Report on MOD Data-CMS 2011A

Andreas Lolos

April 2021

1 General introductory information

Data was taken from $2.3fb^{-1}$ proton-proton (p-p) collisions at center of mass energy of $\sqrt{s} = 7TeV$, during 2011 RunA at the LHC. A sample of $1,785,625$ central jets with transverse momentum (pT) above $375GeV$ is isolated.
The CMS Open Data release from the above LHC 2011A Run, includes detector-simulated Monte-Carlo (MC) samples, as well as samples of hard QCD scattering generated by $PYTHIA$ 6.4.25 with $Z2$ tune. Due to the exploratory nature of this study (pull vector), there may be no need to unfold for detector effects nor estimate systematic uncertainties. A general agreement between CMS 'real' collision data, simulated MC samples and Pythia generated data should suffice for the experimental robustness of this study.
It is crucial to note, that CMS Open Data contain information about reconstructed particle flow candidates (PFCs), which makes pull vector construction possible. Additional information is provided for the primary vertices of events, allowing to mitigate pileup. Pileup involves the contamination of the energy distribution arising from leading vertex by radiation from soft collisions (pileup).
It is important to note at this point that all of the above and below information can be found on [2]. A complementary paper that contains similar info about how data were distilled can be found [3]. Although the latter contains info on 2010 dataset released by CMS, it proved to be useful.

2 CMS Open Data

The CMS Open Data is grouped into Jet Primary Datasets (JPD) that contain a subset of the triggers used for event selection. There are 19 JPDs included in the 2011 release, along with corresponding MC samples. JPDs are provided by CMS collaboration in Analysis Object Data (AOD) format. These datasets provide high-level reconstructed objects used for CMS analyses in Run 1. Distilled MIT Open Data (MOD) are extracted from Jet Primary Dataset [1].
This JPD includes a variety of single jet and dijet triggers and contains $30,726,331$ events spread across $1,223$ AOD files. The 2011A data-taking period is divided into 318 runs which in turn are divided into 109,428 luminosity blocks (LBs). So what is a LB;

A luminosity block is the smallest unit of data-taking for which there is luminosity information and during which the triggers 'work properly'. To put it simply, an event is the recorded data from one beam crossing. A luminosity block is a collection of temporally consecutive events. A run consists of many temporally consecutive LBs.

As mentioned before, PFCs are provided for each event in the AOD format. PFCs are particle-like objects containing the reconstructed four-momentum and particle identification code (PID). Additionally, AK5 jets are provided by CMS. These are jets produced by clustering PFCs with the anti-kt jet algorithm (radius parameter is R=0.5).

3 MIT Open Data (MOD)

According to authors, MOD are specked out in a certain way. The hardest or/and second hardest jets are considered for their analysis. So, in data only exist single or dijet events with pt that is corrected using jet energy correction (JEC) factors. Also, a jet quality cut is imposed, so that the jet's pseudorapidity $|\eta^{jet}| < 1.9$. PFCs are reclustered into AK5 jets and then compared to CMS's AK5 objects. If the four-vector for AK5 jets differs more than 10^{-6} then the jet is dropped along with its PFCs.
Attention is focused to 9 single-jet triggers. In MOD exist events for which HLT_JET300 fired. This trigger is the lowest pt single-jet unprescaled trigger. For $pt > 375GeV$, it is nearly 100% efficient. Single-jet triggers are designed to fire any time an event contains a jet with pt above threshold, as stated [2].
Simulation and generator data exist alongside MOD. These data files are extracted from AOD files with names that correspond to hard-scattering parton pt range. Sim and Gen jets are matched if $\Delta R <= 0.5$. Gen jets are matched to final state parton if $\Delta R <= 1.0$.
Jet studies executed by the authors are based on the 2 hardest jets in an event. This is motivated by the fact that $2 \to 2$ QCD dijet production at leading order yields 2 jets of equal pt. If more than 2 jets are considered then information beyond leading order is needed (don't quite get this). Due to pt threshold being almost $375GeV$, in 'real' collision data, many events, for which one jet exists in data, can be found. Furthermore, after selection of events, simulation dataset, according to TABLE IV, [2] is bulkier than CMS Open Data dataset, by a factor of almost 35.

References

[1] CMS collaboration. *Jet primary dataset in AOD format from RunA of 2011 (/Jet/Run2011A-12Oct2013-v1/AOD). CERN Open Data Portal.* DOI: http://doi.org/10.7483/OPENDATA.CMS.UP77.P6PQ.

[2] Patrick T. Komiske et al. "Exploring the space of jets with CMS open data". In: *Physical Review D* 101.3 (Feb. 2020). ISSN: 2470-0029. DOI: 10.1103/physrevd.101.034009. URL: http://dx.doi.org/10.1103/PhysRevD.101.034009.

[3] Aashish Tripathee et al. "Jet Substructure Studies with CMS Open Data". In: *Phys. Rev. D* 96.7 (2017), p. 074003. DOI: 10.1103/PhysRevD.96.074003. arXiv: 1704.05842 [hep-ph].