DRAFT 0.1

Glen Cowan
21 September, 2016

Ideas on Theory Errors (summary of several conversations with Frank Krauss, Frank Tackmann and the HComb Group)

Starting from conversations at the recent Higgs Hunting conference I wanted to write down some ideas on how theoretical uncertainties could be better treated in our analyses.

Suppose we measure some set of numbers $\mathbf{y}$ and we have a prediction for the probability distribution $P(\mathbf{y})$ from some parametric model. To be concrete assume there is a parameter of interest μ, which could be the signal strength parameter, and a vector $\boldsymbol{\theta}$ of nuisance parameters that may include things like α_s, QCD renormalization scale, factorization scale, parton parameters, etc. In principle one can use this prediction to compare with theory using a likelihood function to test different values of μ,

$$L(\mu, \boldsymbol{\theta}) = P(\mathbf{y}|\mu, \boldsymbol{\theta}) \; .$$

The main idea we started discussing is that for the set of parameters we have currently at our disposal in the prediction (α_s, μ_R, μ_F, etc.) there is no point in this parameter space for which the prediction is "correct".[1] If the parametrization does not include a point that is close to being correct, then any inference about the parameter of interest (discovery significance, parameter estimates) is also incorrect.

A classic example of this kind of situation occurred in estimates of α_s using event-shape variables at PETRA and (early) LEP. Fits were carried out of both α_s and the renormalization scale μ_R using the $\mathcal{O}(\alpha_s^2)$ prediction without leading-log corrections. This resulted in values of μ_R that were far below the energy scale of the reaction and in addition the fitted value of $\alpha_s(M_Z)$ was significantly lower than what was found from other measurements (will find a good reference for this ...). The point is that because of missing higher-order terms in the prediction, there was no point in the (α_s, μ_R) parameter space for which the prediction was correct.

There may be a similar situation in Higgs Physics due to preference of low α_s values from parton fits from DIS (this was emphasized in the talk by R.Boughezal at Higgs Hunting, p. 23).

If no point in the model's parameter space reflects the true prediction, one can simply introduce additional nuisance parameters into the model,

$$L(\mu, \boldsymbol{\theta}) \to L(\mu, \boldsymbol{\theta}, \boldsymbol{\nu}) \; .$$

Now hopefully for some point in the enlarged parameter space the model's prediction is "true", in the sense that the distance by some measure between the model's prediction at its optimal point and the true prediction can be neglected.

At this point there is no longer any systematic uncertainty to deal with. However, because of the correlations between the estimators of all of the parameters, the sensitivity of the

[1] "Correct" here only means it is the true prediction of the theory, e.g., if one could somehow compute to all orders. Whether the theory itself is true is of course a different question.

analysis to the parameter of interest μ will decrease, and will manifest itself, e.g., as an increase in the variance of the estimator $\hat{\mu}$. This is the desired effect of the added nuisance parameters. The systematic error from the incorrect model has been converted into an increased statistical error from a correct one.

The key ingredient of an improved parametrisation is needed for both Bayesian and Frequentist analyses. In a Bayesian context, one simply assigns a prior pdf $\pi_\nu(\boldsymbol{\nu})$ to the nuisance parameters. Bayes' theorem then relates the posterior pdf to the likelihood and prior,

$$p(\mu, \boldsymbol{\theta}, \boldsymbol{\nu}|\mathbf{y}) \propto P(\mathbf{y}|\mu, \boldsymbol{\theta}, \boldsymbol{\nu})\pi(\mu, \boldsymbol{\theta})\pi_\nu(\boldsymbol{\nu}) \ ,$$

and one then marginalizes $p(\mu, \boldsymbol{\theta}, \boldsymbol{\nu}|\mathbf{y})$ over the nuisance parameters to find the posterior pdf for the parameter of interest.

In the Frequentist approach, one can treat the nominal value (best guess?) $\tilde{\nu}$ for each nuisance parameter ν as a measured quantity, e.g., a Gaussian:

$$\tilde{\nu} \sim \text{Gauss}(\nu, \sigma_\nu) \ .$$

Here the notation with a tilde is used to indicate that formally this is a random variable, treated on the same footing as a measured quantity. One may be tempted to call this $\hat{\nu}$, but the hat is used in the profile likelihood ratio later with a different meaning, namely, the maximum-likelihood estimator for ν, so we take $\tilde{\nu}$ here for the "measured" value.[2] In some cases $\tilde{\nu}$ may really be a measurement; in others it is merely a nominal value but from the standpoint of the statistical method it is treated like a measurement.

The set of measured quantities now consists of the main measurements $\mathbf{y}$ and the "control measurements" $\tilde{\boldsymbol{\nu}}$ that constrain the nuisance parameters $\boldsymbol{\nu}$. Therefore the likelihood becomes

$$L(\mu, \boldsymbol{\theta}, \boldsymbol{\nu}) = P(\mathbf{y}|\mu, \boldsymbol{\theta}, \boldsymbol{\nu})P(\tilde{\boldsymbol{\nu}}|\boldsymbol{\nu}) \ .$$

Values of the parameter of interest μ are tested using the profile likelihood ratio as usual:

$$\lambda(\mu) = \frac{L(\mu, \hat{\hat{\boldsymbol{\theta}}}, \hat{\hat{\boldsymbol{\nu}}})}{L(\hat{\mu}, \hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\nu}})} \ .$$

The main point is that to carry out either the Bayesian or Frequentist analysis one must start by parametrizing the prediction in a way that has enough nuisance parameters and thus enough "flexibility" to be correct for some point in its parameter space. From there the analysis task is in effect reduced to a solved problem.

The Gaussian Ansatz for $\tilde{\nu}$ was used above only as an example. It could be motivated, e.g., by the Central Limit Theorem if one thought that the deviation of $\tilde{\nu}$ from its true value was due to a sum of a large number of contributions. On the other hand, if the deviation of $\tilde{\nu}$ from ν was better modeled as the product of a large number of factors, then $\ln \tilde{\nu}$ differs from $\ln \nu$ by a larger number of additive terms and so the Central Limit Theorem would say that $\ln \nu$ follows a Gaussian, i.e., $\tilde{\nu}$ should be modeled by a log-normal distribution. Of course if the measured (nominal) values of the nuisance parameters are correlated then this must be taken into account in the model.

[2] Some people call the measured value ν_0, but to my eye this looks like a constant, not a measured value.

The main question to address, therefore, is how to insert the additional parameters into a model. Suppose one were simply trying to predict a single observable R, which depends on α_s, and by comparing to a measured value $\tilde{R}$ we want to estimate α_s (here, our parameter of interest). To do this we would write down the likelihood function,

$$L(\alpha_s) = P(\tilde{R}|\alpha_s) \ ,$$

which could be e.g. a Gaussian centred about the predicted value of R.

Suppose prediction for R is known to 2^{nd} order, i.e.,

$$R(\alpha_s) = a_0 + a_1\alpha_s + a_2\alpha_s^2 \ .$$

In using this formula, we would be guessing that the unknown coefficients $a_3, a_4, \ldots$ are best approximated as zero. Most probably this is not the case, but based on experience with the calculations of a_0, a_1 and a_2 we may be able to guess something about how big a_3 (or all of the rest of the terms) could be (maybe using Padé approximants?). So one way would be to assign an uncertainty σ_3 to the coefficient a_3 and treat our nominal value $\tilde{a}_3$ as a measured quantity (our "observed" value of $\tilde{a}_3$ is of course 0). The likelihood then becomes

$$L(\alpha_s, a_3) = P(\tilde{R}|\alpha_s, a_3)P(\tilde{a}_3|a_3) \ ,$$

where $P(\tilde{a}_3|a_3)$ could be, e.g., a Gaussian for $\tilde{a}_3$ centred about a_3 with a standard deviation of some given σ_3, or more likely something with longer tails like a Student's t distribution.

If we are not dealing with a single measured number R but rather a distribution, say, dR/dx where x is some kinematic variable, then the missing term is not a single number but rather a function, say, $a_3(x)$. Here one could say that the unknown function (called below $u(x)$ for "unknown") has an expansion in some basis functions of the form

$$u(x) = \sum_n b_n \varphi_n(x) \ .$$

Here the functions φ_n could be functions that have increasing structure (frequency, wobbliness, ...) for increasing n. So on the basis of some understanding of the missing terms one may believe that high-n components are highly suppressed, i.e., that $u(x)$ should be relatively smooth. If x lives on a closed interval one could take, e.g., Bernstein polynomials or some other convenient basis function. So here some minimal set of expansion coefficients b_n become the additional nuisance parameters.

FK's idea to test this procedure is to begin with an example where we know already the higher-order correction (say, 3rd order in the example above). Then various games for estimating the uncertainty using guesses for the missing part can be checked against the known higher order term.

FT's idea was to exploit the fact that some higher-order corrections give a function that passes through zero, and the uncertainty is in effect a question of where the function passes through zero. Here one could simply parametrize a smooth function that passes through zero at a given point and treat that point as the nuisance parameter.