The Official Unoffical CMS Reference Guide for Graduate Students

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

August 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results totally unrelated to the topic of interest and it can be hard to decipher which pages are recent and which are out-of-date (not to mention all of the personal pages with wrong information!).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 Scale Factors

Because of our imperfect simulation, sometimes the efficiency of making a cut on a MC sample will not have the same efficiency as making the same selection on data. For example it might be that requiring electrons to pass the Medium ID selects 70% of real electrons in simulation but only selects 68% of electrons in real data. In order to fix our simulation to match the data, a scale factor, given by the efficiency in data divided by the efficiency in MC is used. In this case that would be $\frac{0.68}{0.7} = 0.971$. So then for every event in simulation where we required an electron to pass a Medium ID and it did, we would multiply the weight of that event by 0.971 so that the MC distribution would match the one in data.

Note that this was a very simplified example. In general the efficiency, and thus the scale factors (SF's), will depend on the pt and eta of the object you are requiring the selection on. Things that scale factors are typically applied to are triggers, ID's, isolation selections, reconstruction efficiencies, b-tagging selections, etc. Often times these scale factors will be provided by the various object groups like Muon POG, EGamma POG etc. Sometimes you have to derive them yourself which can be annoying.

The most common method of measuring scale factors is called Tag and Probe. This page is very outdated but has a good description of Tag and Probe.

1.2 Jet Corrections

1.3 Pileup

1.4 Tagging

1.5 Triggers

There are way too many collisions for CMS to process and record each one. Also the vast majority of collisions are just going to be uninteresting low energy QCD junk anyway. For this reason triggers are used. Triggers are flags that quickly determine if the event is interesting and is worth being reconstructed and saved. Every event that we actually record in data passed some trigger.

CMS uses a two level trigger system. The first level is called the L1 trigger which for every event has 4 microseconds to decide whether it is interesting or not. Generally the L1 trigger has to use very basic detector information because it doesn't have time to do reconstruction. Events that pass the L1 trigger are then given to the HLT trigger which has $\tilde{1}00$ ms per event to decide if it should be kept. So it can afford do a little reconstruction to figure out if the event is interesting or not, but still no where near the time to do a fully detailed reconstruction. If the event then passes the HLT it is actually saved.

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. The 'Menu' of what HLT triggers are available for you use changes for each run.

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.

1.6 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

$$\frac{xsec * lumi}{total\_mc\_weight} \tag{1}$$

where $xsec$ is the cross section, $lumi$ is the lumionsity of data collected, and $total\_mc\_weight$ is total weight of all the produced MC events. In the simple case, where the MC generator gives every event the same weight, than the total weight is just the total number of events. Sometimes however, just to be annoying, generators produce events to have different weights and sometimes events even have negative weights (aMC@NLO does this). So in general the total weight is the sum of the weights of all the produced events.

Note that we cannot just include the weight 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 ignore the fact there are some events we for a given process that we do not reconstruct but still happen. We correct for the 'efficiency' of our selection in other steps (that's what the other corrections to simulation are for). Applying this weight will normalize the simulation to a yield comparable to the data.

1.6.1 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)

The easiest way to look for MC and data samples is with DAS. Which is here. The datasets will generally have the format "/XXXX/YYYY/ZZZZ" .

The XXXX will be something about the actual physics content of the dataset, eg WW_TuneCP5_13TeV-pythia8 is for WW simulation made with Pythia using the tune CP5. For data it will describe the triggers used for that dataset eg "SingleMuon".

The YYYY will be about the production info of that data set. For MC it will be something like "RunIIAutumn18MiniAOD-102X_upgrade2018_realistic_v15-v2" which means it was made during the RunIIAutumn18 round of MC production in the 102X release with some other info. For data it will be something like Run2018A-17Sep2018-v2 which means its from RunA from 2018 data taking and it was reconstructed in the 17Sep2018 batch. Often times data is reconstructed multiple (Re-Reco'ed) times with progressively better calibrations, usually you want the latest one.

The ZZZZ will be the data type like MINIAOD or NANOAOD.

You can use wildcards to help you find things if you don't know the exact name. Once you find a dataset you can look at a list of all its files (its often a good idea to run on one file before you run a full crab job). There is also a link to cross-section database entry for that dataset which sometimes works. Usually this cross section is not the most accurate one, but it is useful as a starting point.

$t\bar{t}$ 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).

3 Useful Resources (Links)

Note that the links don't seem to work in the pdf showed by github. So download it and you should be able to access them.

- **CMS Workbook**: A decent general introduction to many of the different computing tasks. Usually only a little out of date. A good first place to look.
- **CMS Twiki**: General Twiki pages for Physics Object Groups (POG's), like Muons, Jets/MET, and Physics Analysis Groups (PAG's) like B2G, EXO, etc and other groups. Usually groups keep their pages relatively up to date, but it depends on the group. This is the place to look for recommendations on common things like working points, scale factors, etc.
- **CMS Induction Course**: A series of lectures for new people entering the collaboration. Gives a good overview of different areas of CMS, from detectors, to organizational structure.
- **Data Aggregation System(DAS)**: Look up datasets to use. See the section on using DAS.
- **MCM**: Look up information on the production history of different datasets. Can check the configuration or parton shower settings of generator used for the MC production. Can search by DAS name.
- **Hypernews**: A tool for email forums. You should subscribe to the forums related to the things you are working on. These forums are also a good place to ask for help if you have an issue or question with some CMS tool and can't find help on the Twiki. People are usually helpful but it may take a few days to get a response.
- **LPC Tools Documentation**: A good resource if you are having issues with EOS, Condor, grid certificate issues, etc. Anything related to using LPC.
- **Grafana (CRAB job monitoring)**: Check how your crab jobs are doing. Select your user name in the top left box to only see your jobs. The switched to a new service as of Summer 2019 and its still a little buggy. Good idea to check with 'crab status' if something looks weird.

- **CADI**: Look up past and current analysises, contains links of most recent analysis note, paper, where they are in approval process etc.
- **Combine Documentation**: Documentation for combine, the most common fitting tool in CMS.

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.