DRAFT 0.1

Glen Cowan
11 February 2014

Note on estimate of efficiency

Suppose for each event one measures two variables, x and y. The events can correspond to one of two hypotheses, electron or proton (e or p). Suppose that by requiring $y > y_{\rm cut}$ one can achieve a very high electron purity. The goal is to estimate the electron selection efficiency of the cut on y.

Suppose one creates disjoint intervals according to the variable y, e.g.,

$$\begin{aligned} y_0 &\leq y < y_1 \ , \\ y_1 &\leq y < y_2 \ , \\ &\cdots \\ y_{n-1} &\leq y < y_n \ , \\ y_n &\leq y \ , \end{aligned}$$

where for the last interval $y_n = y_{\rm cut}$; this corresponds to the final interval for which we want to know the efficiency.

Within each interval one can use the x values to determine the fractions of electrons and protons by fitting the function

$$f(x|y \in \Delta y_i,\, a_i) = a_i f(x|y \in \Delta y_i,\, {\rm e}) + (1 - a_i) f(x|y \in \Delta y_i,\, {\rm p}) \ . \tag{1}$$

Here the coefficient a_i gives the fraction of electrons and it is assumed here that the pdfs $f(x|y \in \Delta y_i,\, {\rm e})$ and $f(x|y \in \Delta y_i,\, {\rm p})$ can be determined from Monte Carlo; for now the uncertainty in these shapes is not considered. The output of the fit is then a set of estimated values $\hat{a}_i$ with variances $V[\hat{a}_i]$. As the intervals are disjoint, the estimators are uncorrelated.

The number of electrons $N_{{\rm e},i}$ in interval i can be estimated as

$$\hat{N}_{{\rm e},i} = N_i \hat{a}_i \ , \tag{2}$$

where N_i is the total number of events in the ith y interval. The values N_i can be modeled as following a multinomial distribution with probabilities for each bin of y of $p_0, p_1, \ldots p_n$ and a total number of entries (without any cut on y) of $N_{\rm tot}$.

The covariance matrix for the multinomial distribution is

$$\mathrm{cov}[N_i, N_j] = N_{\rm tot} p_i (\delta_{ij} - p_j) \ . \tag{3}$$

Estimates for these values can be obtained by estimating the individual probabilities using the observed numbers of events found. That is, one takes

$$\widehat{\mathrm{cov}}[N_i, N_j] = N_{\mathrm{tot}}\hat{p}_i(\delta_{ij} - \hat{p}_j) \tag{4}$$

with

$$\hat{p}_i = N_i/N_{\mathrm{tot}} \ . \tag{5}$$

The desired efficiency is the expected number of electrons in y-interval n divided by the total number of electrons, i.e.,

$$\varepsilon_{\mathrm{e}} = \frac{N_{\mathrm{tot}}p_n a_n}{\sum_{i=0}^{n} N_{\mathrm{tot}}p_i a_i} \ . \tag{6}$$

This is estimated by using N_i to determine $N_{\mathrm{tot}}p_i$ and replacing the a_i with their corresponding estimators $\hat{a}_i$, i.e.,

$$\hat{\varepsilon}_{\mathrm{e}} = \frac{\hat{N}_{\mathrm{e},n}}{\sum_{i=0}^{n} \hat{N}_{\mathrm{e},i}} = \frac{N_n\hat{a}_n}{\sum_{i=0}^{n} N_i\hat{a}_i} \ . \tag{7}$$

As the covariance for the N_i and $\hat{a}_i$ are available one can use error propagation to determine the corresponding variance of $\hat{\varepsilon}_{\mathrm{e}}$.

In its current form, there would be no loss in simply taking two y intervals, i.e., $y < y_{\mathrm{cut}}$ and $y \geq y_{\mathrm{cut}}$. The formulae above still apply.

By using a larger number of y intervals, however, one could reduce the statistical error in $\hat{\varepsilon}_{\mathrm{e}}$ if it is possible to parameterize the dependence of the electron fraction a on the variable y. That is, suppose one had a function $a(y;\boldsymbol{\theta})$ for some set of parameters $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_m)$, where we assume that the number of parameters m is less than the number of y intervals $(n+1)$. Then one could carry out a standard least squares fit to determine $\boldsymbol{\theta}$ by minimizing

$$\chi^2(\boldsymbol{\theta}) = \sum_{i=0}^{n} \frac{(\hat{a}_i - a(y_i;\boldsymbol{\theta}))^2}{\sigma^2_{\hat{a}_i}} \ , \tag{8}$$

where y_i could be taken as the centre of the ith interval. This fit would provide a covariance matrix $V_{ij} = \mathrm{cov}[\hat{\theta}_i, \hat{\theta}_j]$. One could then use Eq. (7) above but with $\hat{a}_i$ replaced by $a(y_i;\hat{\boldsymbol{\theta}})$. By using error propagation now with the covariance matrices for the N_i and the $a(y_i, \hat{\boldsymbol{\theta}})$ one can determine the variance in $\hat{\varepsilon}_{\mathrm{e}}$.