Multi-range likelihood fits in RooFit: motivating ROOT PR #7719

Jonas Rembser

April 1, 2021

Let's start with some definitions:

- n : number of ranges
- k_i : number of observations in range
- λ_i : number of expected events in range
- l_i : negative log-likelihood (NLL) for range i, normalized separately in each range
- g_i : integral of unnormalized pdf in range i

$$K = \sum_{i=1}^{n} k_i, \qquad \Lambda = \sum_{i=1}^{n} \lambda_i, \qquad G = \sum_{i=1}^{n} g_i$$

In a non-extended mulit-range fit, the total likelihood is currently expressed as follows in RooFit:

$$L' = \sum_{i=1}^{n} l_i. \tag{0.1}$$

However, this is not correct because the normalization should be done over the union of the ranges. I suggest to *undo* the normalization in each range and add a total normalization term instead:

$$L = \sum_{i=1}^{n} l_i - \sum_{i=1}^{n} \log(g_i) + N \log(G). \tag{0.2}$$

That's what this pull request (PR) implements.

The natural question to ask now is: why do we already get the correct result for extended multi-range fits in RooFit even without this normalization correction?

With the normalization correction and an additional Poisson term for the total number of events, the extended likelihood should be:

$$L_{\text{ext}} = \sum_{i=1}^{n} l_i - \sum_{i=1}^{n} n_i \log(g_i) + N \log(G) - N \log \Lambda + \Lambda. \tag{0.3}$$

What is currently implemented in RooFit is no normalization correction, but instead a separate Poisson term for each sub-range:

$$L'_{\text{ext}} = \sum_{i=1}^{n} l_i - \sum_{i=1}^{n} n_i \log(\lambda_i) + \sum_{i=1}^{n} \lambda_i. \tag{0.4}$$

Fortunately, it turns out that adding an extension term to each range is equivalent to the normalization correction plus a global extension term, so no change is required for extended multi-range fits. To prove this, we need to keep in mind that the expected number of events in each range relates to the total number of expected events just like the pdf integral in each range relates to the total integral:

$$\frac{\lambda_i}{\Lambda} = \frac{g_i}{G}.$$

With this, one can show that the implemented extended likelihood in RooFit L'_{ext} is indeed equivalent to the correct extended likelihood L_{ext}. Or in other words, an extension term for each range is equivalent to the normalization correction plus an extension term for the total number of events:

$$\begin{aligned}
L'_{\text{ext}} &= \sum_{i=1}^{n} l_i - \sum_{i=1}^{n} n_i \log(\lambda_i) + \sum_{i=1}^{n} \lambda_i \\
&= \sum_{i=1}^{n} l_i - \sum_{i=1}^{n} n_i \log(\lambda_i) + \Lambda \\
&= \sum_{i=1}^{n} l_i - \sum_{i=1}^{n} n_i \log(g_i \frac{\Lambda}{G}) + \Lambda \\
&= \sum_{i=1}^{n} l_i - \sum_{i=1}^{n} n_i [\log(g_i) - \log(G) + \log(\Lambda)] + \Lambda \\
&= \sum_{i=1}^{n} l_i - \sum_{i=1}^{n} n_i \log(g_i) - N \log(G) + N \log(\Lambda) + \Lambda \\
&= L_{\text{ext}}.
\end{aligned}$$

That's why we don't need to apply the normalization correction term in the extended case. It's unclear to me if this equivalence was considered when extended multi-range fits were implemented, but it is reassuring to see that the mathematically correct exteded likelihood was implemented.