Error Calculation Stuff

Lucas Wilkins

November 2020

Notation:

N	Number of bins
M	Number of model parameters
λ_i	Expected counts in bin (parameter) i
$\lambda = (\lambda_1 ... \lambda_N)$	Vector of expected counts in all bins (parameter)
τ_θ	Time recording at experimental condition θ, e.g. angle, spin, contrast
$\mu_{i\theta}$	Incident flux in bin i for condition θ (parameter)
$r_i(\xi)$	Reflectance for bin i from model with parameters ξ
ξ_j	Parameters of the structure model
g^z	Fisher information for parameterisation z
$\mathcal{X}$	Support of probability distribution describing measurements
X	Random variable describing a full measurement
X_i	Random variable describing a single bin (ith bin)
$x_i \in \mathcal{X}_i$	Neutron count in bin i
$p(x; z)$	Probability distribution for measurements, parameterised by z
$\Pr([\text{stuff}])$	Probability of stuff
s_i	Total number of incident neutrons in bin i

1 Measurement/Model Details

The model describes the reflectance of a material at a given neutron momentum transfer. These values are binned. The measured number of neutrons in bin i is given by:

$$\lambda_i = r_i s_i(\tau) \tag{1}$$

where s_i is the total number of incident neutrons in bin i. s_i is a function of the amount of incident neutrons in the ith bin, at each experimental condition ($\mu_{i\theta}$), and the time each condition is measured for τ_θ.

$$s_i(\tau) = \sum_k \tau_k \mu_{ik}$$

and so

$$\lambda_i(\xi) = \sum_\theta r_i(\xi)\tau_\theta \mu_{i\theta} \tag{2}$$

Ultimately we want to calculate errors on ξ in terms of τ.

2 General Notes

Some basic probability theory (Kolmogorov Axioms).

Probability space is a triple of quantities (X, Σ, μ). X is the support, Σ is the sigma algebra, μ is a probability measure. The support is what the probability is of. So, for a dice, the support would be the possible rolls, $\{1, 2, 3, 4, 5, 6\}$. Don't worry about Σ, it is usually 2^X (the power set), for a dice it would $\{\{\}, \{1\}, \{2\} \ldots \{1, 2, 3, 4, 5, 6\}\}$. A probability of rolling a number less than 4, would be $Pr(\{1, 2, 3\})$.

Fisher information for discrete variables

$$g^z_{jk} = \sum_{\mathcal{X}_i} p(x_i; z) \frac{\partial}{\partial z_j} \log p(x_i; z) \frac{\partial}{\partial z_k} \log p(x_i, z)$$

3 Application

3.1 Support

$\mathcal{X}_i$ is the possible values that the number of counts in a given bin can take, i.e. $\{0, 1, 2, 3...\} = \mathbb{N}_0$ So, the full measurement takes a value from

$$\mathcal{X} = \mathcal{X}_1 \times \mathcal{X}_2 \times \ldots \times \mathcal{X}_N = \mathbb{N}_0^N$$

3.2 Fisher Information about the λ coordinates

The probability distribution of the measurement in one bin is Poisson distributed:

$$\Pr(X_i = x_i; \lambda_i) = \frac{e^{-\lambda_i} \lambda_i^{x_i}}{x_i!}$$

and the corresponding Fisher information for this bin, with respect to λ_i, is

$$g^{\lambda_i} = 1/\lambda_i = 1/\mathbb{E}[\operatorname{var} X_i]$$

The probability distribution for the whole measurement is, by independence:

$$\Pr(X = (x_0 ... x_N)) = \prod_i Pr(X_i = x_i)$$

and the corresponding Fisher information with respect to λ is

$$g^{\lambda}_{jk} = \begin{cases} g^{\lambda_k} & \text{if } j = k \\ 0 & \text{otherwise} \end{cases}$$

i.e. a diagonal matrix with values of g^{λ_i} which happens to equal (in this case, but not in general), $\mathbb{E}[\operatorname{cov} X]^{-1}$.

3.3 Fisher Information about the ξ coordinates.

In general, we can transform the Fisher information using tensor transforms, i.e.

$$g^Z_{ij} = g^Y_{ab} \frac{\partial y_a}{\partial z_i} \frac{\partial y_b}{\partial z_j}$$

So, the Fisher information in terms of ξ is just:

$$g^{\xi}_{ij} = g^{\lambda}_{ab} \frac{\partial \lambda_a}{\partial \xi_i} \frac{\partial \lambda_b}{\partial \xi_j}$$

3.3.1 As a function of τ

Taking the derivative of equation 1 with respect to ξ gives us

$$\frac{\partial \lambda_i}{\partial \xi_j} = s_i(\tau) \frac{\partial r_i}{\partial \xi_j} \tag{3}$$

The derivative $\frac{\partial r_i}{\partial \xi_j}$ is obtained from the model alone (nothing to do with the data), it is the derivative of the reflectance for bin i with respect to the jth ξ parameter.

We can now put everything together. First, we have the initial Fisher information about the λ parameter, which we can write in terms of s_i and r_i

$$\mathbf{g}^{\lambda} = \begin{bmatrix} 1/s_1 r_1 & 0 & \cdots & 0 \\ 0 & 1/s_2 r_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1/s_N r_N \end{bmatrix}$$

Equation 3 can be re-written in matrix form in terms of a diagonal matrix $\mathbf{S}$ and the Jacobian matrix for the N modelled reflectance points, with respect to the M ξ parameters, $\mathbf{J}$.

$$\mathbf{S} = \begin{bmatrix} s_1 & 0 & \cdots & 0 \\ 0 & s_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & s_N \end{bmatrix}$$

$$\mathbf{J} = \begin{bmatrix} \frac{\partial r_1}{\partial \xi_1} & \frac{\partial r_1}{\partial \xi_2} & \cdots & \frac{\partial r_1}{\partial \xi_M} \\ \frac{\partial r_2}{\partial \xi_1} & \frac{\partial r_2}{\partial \xi_2} & \cdots & \frac{\partial r_2}{\partial \xi_M} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial r_N}{\partial \xi_1} & \frac{\partial r_N}{\partial \xi_2} & \cdots & \frac{\partial r_N}{\partial \xi_M} \end{bmatrix}$$

The tensor transformation of $\mathbf{g}^{\lambda}$ in this notation is then:

$$\mathbf{g}^{\xi} = (\mathbf{SJ})^T \mathbf{g}^{\lambda} (\mathbf{SJ}) = \mathbf{J}^T \mathbf{S} \mathbf{g}^{\lambda} \mathbf{S} \mathbf{J}$$

$$(M \times M) = (M \times N)(N \times N)(N \times N)(N \times N)(N \times M)$$

and the matrix $\mathbf{S}\mathbf{g}^{\lambda}\mathbf{S}$ is a composition of diagonal matrices, and is equal to

$$\begin{bmatrix} s_1/r_1 & 0 & \cdots & 0 \\ 0 & s_2/r_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & s_N/r_N \end{bmatrix}$$

4 Summary

In summary, the Fisher information about ξ is given by:

$$\mathbf{g}^{\xi} = \mathbf{J}^{\mathbf{T}}\mathbf{M}\mathbf{J}$$

where, $\mathbf{J}$ is the Jacobian of the reflectances, r_i, with respect to the parameters, xi. $\mathbf{M}$ is a diagonal matrix with entries $(s_0/r_0, s_1/r_1 \ldots s_N/r_N)$ and s_i is the incident photon flux, which depends on the experimental condition k and the time spent measuring it, τ_k

$$s_i(\tau) = \sum_k \tau_k \mu_{ik}$$

5 Confidence Intervals

To calculate a the confidence intervals in the general case, we can find the set of parameters that differ from the estimate by a certain number of "standard deviations". To do this we need to know the length of a vector in parameter space in terms of the number of standard deviations. This is what the Fisher information does (technically, it gives a linear approximation to the informational distance between distributions, but they're related). The length of a vector in these terms is just:

$$\sqrt{\Delta\xi^T \mathbf{g} \Delta\xi}$$

and so if we want to find a vector with a given length, k, we solve

$$k^2 = \Delta\xi^T \mathbf{g} \Delta\xi$$

It so happens that k in the above equation can be interpreted as "number of standard deviations", so a 2σ error bar will have $k = 2$.

In practice, it is useful to fix a direction, and calculate the magnitude of the vector needed to reach the threshold. i.e. let $\Delta\xi = \epsilon\widehat{\Delta\xi}$ where $\widehat{\text{hat}}$ denotes a unit vector.

6 Confidence Ellipses

In 2D, the unit vectors can be written as $(\sin\vartheta, \cos\vartheta)$. So, to calculate the ellipse we can just solve the following for ϵ over a sample of angles:

$$k^2 = \epsilon^2 \underbrace{\left(\left[\sin\vartheta, \cos\vartheta\right] g^\xi \begin{bmatrix}\sin\vartheta \\ \cos\vartheta\end{bmatrix} \right)}_{\text{scalar}}$$

which then can be plotted as $(\epsilon(\vartheta)\sin\vartheta, \epsilon(\vartheta)\cos\vartheta)$. Another, computationally more efficient approach is to use SVD decomposition to get the 'square root' of g.

7 Sanity Check

If there is a single measurement condition, we have $s_i = \mu_i\tau$ where τ is the total time. Then, we see that τ is a factor of all the s_i's and so of the matrix $\mathbf{M}$ and thus of $\mathbf{g}^\xi$. So, we can write $\mathbf{g}^\xi = \tau\mathbf{f}^\xi$. The size of an "error bar", ϵ at a threshold k in a direction $\widehat{\Delta\xi}$ is given by $k^2 = (\epsilon\widehat{\Delta\xi})g^\xi(\epsilon\widehat{\Delta\xi})$, thus the error bar size ϵ is proportional to $1/\sqrt{\tau}$.

About the threshold. Consider the 1D case. $k^2 = \epsilon^2 g$, g is analogous to the inverse variance, so we have $k^2 = \epsilon^2/\sigma^2$, so, if we want to know where $\epsilon = 2\sigma$, we would have $k^2 = (2\sigma)^2/\sigma^2$, thus $k = 2$.

The notation using $\widehat{\Delta\xi}$ is a more general form than the 1D case as it allows us to probe covariance between parameters, e.g. to draw the ellipses which satisfy the equation. This makes it directly comparable with the sampling methods, and more powerful than simply taking a single value error bar for each parameter.

8 Point Estimates vs Posterior Distributions

8.1 Estimators

An estimator is a function of sampled data that provides an estimation of a parameter. Estimators are usually written with a $\widehat{\text{hat}}$. In the case of a normal distribution, one has standard parameters μ and σ^2, and the variable

X has a probability distribution:

$$p(x; \mu, \sigma^2) = k \exp - \left(\frac{x - \mu}{\sigma} \right)^2 /2$$

and the usual estimators for the standard normal parameters are

$$\hat{\mu}(x_1 \dots x_N) = \frac{1}{N} \sum_i x_i$$

$$\hat{\sigma^2}(x_1 \dots x_N) = \frac{1}{N-1} \sum_i (x_i - \hat{\mu})^2$$

but in general, estimators can be any function of the data, it's just that some estimators are good and others are bad.

In Frequentist statistics, one is concerned with the probability distributions of *estimators*, not of the parameters themselves (as would be the case in Bayesian statistics). For the mean a normal distribution, statistical tests are based on the probability distribution of μ given a certain number of samples N from the normal distribution. When x is normally distributed, the distribution of $\hat{\mu}(x_1 \dots x_N)$ above is also normally distributed, and $\hat{\sigma}(x_1 \dots x_N)$ is Chi squared distributed (which approaches a normal distribution for a large number of samples).

8.1.1 Cramér-Rao Bound

The distribution of estimators often has a variance. When it does, and when it is unbiased, i.e.

$$\mathbb{E}[\hat{\xi}] = \xi$$

then it is related to the Fisher information by the Cramér-Rao bound:

$$\operatorname{var} \hat{\xi} \succcurlyeq \frac{1}{N} (g^{\xi})^{-1}$$

The variance is always "bigger" than the inverse Fisher information. Bigger in this case is a generalisation of numerical inequality to matrices: the differences are positive-definite, i.e.

$$A \succcurlyeq B \iff A - B \text{ is positive definite}$$

This might seem like a weird definition, but it means that if you measure the square length of any vector x by using $x^T A x$ it will result in a bigger number than if you use $x^T B x$. Equivalently, if we think about A and B as representing ellipsoids, it means that B will be contained withing A.

8.2 Bayesian Estimations

The question then is about how to relate this result to the distributions that result from Bayesian methods. The results from Bayesian methods are probability distributions over parameters:

$$\Pr(\xi|x)$$

To make this compatible with the Frequentist approach, we need to derive a estimator (a point estimate) from this distribution. The most obvious way of doing this is by looking at the mean based on the parameter being distributed in this way:

$$\hat{\xi} = \int_{\Xi} \xi \Pr(\xi|x) d\xi$$

Note that under the assumption that this distribution is correct, this estimator is unbiased by definition. The corresponding variance will be

$$\operatorname{var} \hat{\xi}_{Fisher} = \int_{\Xi} (\xi - \hat{\xi})^2 \Pr(\xi|x) d\hat{\xi}$$

Compare to the variance which is the very similar

$$\operatorname{var} \hat{\xi}_{\text{Bayes}} = \int_{\Xi} (\xi - \hat{\xi})^2 \Pr(\xi|x) d\xi$$

8.2.1 A Difficulty

There is a question about whether the probability distribution of the parameters is reflective of the distribution one would have if one makes multiple estimates based on new data.

8.3 A Different Test for the Fisher Information Approach

If we run multiple realisations of the simulated input data, this should relate directly to the Fisher information in a way that the Bayesian approach might not.