Statistical Modeling of X-ray Scattering Experiment in Low Photon Count Regime

September 23, 2024

1 Variance of a Gamma/Gamma/Poisson random variable

This section describes the underlying statistical model used to simulate an X-ray scattering experiment at a X-ray free electron laser (XFEL) source.

A random variable (RV) is a function that maps from a set of all possible outcomes from a sample space to real-valued outcomes. Vectors in $\mathbb{R}^n$ can be seen as functions defined on the set $\{1, 2, \ldots, n\}$ mapping to real values. Specifically, a vector $\mathbf{v} = [1, 2, 3]$ defines a function X such that $X(i) = v_i$ for $i \in \{1, 2, 3\}$. Here, $\mathbf{v} = [v_1, v_2, v_3]$ and $X : \{1, 2, 3\} \to \mathbb{R}$, where each index i maps to the corresponding vector component v_i.

In this section, we derive the variance of a three-stage (compound) random variable (RV) driven as follow:

- The first stage of the compound process is to draw two independent RV X and Y such that

$$X \sim \mathcal{G}(1, L) \qquad \text{and} \qquad Y \sim \mathcal{G}(\mu, k) \tag{1}$$

with the following form of the Gamma probability distribution density (PDF)

$$f_Y(y\,;\ \mu, k) = \frac{1}{\Gamma(k)} \frac{k}{\mu} \left(\frac{ky}{\mu}\right)^{k-1} e^{-\frac{k}{\mu}y}. \tag{2}$$

The statistical distribution of incoming X-ray intensity variations can be described by the Gamma distribution as shown in Eq. (**??**). With this parametrization, μ represents the mean intensity, while the shape of the distribution k drives the variance of the RV since

$$\mathrm{VAR}[Y] := \langle (Y - \langle Y \rangle)^2 \rangle = \frac{\mu^2}{k}.$$

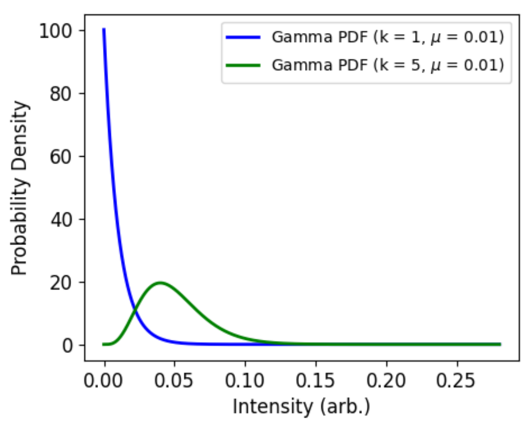

Figure 1: Gamma PDF with the same mean and different shape parameter $k = 1$ and $k = 5$ denoted by purple and green respectively.

When $k = 1$, the Gamma distribution simplifies to the negative Exponential distribution, Fig.**??**. This is analogous to the intensity fluctuations observed due a Self-Amplified Spontaneous Emission (SASE) process, where the SASE FEL intensity distribution exhibits a high frequency of extremely weak shots (causing the distribution to peak near zero) and infrequent shots where the intensity approaches the mean value. Likewise, for self-seeded FEL, k parameter would be greater than 1 with an intensity distribution that is more Gaussian[1]r. This incoming X-ray intensity distribution is represented by the RV Y.

The Gamma RV X encapsulates two key aspects of an experiment:

- The mean of X represents the mean intensity due to sample dynamics, taking the role of a sample response function. While this mean need not be 1, it is inputted as 1 for simplicity.
- L is a way to control the visibility of speckle of the recorded diffraction patterns. The physical phenomena that contributes towards reduction of visibility are: partial longitudinal and transverse, detector pixel size being too large to resolve speckle, and sample or beam instabilities. Thus, the inverse of L is the contrast of your system.

Thus, the parameters μ, k, and L approximately capture various distinct physical attributes of an X-ray beam and the experimental set-up that influence the dynamical parameters, such as the correlation coefficients, contrast and time constants of an X-ray scattering experiment. The RV X represents the combined effects of the sample dynamics, the coherence properties of the beam and imperfect experimental conditions, while RV Y captures the temporal fluctuations of the beam intensity.

- The second stage is to define a second RV Z that is the product of the two RV defined above, i.e., $Z := X \times Y$. We have then

$$Z \sim \mathcal{K}(\mu, k, L) \tag{3}$$

with the three-parameter $\mathcal{K}$-distribution defined as

$$f_Z(z\,;\ \mu, k, L) = \frac{1}{\Gamma(k)\Gamma(L)} \frac{2kL}{\mu} \left(\frac{kLz}{\mu}\right)^{\frac{k+L}{2}-1} K_{L-k}\left[\left(2\frac{kLz}{\mu}\right)^{\frac{1}{2}}\right]. \tag{4}$$

where K is the Bessel function of the second kind.

The RV Z, which is $\mathcal{K}$-distributed, is a product of the two functions that pertain to the sample response and the beam's temporal fluctuations and coherence properties. As such, its mean is the product of the sample response function's mean with beam intensity fluctuation distribution's mean and inherits the same shape parameters k and L. It describes the statistical distribution of the intensity due to the interactions between the sample and the beam, **but without any Poisson photon counting statistics involved**, which means no discretization of photon counts have occurred yet, Fig. **??**.

- The third (and last) stage consist in drawing a Poisson RV W with a mean parameter λ driven by the value of the outcome $Z = z$, i.e.,

$$W \,|\, Z \sim \mathcal{P}(\lambda = z). \tag{5}$$

The last statistical distribution of this three-stage compound process incorporates Poisson statistics, Fig. **??**. The statistical distribution of the number of photons detected in a given time interval on the detector can be modeled by a Poisson distribution. In the extremely

[1] It is actually a standard result that the Gamma PDF converges to the Gaussian PDF as k grows, see for instance [**?**, Sec. 3.3]

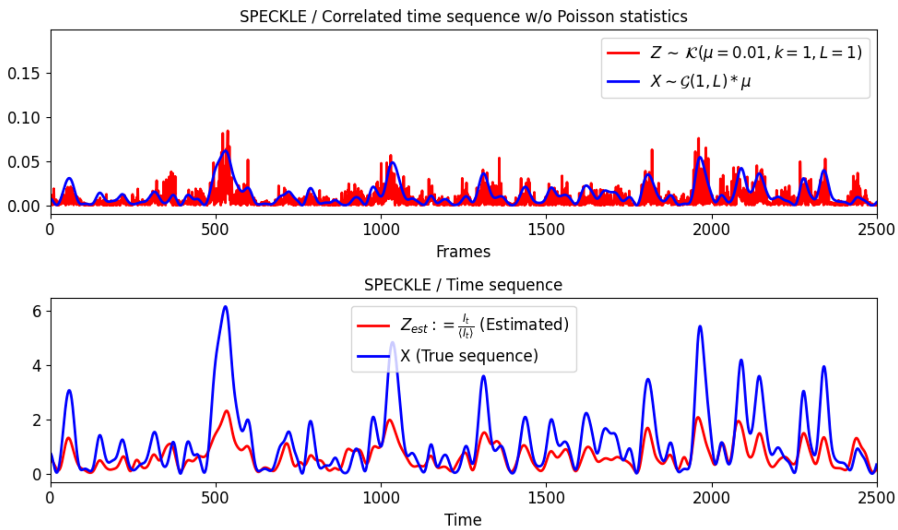

Figure 2: Generated speckle time series of a single pixel with $\mu = 10^{-2}$ mean photon count without Poisson counting statistics. Top: Intensity observed at the detector, RV Z (red) and input intensity RV X modulated by μ (blue). Bottom: Ratio of recorded intensity pattern RV Z normalized by the mean of each frame (red) and the RV X (blue).

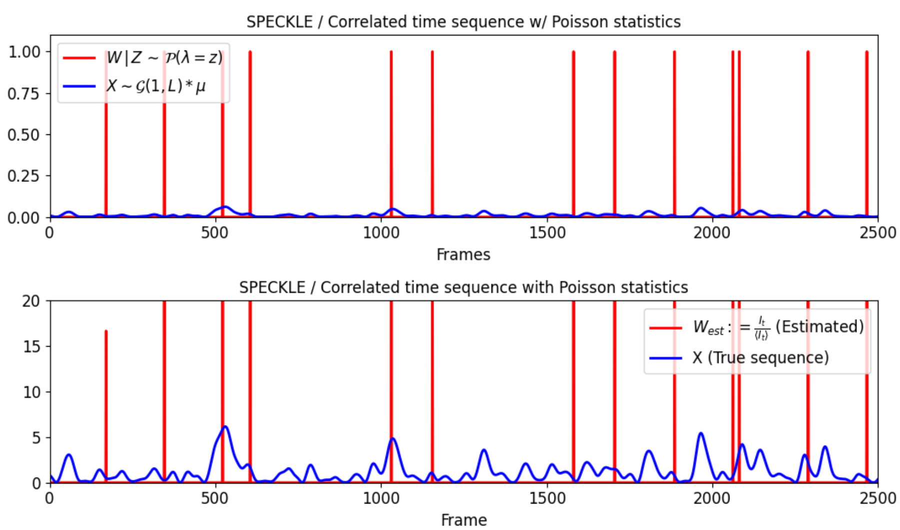

Figure 3: Generated speckle time series of a single pixel with $\mu = 10^{-2}$ mean photon count with Poisson counting statistics. Top: Intensity observed at the detector, W (red) and input intensity X modulated by μ (blue). Bottom: Ratio of recorded intensity pattern W normalized by the mean of each frame (red) and the RV X (blue).

photon-starved regime, Poisson counting statistics becomes prominent because it is rare to detect multiple photons within the time frame of a single data collection period. So W is a conditional probability distribution, describing the outcomes you would obtain with Poisson statistics involved once the outcomes from the product RV Z are considered. Consequently, W represents the intensity distribution of the diffraction patterns recorded, which is a direct observable in an experiment.

Finally, we deduce the PDF of the observed XFEL-count W by marginalizing the distribution of (W, Z),

$$f_W(\omega\,;\,\mu, k, L) = \int f_{W|Z}(\omega \,|\, Z = z) \times f_Z(z; \mu, k, L)\,\mathrm{d}z \tag{6}$$

where the first and second PDFs under the integral are given by (**??**) and (**??**), respectively. An explicit form of this PDF can be found in [**?**, Eq. (4.3)]. The explicit form of Eq. (**??**), resulting from the convolution of $\mathcal{K}$-distributed RV Z with a Poisson distribution, which produces RV W, is only necessary when we want to predict the distribution of intensity detected without specifying the value of the outcome z. Since the knowledge of variance and expectation of W is sufficient to estimate the contrast of the system, we can use the two convolved distributions that make up W to compute the moments of W as follows.

We can compute the expectation and variance of W *via* the first two order moments of the Poisson and $\mathcal{K}$ distributions. In particular, it is easy to derive from (**??**) and (**??**) that

$$\begin{aligned} \langle W \rangle &:= \int \omega f_W(\omega)\,\mathrm{d}\omega \\ &= \int z f_Z(z; \mu, k, L)\,\mathrm{d}z \\ &\equiv \langle Z \rangle. \end{aligned} \tag{7}$$

The expectation of the conditional probability of Poisson RV W given $Z = z$ depends on Z. Therefore, the expectation of W is also the same as that of Z. Intuitively, the convolution of a $\mathcal{K}$-distributed RV Z with a Poisson distribution preserves the mean of the $\mathcal{K}$ distribution since Poisson statistics essentially discretize photon counts without changing the overall mean of the system. The expectation of the $\mathcal{K}$-distributed RV Z (**??**) is μ, which leads to

$$\langle W \rangle = \mu. \tag{8}$$

Similarly, it is not difficult to establish that

$$\begin{aligned} \mathrm{VAR}[W] &\equiv \langle W^2 \rangle - \langle W \rangle^2 \\ &= \mathrm{VAR}[\mathrm{Poisson}(\mu)] + \mathrm{VAR}[Z] \\ &= \mu + \mathrm{VAR}[Z] \end{aligned} \tag{9}$$

Given W is the outcome of a Poisson process given the outcome of $\mathcal{K}$ component, the variance of W should reflect both the intrinsic variance of $\mathcal{K}$ and the Poisson process. For the Poisson distribution, the variance and mean are the same. Because the variance of the RV $Z \sim \mathcal{K}(\mu, k, L)$ is

$$\begin{aligned} \mathrm{VAR}[Z] &\equiv \langle Z^2 \rangle - \langle Z \rangle^2 \\ &= \mu^2 \left(\tfrac{(k+1)(L+1)}{kL} - 1 \right) \end{aligned} \tag{10}$$

we obtain

$$\mathrm{VAR}[W] = \mu + \mu^2 \left(\frac{(k+1)(L+1)}{kL} - 1 \right). \tag{11}$$

We can retrieve L using Eq. (**??**) if we know the beam intensity distribution shape parameter k, since estimate of $\mathrm{VAR}[W]$, $\widehat{v}_W$, can be directly computed from the recorded diffraction patterns.

Similarly, the product kL can be determined if k is unknown. The parameter k can be estimated by fitting a Gamma distribution for its shape parameter from the histogram constructed using the recorded incoming X-ray intensities, which is already something you record in an X-ray scattering experiment. For the usual (fully coherent) case $L = 1$, this relation reads

$$L = 1 \quad \Rightarrow \mathrm{VAR}[W] = \mu + \mu^2 \left(\frac{2}{k} + 1 \right). \tag{12}$$

Finally, the general relation (**??**) can be used to derive a (moment-based) estimate for the parameter L. More specifically, an estimator for L reads

$$\widehat{L} := \frac{\mu^2(1+k)}{k(\widehat{v}_W - \mu) - \mu^2} \tag{13}$$

where $\widehat{v}_W$ is the usual empirical variance computed from a series of outcomes of the RV W.

In summary, this statistical model based on the three-stage compound process captures the various stochastic processes involved in an XRD experiment. It incorporates the fluctuations induced by the X-ray beam and the sample, as well as a way to control the visibility of speckle that affect the dynamical quantities of interest. In particular, the variance of these combined processes is crucial for extracting contrast as in Eq. (**??**).

More specifically, Eq. (**??**) estimates the contrast of the underlying system where the reciprocal of $\widehat{L}$ represents the system's contrast, which is the intercept of the correlation coefficient function at $\Delta t = 0$. This approach overcomes the need to detect two-photon events and to bin data based on incident beam intensity to account for the beam intensity fluctuations, because the analysis take into account of all pixels and how they evolve over time. Consequently, it allows for every shot to contribute towards the contrast estimate easily, enabling the extraction of contrast even with extremely low mean photon count rates, further lowering the detection limit. So far in theory :')

2 Correlation Coefficients in XPCS

Let $I(t)$ be the intensity recorded in a given pixel at time t. We can compute the correlation of $I(t)$ with $I(t+\tau)$ after some time delay τ the following way

$$R^{(1)}_{t,t+\tau} := \frac{\langle I(t) I(t+\tau) \rangle}{\langle I(t) \rangle^2}. \tag{14}$$

Note that $\langle \cdot \rangle$ denotes the expectation operator (i.e., the ensemble averaging). In practice, an estimation of this expectation is obtained by averaging the camera pixels over the user defined ROI, and over various sets of pulse trains provided during the XFEL experiment.

If $\langle I(t) \rangle$ is actually changing w.r.t time (e.g., in the XFEL SASE mode), we need to adapt the relation above

$$R^{(2)}_{t,t+\tau} := \frac{\langle I(t) I(t+\tau) \rangle}{\langle I(t) \rangle \langle I(t+\tau) \rangle} = \left\langle \frac{I(t)}{\langle I(t) \rangle} \frac{I(t+\tau)}{\langle I(t+\tau) \rangle} \right\rangle, \tag{15}$$

the second equality holding because of the linearity of the expactation operator. The Eq. **??** is the same as the TTCF equation in XPCS literature. For notational simplicity, let $I(t) \equiv X$ and $I(t+\tau) \equiv Y$ so that Eq. **??** takes a more compact form

$$R^{(2)}_{X,Y} = \frac{\langle XY \rangle}{\langle X \rangle \langle Y \rangle} = \left\langle \frac{X}{\langle X \rangle} \frac{Y}{\langle Y \rangle} \right\rangle \tag{16}$$

As we explain below, $R^{(1)}$ or $R^{(2)}$ are not the usual correlation coefficient as defined in applied statistics.

3 Correlation Coefficients in Statistics

We introduce the correlation coefficient as it is defined in the statistical literature [**?**, Sec. 27.8] as

$$R_{X,Y} := \frac{\mathrm{Cov}(X,Y)}{\sigma_X \sigma_Y} \tag{17}$$

where $\mathrm{Cov}(X,Y) := \langle (X - \langle X \rangle)(Y - \langle Y \rangle) \rangle$ and $\sigma_X = \sqrt{\mathrm{Var}(X)}$. Since $\mathrm{Cov}(X,Y)$ can also be defined as $\mathrm{Cov}(X,Y) = \langle XY \rangle - \langle X \rangle \langle Y \rangle$ the correlation coefficient above also reads

$$R_{X,Y} := \frac{\langle XY \rangle}{\sigma_X \sigma_Y} - \frac{\langle X \rangle \langle Y \rangle}{\sigma_X \sigma_Y}. \tag{18}$$

How does subtracting the mean (ROI) pixels intensity of each frame, $\langle X \rangle$, from each corresponding recorded pixel intensity at time t, X, before computing correlation coefficients compares to other XPCS methods or affect the resulting correlation coefficients? [Should we suppress this paragraph? (it cannot be understood without knowing that the mean of a exponential RV is equal to its STD, which is stated in the next section).]

Important property of this correlation coefficient is that it satisfies similar properties as an inner product. If two random variables X and Y are of zero mean, then $\mathrm{Cov}(X,Y)$ is the dot product of X and Y, where the "length" of X and Y is σ_X and σ_Y respectively and thus, the correlation is the cosine of the angle between two vectors. That is:

$$R_{X,Y} = \frac{X \cdot Y}{\|X\| \|Y\|} = \frac{\text{Cov(X,Y)}}{\sigma_X \sigma_Y} = \cos(\theta) \tag{19}$$

This means correlation Eq. **??** is a measure of similarity between random variables. From Cauchy-Schwartz inequality [REF], which states that the absolute value of the dot product of two vectors is less than or equal to the product of the norm of the two vectors, we have

$$|\text{Cov(X,Y)}| \leq \sigma_X \sigma_X \iff -1 \leq R_{X,Y} \equiv \frac{\mathrm{Cov}(X,Y)}{\sigma_X \sigma_Y} \leq 1.$$

Clearly, Eq. **??** is an alternative way to construct an element of 2D correlation matrix, TTCF. For uncorrelated random variables (X,Y), $\mathrm{Cov}(X,Y) = 0$, hence, $R_{XY} = 0$ and the maximum correlation value you can obtain is 1 in this construction, different than as in time-averaging of TTCF to obtain g_2.

4 Relationship between the two correlation coefficients

So how is the correlation coefficient used in XPCS analysis, $R^{(2)}_{XY}$ (**??**), related to the statistical correlation coefficient, R_{XY}(**??**)?

First, let us establish the following relations between the standard deviation σ_X and its mean $\langle X\rangle$ for a given probability distribution function (PDF). PDF of pure speckle follows negative exponential, so its mean = standard deviation. Pure speckle, R^{SPE}_{XY}, implies an element of the traditional 2D correlation TTCF matrix in XPCS analysis.

For independent r.v. X and Y, the variance of their sum or difference is the sum of their variances. Thus, when you include Poisson photon counting statistics to pure speckle case, we can simply add the respective variances from each PDF and square root that quantity to obtain the standard deviation of the combined PDF.

PDF for X	Speckle	Poisson	Speckle/ Poisson
Mean	$\langle X\rangle$	$\langle X\rangle$	$\langle X\rangle$
σ_X	$\langle X\rangle$	$\sqrt{\langle X\rangle}$	$\sqrt{\langle X\rangle+\langle X\rangle^2}=\langle X\rangle\sqrt{(1+\frac{1}{\langle X\rangle})}$

The usual statistical correlation coefficient, R_{XY} (**??**), reads under each PDF assumption given in the table:

- ⋄ Speckle, negative exponential ($\sigma_X=\langle X\rangle$) $\implies R_{XY}=\frac{\langle XY\rangle}{\langle X\rangle\langle Y\rangle}-1\equiv R^{\text{SPE}}_{XY}$
- ⋄ Poisson $\left(\sigma_X=\sqrt{\langle X\rangle}\right)\implies R_{XY}=\frac{\langle XY\rangle}{\sqrt{\langle X\rangle\langle Y\rangle}}-\sqrt{\langle X\rangle\langle Y\rangle}\equiv R^{\text{SPE}}_{XY}\sqrt{\langle X\rangle\langle Y\rangle}$
- ⋄ SPE+Poisson $\left(\sigma_X=\sqrt{\langle X\rangle^2+\langle X\rangle}\right)\implies R_{XY}=R^{\text{SPE}}_{XY}\left(1+\frac{1}{\langle X\rangle}\right)^{-1/2}\left(1+\frac{1}{\langle Y\rangle}\right)^{-1/2}$

To summarize:

$$\alpha(\langle X\rangle,\langle Y\rangle)=1 \qquad \text{if PDF = SPE} \tag{20}$$

$$\alpha(\langle X\rangle,\langle Y\rangle)=\sqrt{\langle X\rangle\langle Y\rangle} \qquad \text{if PDF = Poisson} \tag{21}$$

$$\alpha(\langle X\rangle,\langle Y\rangle)=\left(1+\frac{1}{\langle X\rangle}\right)^{-1/2}\left(1+\frac{1}{\langle Y\rangle}\right)^{-1/2} \qquad \text{if PDF = Speckle + Poisson} \tag{22}$$

Since $R^{\text{SPE}}_{XY}:=\frac{\langle XY\rangle}{\langle X\rangle\langle Y\rangle}-1$, when considering SPE+Poisson case:

$$R^{\text{SPE}}_{XY}\equiv\frac{\langle XY\rangle}{\langle X\rangle\langle Y\rangle}=1+\frac{R_{XY}}{\alpha(\langle X\rangle,\langle Y\rangle)} \tag{23}$$

Using XPCS notation:

$$\begin{aligned} R^{(2)}_{t,t+\tau} &:= \frac{\langle I(t)I(t+\tau)\rangle}{\langle I(t)\rangle\langle I(t+\tau)\rangle} \\ &= 1+\frac{R_{I(t),I(t+\tau)}}{\alpha\left(\langle I(t)\rangle,\langle I(t+\tau)\rangle\right)} \end{aligned} \tag{24}$$

5 Numerical Simulation of an X-ray Scattering Experiment