CMS Data Analysis Tutorial

January 4, 2016

Introduction

This tutorial is based on a fraction of (50 pb^{-1}) real CMS data that is collected at 7 TeV of proton-proton collisions at the LHC, and is released for educational purposes (see http://ippog.web.cern.ch/resources/2012/cms-hep-tutorial). Also, the exercises we will follow are very similar to that of followed in CMS HEP Tutorial by Dr. Christian Sander and Dr. Alexander Schmidt who are the designers of this tutorial. In this tutorial, we are going to consider pair-produced top quark events, where one of the top quarks decays to bottom-quark and W^+, then W^+ decays to a lepton(electron, muon, tau) and its neutrino. The other top quark (actually it is an anti-top) decays to anti-bottom and W^-, then W^- decays to two one b-quark and two quarks. Each of these quarks pronounces themselves as collimated spray of particles which we call as jet. This cascade of decay is called as the "semi-leptonic" decay channel of a top quark pair.

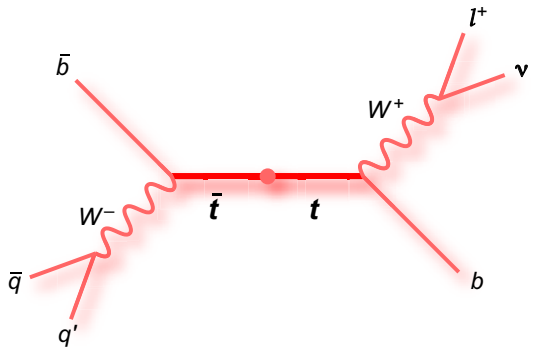

Figure 1: semi-leptonic decay topology.

filename	type	#events	x-section	int. lumi.	trig. only
`data.root`	data	469384		50 pb^{-1}	yes
`ttbar.root`	sim. $t\bar{t}$ signal	36941	165 pb	50 pb^{-1}	no
`wjets.root`	sim. W plus jets background	109737	31300 pb	50 pb^{-1}	yes
`dy.root`	sim. Drell-Yan background	77729	15800 pb	50 pb^{-1}	yes
`ww.root`	sim. WW background	4580	43 pb	50 pb^{-1}	yes
`wz.root`	sim. WZ background	3367	18 pb	50 pb^{-1}	yes
`zz.root`	sim. ZZ background	2421	6 pb	50 pb^{-1}	yes
`single_top.root`	sim. single top background	5684	85 pb	50 pb^{-1}	yes
`qcd.root`	sim. QCD multijet backgr.	142	10^8 pb	50 pb^{-1}	yes

Table 1: Data and simulated Monte Carlo samples.

In an experiment, we only have the final particles of a decay as detected and reconstructed. However, there are different possible elementary physical processes which give the same final products. We call these events as "background events". Since nobody tells us about the origin of the final particles, we have to find a way to differentiate our "signal events" from those background events. To this purpose, by using Monte Carlo (MC) simulation techniques, we produce sample of events based on different physics processes that may give the same final state as the signal process. For the tutorial, we have several already-produced MC samples (see Table 1). We need to study the differences in kinematical properties of the signal and background events with the help of MC event samples. Using these differences, we should apply some selections on the data events to suppress the background events. These selections, of course, discard some of the signal events. One should try to optimize his/her selections - so called as "cuts" in High Energy Physics (HEP) world - in order to suppress the signal events as least as possible while suppressing the background events as much as possible. The number of simulated events can be much smaller or much larger than required for an integrated luminosity of 50 pb^{-1}, so these have to be weighted accordingly. The weight is stored in the `EventWeight` variable in each sample.

1 Warm Up

In order to get used to the tree structure of the data and sample files, we will get some variables out of the trees, and plot some observables. Here is the list of tasks for the warm up session:

- Make a plot of i) isolated muon multiplicity, ii) isolated electron multiplicity. For this exercise, the isolation criterium can be taken as `Muon_Iso[i]>0.125` and `Electron_Iso[i]>0.1.`
- Make a plot of dimuon invariant mass for each MC sample. You can loop over the muons for each event, and construct TLorentzVectors for those muons. Then, you should calculate the invariant mass of two oppositely charged muons leading in pT to fill a histogram with these mass values.
- Make a plot of b-jet multiplicity. It is not possible to identify the quark origin of a jet in general. However, if it is originated by a b-quark, it may be said something about whether it is a b-jet or not. b-quarks first hadronise into a B-Meson which has a relatively long lifetime of 1.638 ± 0.011 x 10^{-12} s. This time interval helps us to measure the distance between the primary vertex (vertex of the hard pp interaction) and the decay vertex of the B-Meson. Based on this measurement, a discriminator (b-tag discriminator) variable is calculated which indicates the probability for a jet is originated by a b-quark. The higher the value, the more likely the jet is a b-jet. In this exercise, we may use Jet_btag[i] > 0.8.
- Compare your results to MC simulation by drawing your histograms for the data and for the simulation on the same canvas. Make sure that you select only triggered events for the MC samples.

2 How to select $t\bar{t}$ events

As it is mentioned in the introduction part, we have to find some cuts to reject background events that might look like the signal events while accepting the signal events efficiently.

- Just by looking at the semi-leptonic decay topology, it is easy to see that we need at least one isolated lepton in the final state. Compare several other distributions of event variables for simulated signal (events) and background.
- Try to find variables which are especially sensitive to separate signal from background (jet multiplicity, transverse momenta of jets and leptons, lepton isolation, b-tagging, missing transverse energy, angular distributions like η or ϕ of muons, electrons and jets). Fill all these distributions into histograms and compare between signal, background and data.
- Try to find cuts on these histograms that accept more signal and less background. Among several set of cuts, select the optimum cut set that gives the best signal to background ratio (purity).
- Apply the same cuts on the data. Compare the selection efficiency on data vs. on MC.

3 $t\bar{t}$ Cross Section

If we are sure that $t\bar{t}$ signal events are efficiently selected, the next step will be the calculation of cross-section of top quark pair production at the LHC.

- The first ingredient is the trigger efficiency ε_{trig}. We may rely on the MC simulation to reproduce this efficiency correctly. Plot trigger "turn-on" curve which shows the trigger efficiency depending on the muon transverse momentum p_T. You may use `TEfficiency` class. In order to utilize this class, you have to create two histograms i.e. `h_pass` and `h_total` where the first one is the isolated muon p_T histogram for triggered events (`triggerIsoMu24`==1) and the second one is for all of the events. Calculate the efficiency of triggering top quark events with a reconstructed and isolated muon of $p_T > 25$ GeV.

- The second ingredient is the acceptance ε_{accept} (not including the trigger). This includes the fact that we only select semi-leptonic top quark decays with muons that is almost 30% of all kind of decays of $t\bar{t}$ pairs. You can calculate the acceptance by comparing the number of generated top quark events with the number of selected events, after all your cuts. The number of generated and selected events must be calculated by summing `EventWeights` for each event, not the bare numbers.

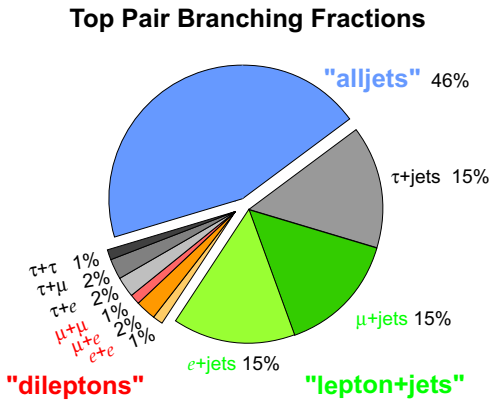

Figure 2: Top pair branching fractions.

- background subtraction: we also trust the simulation to correctly predict the number of back- ground events after selection. Subtract the expected background from the observed (selected) data events. That number is your number of observed signal events.

- By using the relation:

$$N_{obs}^{signal} = \varepsilon_{trigg} * \varepsilon_{accept} * \mathcal{L}_{int} * \sigma_{t\bar{t}}$$

  You can calculate the cross section $\sigma_{t\bar{t}}$.

- Compare your result with measurements performed by the ATLAS and CMS Collaborations at $\sqrt{s}$ =7 TeV of pp collisions. What could be the sources of difference between your results and the others?

4 Top Quark Mass

After selecting the signal events, you can construct the fourvectors of top quark by using the final products of $t\bar{t}$ decay.

- First of all, let's determine the mass of the top quark in MC simulation (in $t\bar{t}$ events)? You should use the generator-level truth information stored in the trees to form the top quark fourvector both for the hadronic and leptonic branch.

- Now, try to use detector objects only in $t\bar{t}$ signal events. You should start by forming the fourvectors of jets that are not b-tagged. These jets are more likely originated from the W boson.

- Once you have the W boson fourvector in the event, add this fourvector to a b-jet four vector. As there are two b-jets, calculate the angular distance between the W and b-jets in $\eta - \phi$ space ($\Delta R = \sqrt{(\Delta\eta)^2 + (\Delta\phi)^2}$), and use the closest b-jet to combine with W.