Derivation: From RAW to Signal Intensity

First, consider how the analysis program RAW (& most other analysis softwares) determines the error on the Rg from the Guiner analysis, which is reported as 'uncertainty' in RAW. RAW uses a basic linear least squares approach, which can be applied to linear or non-linear functions of the form:

$$y(x) = \sum_{k=0}^{M-1} a_k X_k(x) \tag{1}$$

where a_k and X_k represent a set model parameters and basis functions, respectively. For these linear models, a merit function is defined as follows:

$$\chi^2 = \sum_{i=1}^{N-1} (\frac{y_i - \sum\limits_{k=0}^{M-1} a_k X_k(x)}{\sigma_i})^2 \tag{2}$$

To discuss the methods for minimizing the merit function (χ^2), some convenient notation will be introduced. Let A be a matrix whose N x M components are constructed from the M basis functions evaulated at the N abscissas x_i (point from y-axis) and from the N measurement errors (σ_i):

$$Design\ Matrix = A_{ij} = \frac{X_j(x_i)}{\sigma_i} \tag{3}$$

The design matrix (A) has the intuitive constraint of $N \geq M$ because you of course must have more data points than parameters.

$$\begin{pmatrix} \mathrm{X}_0(x_0)/\sigma_0 & \mathrm{X}_1(x_0)/\sigma_0 & ... & \mathrm{X}_{M-1}(x_0)/\sigma_0 \\ \mathrm{X}_0(x_1)/\sigma_1 & \mathrm{X}_1(x_1)/\sigma_1 & ... & \mathrm{X}_{M-1}(x_1)/\sigma_1 \\ . & . & . & . \\ . & . & . & . \\ . & . & . & . \\ \mathrm{X}_0(x_{N-1})/\sigma_{N-1} & \mathrm{X}_1(x_{N-1})/\sigma_{N-1} & ... & \mathrm{X}_{M-1}(x_{N-1})/\sigma_{N-1} \end{pmatrix}$$

The design matrix permits and requires additional variable assignments for the end-goal of the minimization of the merit function(χ^2):

$$b_i = \frac{y_i}{\sigma_i} \tag{4}$$

$$'M'\ Vector = [a_0, ..., a_{M-1}] \tag{5}$$

Note, the M-vector consists of parameters to be fitted.

Minimization of the $\chi^{squared}$ metric occurs where the derivative vanishes with respect to all M fitting parameters (a_k).

$$0 = \sum_{i=0}^{N-1} \frac{1}{\sigma^2}(y_i - \sum_{j=0}^{M-1} a_k X_j(x_i))X_k(x_i) \tag{6}$$

where k runs from 0 to M-1. Interchanging sums allows this to be rewritten in matrix notation:

$$\sum_{j=0}^{M-1} \alpha_{kj} a_j = \beta_k \tag{7}$$

$$\alpha_{kj} = \sum_{i=0}^{N-1} \frac{X_j(x_i)X_k(x_i)}{\sigma_i^2} = A^T \cdot A \tag{8}$$

$$\beta_k = \sum_{i=0}^{N-1} \frac{y_i X_k(x_i)}{\sigma_i^2} = A^T \cdot b \tag{9}$$

Therefore, β is a vector of length M, because A contains M-1 columns and b (Eq. 4) is a column vector of length N. Furthermore, equation 7 can be rewritten as follows:

$$\sum_{j=0}^{M-1} (A^T \cdot A) a_j = A^T \cdot b_k \tag{10}$$

or equivalently:

$$\alpha \cdot a = \beta \tag{11}$$

The inverse matrix (C $= \alpha^{-1}$) is the **covariance matrix** which is where we will extract our errors from. To make sense of that, consider(simple rearrangement of equation 7):

$$a_j = \sum_{k=0}^{M-1} \alpha_{jk}^{-1} \beta_k = \sum_{k=0}^{M-1} C_{jk} (\sum_{i=0}^{N-1} \frac{y_i X_k(x_i)}{\sigma_i^2}) \tag{12}$$

It also known (see 'Numerical Recipes: Chapter 2'. *Press, William H., et al. Numerical recipes 3rd edition: The art of scientific computing. Cambridge university press, 2007.*) but not derived here that:

$$\sigma^2(a_j) = \sum_{i=0}^{N-1} \sigma_i^2 (\frac{\partial a_j}{\partial y_i})^2 \tag{13}$$

Note, the covariance matrix is independent of y_i. Now, expand the differential term in the above expression.

$$(\frac{\partial a_j}{\partial y_i}) = \sum_{k=0}^{M-1} \frac{C_{jk} X_k(x_i)}{\sigma_i^2} \tag{14}$$

Substitute this result back into equation 13 and obtain:

$$\sigma^2(a_j) = \sum_{k=0}^{M-1} \sum_{l=0}^{M-1} C_{jk} C_{jl} (\sum_{i=0}^{N-1} \frac{X_k(x_i) X_l(x_i)}{\sigma_i^2}) \tag{15}$$

Where the term in the parenthesis is equal to $\alpha_{kl} = C_{jk}^{-1}$ and equation 15 reduces to:

$$\sigma^2(a_j) = C_{jj} \tag{16}$$

The diagonal and off-diagonal elements of the matrix C_{jj} correspond to the variance of the fitted parameters and the covariance between parameters a_j and a_k, respectively.

The analytical solution for a straight line (i.e. an analytical expression for the error in the slope and intercept) requires the introduction of the following notation:

$$S = \sum_{i=0}^{N-1} \tfrac{1}{\sigma_i^2} \ ; \ S_x = \sum_{i=0}^{N-1} \tfrac{x_i}{\sigma_i^2} \ ; \ S_y = \sum_{i=0}^{N-1} \tfrac{y_i}{\sigma_i^2}$$

$$S_{xx} = \sum_{i=0}^{N-1} \tfrac{x_i^2}{\sigma_i^2} \ ; \ S_{xy} = \sum_{i=0}^{N-1} \tfrac{x_i y_i}{\sigma_i^2} \ ; \ \Delta = S_{xx} S - (S_x)^2$$

$$a = \tfrac{S_{xx} S_y - S_x S_{xy}}{\Delta} \ ; \ b = \tfrac{S_{xy} S - S_x S_y}{\Delta}$$

But we must reconsider this notation in terms of our experimental system where I is the intensity of the measurement, c is the concentration in units $\frac{mg}{ml}$, M is the molecular weight in kDa, $\Delta\rho$ is the excess scattering length (cm^{-2}), t is the exposure time in seconds, and F is the flux in photons per second. Lastly, any variables with at ŝymbol indicate that dependencies have been factored out. For clarity in the derivation, we will let $k = (c \cdot M \cdot \Delta\rho^2)$ and $\tau = (t \cdot F)$

The intensity of the SAXS measurement, I(q), is obtained by subtracting the buffer intensity from the protein intensity and the subtracted intensity is a mix of sample and buffer (Eq.17). Simple propagation of errors states the error (σ^2) of the intensity of the subtracted measurement is the sum of the squares of the errors from the buffer and protein measurements(Eq.18):

$$I(q) = I_{protein} + I_{protein-buffer} - I_{bulk-buffer} \tag{17}$$

$$\sigma^2 = \sigma^2_{sample} + \sigma^2_{buffer} \tag{18}$$

Recalling that each SAXS measurement is considered from the perspective of Poisson counting, where the variance is equal to the counts:

$$\sigma^2 = k\tau\hat{\sigma}^2_{protein} + 2\tau\hat{\sigma}^2_{buffer} \tag{19}$$

The reader should take particular notice the error in the bulk buffer measurement has no dependence on k and a resulting additional factor of 2. As a proxy for the signal *usability* we are interested in the Guiner fit, a linear fit of the SAXS intensity in the low-angle regime of the form:

$$ln(I(q)) = b \cdot q^2 + a \tag{20}$$

where the slope, b $= -(\frac{1}{3})R_g^2$, and the y-intercept, a $= ln(I(0))$. The transformation into log space requires a transformation of the variances from basic principles of error propagation:

$$\sigma_{ln(I_i)} = \frac{\sigma_i}{I_i} = \frac{1}{k\sqrt{\tau}}(\frac{1}{\hat{I}_i})(k\hat{\sigma}^2_{protein} + 2\hat{\sigma}^2_{buffer})^{\frac{1}{2}} \tag{21}$$

In limit of low concentration, and therefore low k, the k dependence of the square-root term in Eq.21 may be ignored. The validity of the assumption was tested **HOW?!**. **Wouldn't this also imply the variance from the protein would drop out?!**. Utilizing the above information, the notation for calculation of the error associated with the linear fit terms in terms of the standard error formula become:

$$S \equiv \sum_i \frac{1}{\sigma^2_{ln(I_i)}} \quad ; \quad S_{q^2} \equiv \sum_i \frac{q^2}{\sigma^2_{ln(I_i)}} \equiv k^2\tau \sum_i \frac{q^2}{(\frac{\hat{\sigma_i}}{\hat{I}_i})^2}$$

$$S_{q^2q^2} \equiv \sum_i \frac{q_i^4}{\sigma^2_{ln(I_i)}} \equiv k^2\tau \sum_i \frac{q^4}{(\frac{\hat{\sigma_i}}{\hat{I}_i})^2}$$

$$\Delta \equiv S \cdot S_{q^2q^2} - S^2_{q^2} \equiv k^4\tau^2(\hat{S} \cdot \hat{S}_{q^2q^2} - \hat{S}^2_{q^2}) \equiv k^4\tau^2\hat{\Delta}$$

The standard error in the slope (b) can then be written as follows:

$$\frac{\sigma_b}{b} = \frac{1}{b}\sqrt{\frac{S}{\Delta}} \tag{22}$$

Substituting in the definition of R_g from the Guiner fit ($R_g = \sqrt{-3b}$) yields:

$$\sigma_{R_g} = |\frac{R_g \sigma_b}{2b}| \tag{23}$$

Therefore we obtain a final expression for the relative error in R_g as a function of the slope of the Guiner fit and the parameters of interest: k and τ.

$$|\frac{\sigma_{R_g}}{R_g}| = |\frac{\sigma_b}{2b}| = \frac{1}{2b}\sqrt{\frac{S}{\Delta}} = \frac{1}{2b}(\frac{1}{k\sqrt{\tau}})\sqrt{\frac{\hat{S}}{\hat{\Delta}}} \tag{24}$$

This analytical model provides a framework for predictive data quality assessment prior to collecting experimental data at a synchrotron beamline thus providing experimentalists with a more efficient way of initially preparing and characterizing the optimal sample conditions for their system of interest. Specifically, it allows for a relative perspective on the dependence of the quality of acquired data on sample concentration, sample molecular weight, and experimental pressure. Furthermore, the form presented in Eq.24 is utilized in the most recent release of SAXSProf, a SAXS measurement simulator, available at (`https://github.com/Mill6159/SAXSProf_Desktop_GUI`).