The Official Unoffical CMS Reference Guide for Graduate Students

Oz Amram, Lucas Corcodilos, Cristina Mantilla Suarez,
and other JHU graduate students

July 22, 2019

> That's one of those pieces of information in the *Book of Everything You Need to Know* that nobody wrote.
>
> A senior graduate student

The introductory quote (or some form of it) has been said too many times by students who know enough to fill a book but are too busy to write it. This project is meant to remedy that issue for physics at CMS. There are simply too many pieces of the collider, detectors, software, and beaurecratic structure for any one person to know everything. The success of the experiments and the analyses they perform rely on every contributor's ability to be an expert on a few things, to share their expertise, and to learn about areas of inexperience from other experts when necessary.

One of the greatest difficulties of the final responsibility is finding the proper documentation. Even if a group within CMS does an excellent job of documenting their procedures and providing easy-to-use tools, there may not be an obvious way to look for their resources! While searching the CMS TWiki or Google is certainly an option, it typically provides search results totally unrelated to the topic of interest and it can be hard to decipher which pages are recent and which are out-of-date (not to mention all of the personal pages with wrong information!).

This guide attempts to remedy this issue by crowd sourcing references, links, explanations, and sub-guides from the experts in one searchable document. Whether you are enitrely new and need an encyclopedia to reference or are an old hand that just needs to learn about the latest lepton scale factors, there's information for everyone. And if the information you need isn't here, go look for it and become the new resident expert by contributing what you've learned to this document!

Sharing is caring

Sharing knowledge is the key to this project's success so please consider contributing what you know. There's no need to write pages all at once. Just keep this document in mind after you've discovered a nuianced issue with triggers or write an email to an undergraduate describing jet reclustering. Those are excellent moments to contribute because the information is fresh in your mind and most of the work of writing it out may already be done. Simply edit the LaTeX yourself or just drop the text into an issue on GitHub and let others do it for you.

Enjoy and feel free to ask questions!

1 Corrections, Weights, and Scaling

This section covers different types of corrections, weights, re-weights, scaling, and other multiplicative factors required because of either differences in simulation and data or known inconsistencies in reconstruction algorithms that affect objects in both simulation and data. Each section briefly covers the reason for the corrections, how they are derived, and where to find the latest values and uncertainties.

1.1 Jet Corrections

1.2 Pileup

1.3 Tagging

1.4 Triggers

Using High Level Triggers (HLTs)

High Level Triggers are set to 0 or 1 depending if their conditions are satisfied. Some of these are very simple like HLT_PFHT1050 which says "true if the transverse hadronic activity/energy is greater than 1050 GeV" (PF means the variables are reconstructed using the Particle Flow algorithm). Others are more complicated like AK8PFJet420_TrimMass30 which says "true if there's an AK8 jet with pt greater than 420 GeV and it's mass after 'trimming' is greater than 30 GeV." Analyses usually use a combination of logical ORs of these triggers.

In some cases, the simulation and data don't have all of the trigger bits. That means the bits won't get saved. In this case, the trigger is treated as if it's false.

1.5 Cross Sections and Luminosity

Ignoring the fact that we cannot simulate physics exactly, one still cannot directly compare simulation against data because the number of simulated events for process X will not match the number of events where process X actually occured in data. To correct for this, we renormalize the yield to the cross section of the process and luminosity of the data. This is weight is derived with

$$\frac{xsec * lumi}{nevents} \tag{1}$$

where $xsec$ is the cross section, $lumi$ is the lumionsity of data collected, and $nevents$ is the number of events generated. Note that $nevents$ is NOT the number of events remaining after making a selection - it has to be the number of events generated before any selection is made. Making a selection first would introduce an efficiency term which we don't care to consider at this step (that's what the other corrections to simulation are for).

Applying this weight will normalize the simulation to a yield comparable to the data.

1.5.1 Signal simulation

Signal simulation is treated uniquely relative to the simulation of backgrounds because the backgrounds have been studied and their cross sections are known (with an uncertainty of course). With signal, we typically want to solve for the cross section. This doesn't mean you can't use the theoretical cross section though. In fact, using it can be useful to set the scale and allow one to solve for a unitless normalization of the simulation template called the `signal strength`.

The signal strength is fit for when comparing data against a background estimate. In the backgroud-only hypothesis, it is fixed to 0 because the hypothesis assumes no signal exists. In the so-called signal+background hypothesis, the signal strength is left to float and the fit tries to "fill-in" parts of the distribution with it. If the signal simulation template is normalized to its theoretical cross section and the luminosity of the data being analyzed, then a signal strength of 1 means the template is exact. A value of 2 means there is twice as much signal as the simulation (including the cross section value) predicts and so-on.

If the signal simulation is only scaled to the luminosity (this means the cross section is effectively set to 1), then fitting for the signal strength is equivalent to fitting for the true cross section. This may be more desirable in certain circumstances and both methods can be used to check that the fit is stable and finds the same physical answer in both scenarios.

2 Accessing Data and Monte Carlo Sets

Using DAS (Data Aggregation System)

$t\bar{t}$ Simulation

- For 2016, $t\bar{t}$ MC was generated for inclusive decays (meaning, all possible branchings of the two top quarks were allowed). For 2017 and 2018, decays were split into all-hadronic, semi-leptonic, and leptonic (denoted 2L2Nu).

Glossary

N-1 Plot

A plot as a function of variable X where a selection has been applied to all variables except X (N total variables with N-1 cuts applied). These are typically used to either compare the shapes of signal and background as a function of X or to scan for an optimal point to place a cut to maximize the (cumulative) $S/\sqrt{(}B)$. It's not uncommon to also do N-2 or N-3 plots depending on the scenario.

Signal strength/r/μ

A normalization factor that is fit for when comparing data against a background estimate. In the backgroud-only hypothesis, this is 0 because the hypothesis assumes no signal exists. In the so-called signal+background hypothesis, the signal strenght is left to float and "fill-in" any peaks in the distribution. If the signal simulation template is normalized to its theoretical cross section and the luminosity of the data being analyzed, then a signal strenght of one means the template is exact. A value of two means there is twice as much signal as the simulation (including the cross section value) predicts. If the signal simulation is only scaled to the luminosity (this means the cross section is effectively set to 1), then fitting for the signal strength is equivalent to fitting for the true cross section.