DRAFT 0.3

Glen Cowan
17 November 2020

Note on Obtaining a Response Matrix from MC

1. General procedure

Suppose we have a Monte Carlo model that generates events characterized by a true value y, and then simulates the measurement of this variable resulting in an observed value x. By generating (x, y) pairs one can produce the scatter plot shown in Fig. 1, which reflects the joint pdf $f(x, y)$.

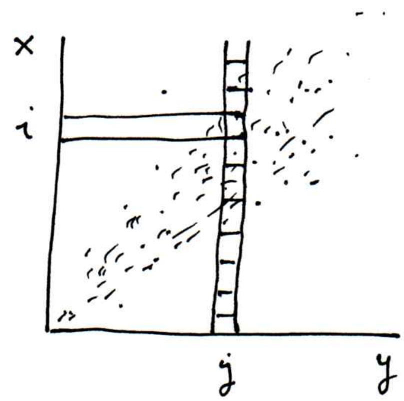

Figure 1: Illustration of a scatter plot of an observed variable x versus the true value y.

Let n_{ij} denote the number of events found in the cell at row i and column j. Suppose the scatter plot has M bins for the true values and N for the observed ones. If an event is generated in column j but not found anywhere, one can put this in an overflow bin, e.g., n_{0j}. In an unfolding problem, one needs the response matrix

$$R_{ij} = P(x \text{ found in bin } i | y \text{ in bin } j) = \frac{n_{ij}}{\sum_{j=0}^{M} n_{ij}} \; . \tag{1}$$

It is in general difficult to find this matrix using MC events alone, since many cells may have few or no entries. One would like therefore to smooth the distribution in some way, e.g., by fitting a parametric function to the scatter plot and then integrating this function over the appropriate regions to determine the response matrix.

The problem of finding an appropriate parametric pdf may be come simpler if one first defines the relative deviation between the observed and true values as

$$u = \frac{x - y}{y} \tag{2}$$

A scatter plot of u versus y could then have the form as shown in Fig. 2.

In many physical situations, the conditional pdf of u given y will be a reasonably bell-shaped curve whose width and other shape parameters will vary only slowly with the true value y. One could therefore hope to parametrize such a density with a function $g(u|y, \boldsymbol{\theta})$.

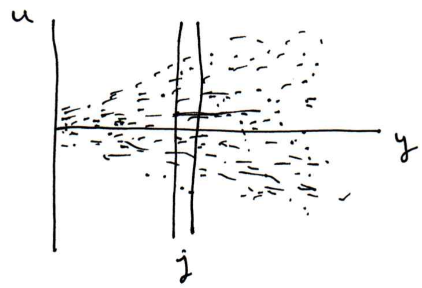

Figure 2: Illustration of a scatter plot of the relative deviation $u = (x - y)/y$ versus the true value y.

As a simple example, one might suppose that this is given by a Gaussian whose mean and width vary smoothly as a function of y, e.g.,

$$g(u|y) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(u-\mu)^2/2\sigma^2} , \quad (3)$$

$$\mu = \theta_1 + \theta_2 y , \quad (4)$$

$$\sigma = \theta_3 + \theta_4 y + \theta_5 y^2 . \quad (5)$$

In this way the entire joint density is specified by the parameters $\boldsymbol{\theta} = (\theta_1, \ldots \theta_5)$. Regardless of the details let us suppose that one succeeds in finding a parametric density $g(u|y, \boldsymbol{\theta})$ and that the parameters are estimated using the MC data.

To fit the response function's parameters, we can use the fact that the joint density of u and y is

$$g(u, y) = g(u|y)g(y) , \quad (6)$$

where $g(y)$ is the marginal pdf of y. Given a sample of (u, y) points, the log-likelihood function for $\boldsymbol{\theta}$ is

$$\ln L(\boldsymbol{\theta}) = \sum_i \ln g(u_i, y_i|\boldsymbol{\theta}) = \sum_i \ln g(u_i|y_i, \boldsymbol{\theta}) + C , \quad (7)$$

where $C = \sum_i \ln g(y_i)$ is a constant in the sense that it does not depend on any of the detector's response parameters $\boldsymbol{\theta}$. By exploiting this fact one can use the pairs of generated points (u_i, y_i) to construct the log-likelihood function (7) and use it to estimate $\boldsymbol{\theta}$.

The equivalent unbinned maximum-likelihood fit could be made using the original density $f(x, y)$. But by transforming to u and y one can construct histgrams of u in bins of y and fit these simultaneously. In this way the individual histograms can be inspected and their goodness-of-fit evaluated.

Once $g(u|y)$ is known in parametric form, one can transform back to the pdf for x using

$$f(x|y) = g(u(x)|y) \left| \frac{du}{dx} \right| = \frac{1}{y} g\left(\frac{x-y}{y} \middle| y \right) . \quad (8)$$

The response matrix is found by integrating $f(x|y)$ over the bins in x while averaging over the bins in y, i.e.,

$$R_{ij} = P(x \in i | y \in j) = \frac{P(x \in i \cap y \in j)}{P(y \in j)} = \frac{\int_j dy \int_i f(x,y)\,dx}{\int_j dy \int f(x,y)\,dx} \tag{9}$$

$$= \frac{\int_j dy\, f(y) \int_i f(x|y)\,dx}{\int_j f(y)\,dy} \approx \int_i f(x|\langle y \rangle_j)\,dx \;, \tag{10}$$

where $\langle y \rangle_j$ is the mean value of y within bin j. That is, we approximate the value of $f(x|y)$ averaged over the bin j with respect to the marginal density $f(y)$ (same as $g(y)$ above) by evaluating $f(x|y)$ with $y = \langle y \rangle_j$, which is exact only if $f(x|y)$ is a linear in y over the bin. For sufficiently small bins this should be an adequate approximation.

2. Determining the response matrix from binned data

If the number of generated events is very large, a single function evaluation of $\ln L(\boldsymbol{\theta})$ may become prohibitively slow. In that case one can bin the data in both u and y. Suppose the full sample has $n_{\rm tot}$ events, and n_{ij} represents the number found in cell (i,j). Here as usual the first index will represent the row, i.e., the u value, and the second index gives the column or y.

The predicted number of events in cell (i,j) is

$$\nu_{ij}(\boldsymbol{\theta}) = n_{\rm tot} \int_{(i,j)} g(u,y)\,du\,dy \approx n_{\rm tot} g(u_i, y_j) \Delta u_i \Delta y_j \tag{11}$$

where Δu_i and Δy_j give the size of the cell and $g(u_i, y_j)$ is evaluated with the u_i and y_j in the centre of cell (i,j). This holds if the distribution changes linearly over a cell, which should be a reasonable approximation for sufficiently small bins.

Since the events are independent and the total number $n_{\rm tot}$ is fixed, the set of n_{ij} follow a multinomial distribution [1],

$$P(\mathbf{n}; \boldsymbol{\theta}) = \frac{n_{\rm tot}!}{\prod_{i,j} n_{ij}!} \prod_{i,j} \left(\frac{\nu_{ij}(\boldsymbol{\theta})}{n_{\rm tot}} \right)^{n_{ij}} \;. \tag{12}$$

The log-likelihood function is therefore

$$\ln L(\boldsymbol{\theta}) = \sum_{i,j} n_{ij} \ln \nu_{ij}(\boldsymbol{\theta}) + C \;, \tag{13}$$

where C represents terms that do not depend on the parameters $\boldsymbol{\theta}$ and can be dropped.

In doing the maximum-likelihood fit, it is useful to obtain a statistic that will provide a measure of goodness of fit. With the multinomial likelihood function this can be done by minimizing [1, 2]

$$\chi^2_{\rm M}(\boldsymbol{\theta}) = -2 \ln \frac{P(\mathbf{n}; \boldsymbol{\nu})}{P(\mathbf{n}; \mathbf{n})} = 2 \sum_{i,j\,:\,n_{ij} \neq 0} n_{ij} \ln \frac{n_{ij}}{\nu_{ij}(\boldsymbol{\theta})} \;. \tag{14}$$

This quantity is the same as $-2\ln L(\boldsymbol{\theta})$ up to an additive constant so minimizing it gives maximum-likelihood estimators for $\boldsymbol{\theta}$. In the large-sample limit the mimized value of $\chi^2_{\rm M}$ follows a chi-square pdf for a number of degrees of freedom equal to the number of cells minus one, so it can be used as a measure of goodness-of-fit.

References

[1] G. Cowan, *Statistical Data Analysis*, Oxford University Press, 1998.

[2] G. Cowan, *Statistics*, in P.A. Zyla et al. (Particle Data Group), Prog. Theor. Exp. Phys. 2020, 083C01 (2020).