Preprint typeset in JHEP style - HYPER VERSION

hep-stat/

Quantifying systematics associated to modeling imperfections

K Cranmer[1]

[1] *Center for Cosmology and Particle Physics, New York University, 4 Washington Place, New York, NY 10003, United States of America*

Abstract: Particle physicists have detailed and computationally expensive simulation tools to describe mainly numerically (in a non-analytic way) the behavior of the detectors. A new class of systematic effects arising from deviations in the statistical model being used in for fitting does not accurately describe the full simulation is addressed using a modified asymptotic distribution for the test statistic and by evaluating the full simulation one additional time at the best fit point.

Keywords: Model Selection.

Contents

1. Introduction

We want to assess the systematics associated to known deficiencies in the statistical model being used directly by the fitting framework (ie. the RooFit/RooStats workspaces) and the full simulation and analysis recommendations used to describe the baseline model (aka Geant + some corrections and/or smearing).

Let us take the "full simulation" model as $f(x|\alpha)$ and the parametrized statistical model $g(x|\alpha)$, where $\alpha = (\mu, \theta)$ includes both the parameters of interest μ and nuisance parameters θ. Typically we make confidence intervals and inference on μ by eliminating the nuisance parameters θ by marginalization or profiling. Let us focus on the profiling approach, in which the asymptotic theory states that the distribution of the profile likelihood ratio $\lambda(\mu)$ follows Wilks $f(-2\log\lambda(\mu)|\mu,\theta) = \chi^2_{\dim\mu}$ when evaluated at the true value and the more general result from Wald that when $\mu \neq \mu'$ the distribution follows a non-central chi-squared distribution with non-centrality parameter Λ, viz. $f(-2\log\lambda(\mu)|\mu',\theta) = \chi^2_{\dim\mu}(\Lambda)$.

Ideally, the full simulation $f(x|\alpha)$ could be used directly as a likelihood function, but that is computationally intractable currently. Instead, we typically estimate $f_i(x|\alpha_i)$ with i in some a priori defined set of variations: typically one-at-a-time evaluation of each component of α by "$1\,\sigma$". We can take these as accurate estimates of the simulation at those points. From these points we construct an interpolated or parametrized estimate $g(x|\alpha)$, which may or may not satisfy $f_i(x|\alpha_i) = g(x|\alpha_i)$. The real problem is for all the points α away from the sampled points, the interpolation algorithm simply doesn't have enough information to represent the full simulation there. We can write generically that

$$f(x|\alpha) = g(x|\alpha) + e(x|\alpha), \tag{1.1}$$

where $e(x|\alpha)$ is the error in the g's approximation of f.

Before stating the problem, let us imagine an artificially merged family of probability density functions like those used for merging non-nested models. There are two common approaches:

$$h(x|\alpha,\epsilon) = \epsilon f(x|\alpha) + (1-\epsilon)g(x|\alpha) \tag{1.2}$$

and

$$h(x|\alpha,\epsilon) = f(x|\alpha)^{\epsilon} g(x|\alpha)^{(1-\epsilon)} \ . \tag{1.3}$$

The latter approach was preferred by Cox and others, because if f and g are both in the exponential family, then so is h. For (perhaps temporary) technical reasons described below, I will proceed with the first approach (the mixture) in which $h(x|\alpha,\epsilon=0) = g(x|\alpha)$ and $h(x|\alpha,\epsilon=1) = f(x|\alpha)$. Note, these choices are related to what Amari and the information geometry folks refer to as (exponentially) e-flat and (mixture) m-flat, which involves some fairly technical differential geometry.

Statement of the Problem: How does one provide approximately calibrated confidence intervals in the parameter of interest μ for all values of θ when $\epsilon = 1$ is true, but the likelihood function can only be evaluated at $\epsilon = 0$?

That is to say, how do we obtain intervals as if we were using f, even though we only have access to g and perhaps severely limited access to f. Denote the profile likelihood restricted g as: $\lambda_g(\mu) = g(x|\mu,\hat{\hat{\alpha}}(\mu))/g(x|\hat{\mu},\hat{\alpha}(\mu))$, with the typical notation for the conditional and unconditional maximum likelihood estimators. We want:

$$f(-2\ln\lambda_h(\mu)|\mu,\epsilon=1,\theta) = f(-2\ln\lambda_f(\mu)|\mu,theta) \tag{1.4}$$

Approach: To approximate this, we can think of ϵ as an additional parameter of interest where we test at $(\mu,\epsilon=0)$ and assume $(\mu,\epsilon=1)$ is true. We can use Wald's theorem for when the true and tested values are not equal

$$f(-2\ln\lambda_h(\mu,\epsilon=0)|\mu,\epsilon=1,\theta) = \chi^2_D(\Lambda) \ , \tag{1.5}$$

where again we have a non-central chi-square[1]. There are two issues to explore: a) how similar are $\lambda_g(\mu)$ and $\lambda_h(\mu,\epsilon=0)$, and b) how do we estimate the non-centrality Λ?

The Fisher information matrix is defined by

$$I_{ij}(\mu,\epsilon,\theta) = E\left[\partial_i \log L \, \partial_j \log L|\mu,\epsilon,\theta\right] \ , \tag{1.6}$$

which has two important properties. First, it can be used to define a metric on the space of the parameters, which implies geodesics and a natural distance measure between $\epsilon = 0$ and $\epsilon = 1$. Secondly, when the asymptotics hold, the Fisher information is a good estimate for the covariance of the parameters and can be used to

[1]The number of degrees of freedom D, is normally the dimensionality of the parameters of interest. If f and g are very similar, then number of effective degrees of freedom may still be the dimensionality of μ. In particular, if the variance of $\hat{\epsilon}$ may be very large, then the asymptotic convergence has not set in and we have an effect more similar to the energy scale systematic uncertainty that gave a mild "look-elsewhere" type modification to the distribution. Maybe that is actually a better parametrization, but here we use simply that $D = \dim\mu$.

find the non-centrality parameter. For the one parameter case, the relationship is $\Lambda = (\epsilon - \epsilon')^2/\sigma^2$, where σ^2 is the variance of $\hat{\epsilon}$.[2] Of course, we care about the cases $\epsilon = 0$ and $\epsilon' = 1$, thus $\Lambda = 1/\sigma = I_{\epsilon\epsilon}$. We don't have the variance of $\hat{\epsilon}$ and we don't have immediate access to the Fisher information matrix. However, as long as the Fisher information changes very little between $\epsilon \in [0, 1]$, then we can approximate the the distance along the geodesic, D_F, connecting the points $\mu, \epsilon = 0, \hat{\alpha}$ and $\mu, \epsilon = 1, \hat{\alpha}$ with the line integral along the straight path (with fixed $\mu, \hat{\alpha}$) connecting those points

$$D_F = \int_0^1 dt \sqrt{I_{ij}(\alpha(t))\dot{\alpha}_i(t)\dot{\alpha}_j(t)} \approx \int_0^1 d\epsilon \sqrt{I_{\epsilon\epsilon}} = \sqrt{I_{\epsilon\epsilon}} \,, \tag{1.7}$$

It is known that for close by distributions the Kullback-Leibler divergence via:

$$\sqrt{2KL(p||q)} \to D_F \tag{1.8}$$

as $p \to q$ and

$$KL(p||q) = \int dx \, p(x) \ln \frac{p(x)}{q(x)} \,. \tag{1.9}$$

Thus, we arrive at the main result

$$\Lambda \approx 2 \int dx \, f(x|\mu, \hat{\theta}) \ln \frac{f(x|\mu, \hat{\theta})}{g(x|\mu, \hat{\theta})} \tag{1.10}$$

By evaluating $f(x|\mu, \hat{\theta})$ we can incorporate the systematic associated to imperfect modeling.

Idea If we have the ability to evaluate $f(x|\mu, \hat{\theta})$, then we relate

$$\lambda_g(\mu) = \frac{g(x|\mu, \hat{\hat{\theta}})}{g(x|\hat{\mu}, \hat{\theta})} \tag{1.11}$$

$$= \frac{g(x|\mu, \hat{\hat{\theta}})}{h(x|\mu, \hat{\epsilon}, \hat{\hat{\theta}})} \frac{h(x|\mu, \hat{\epsilon}, \hat{\hat{\theta}})}{g(x|\hat{\mu}, \hat{\theta})} \tag{1.12}$$

$$= \lambda_h(\hat{\mu}, \epsilon = 0) \frac{h(x|\hat{\mu}, \hat{\epsilon}, \hat{\hat{\theta}})}{g(x|\hat{\mu}, \hat{\theta})} \tag{1.13}$$

$$\text{fix: will need h at both MLE and CMLE} \tag{1.14}$$

Recall that $h(x|\hat{\mu}, \epsilon, \hat{\hat{\theta}}) = \epsilon f(x|\hat{\mu}, \hat{\theta}) + (1 - \epsilon) g(x|\hat{\mu}, \hat{\theta})$, so this one parameter family can be easily optimized to find $h(x|\hat{\mu}, \hat{\epsilon}, \hat{\theta})$.

Throughout this draft document there have been approximations regarding the maximum likelihood estimators under h and under ff or g. Those differences are hopefully small, but not yet properly worked through in this document.

[2] Maybe if $\hat{\epsilon}$ is not very normal or the issue that non-convergence to the asymptotic distributions is going to cause a problem, but let us proceed with the logic.