show that $\left\langle \left|\mathbf{f}_j \exp\left(2\pi i \mathbf{h}\cdot\delta_j\right) - d_j\,\mathbf{f}_j\right|^2 \right\rangle = f_j^2\left(1-d_j^2\right)$

Effect of anomalous scatterers on covariances

Exercise 4 (advanced)

Work out the complex covariance between two Bijvoet-pair related reflections: $\left\langle \mathbf{F}^{+}\mathbf{F}^{-}\right\rangle$. (Actually, we use the complex conjugate of the minus hand, because that is more correlated with the plus hand, but then the complex conjugates cancel out in the complex covariance expression.) Note that the scattering factor for an anomalous scatterer is a complex number: $\mathbf{f} = f_0 + f' + i\,f''$. You should find that the complex covariance is also a complex number. Why is this different from the (real) covariance between the true and calculated structure factors: $\left\langle \mathbf{F}\,\mathbf{F}_C^{*}\right\rangle$? What does it imply about the relationship between the phases of the plus and minus hands? What is the physical meaning of this relationship?