PROCEEDINGS OF SCIENCE

Comparison studies for b tagging variables in data and simulation

Garyfallia Paspalaki* **On behalf of the CMS collaboration**
Nat. Cent. for Sci. Res. Demokritos
Institute of Nuclear Particle Physics
E-mail: garyfallia.paspalaki@cern.ch

The identification of jets originated from b quarks is crucial for a broad range of physics analyses. Various b tagging algorithms exist at CMS and are further developed with the use of the machine learning techniques. Constant monitoring of the basic quantities provided to the high-level taggers is fundamental to ensure a good tagging performance and to spot potential issues in the data taking. We present a comparison between the proton-proton collision data collected by the CMS detector in 2016 and simulation. The comparison is between the input variables used by the heavy flavour tagging algorithms and the taggers distributions in several event topologies.

Sixth Annual Conference on Large Hadron Collider Physics (LHCP2018)
4-9 June 2018
Bologna, Italy

*Speaker.

PoS(LHCP2018)077

1. Introduction

The efficient identification of jets originating from bottom quarks (heavy-flavour jets) is fundamental for measurements and beyond the standard model searches at the LHC. Different b-tagging techniques are defined in CMS which benefit from the long life time, high mass and large momentum fraction of the b-hadron produced in b-quark jet. In order to validate the tagger performance it is necessary to compare the simulated input variables and the tagger distributions with the data.

2. Event Topologies

Different event topologies were used to compare data with simulation. Each topology consists of a different jet flavour composition and therefore it is possible to assess the agreement between data and simulation for several cases . The events were selected according to the following criteria.

- **Inclusive multijet sample:** This sample is enriched in light jets and contains also a contribution of jets from pileup interactions. The events are required to have at least one AK4 jet with $p_T > 40$ GeV and pass the trigger requirements. The data are compared to simulated multijet events using jets with $50 < p_T < 250$ GeV.
- **Muon-enriched jet sample:** For this sample the presence of a muon is required and therefore this topology is dominated by jets containing heavy-flavour hadrons. Events are considered if they satisfy an online selection with at least two AK4 jets with $p_T > 40$ GeV of which at least one contains a muon with $p_T > 5$ GeV. A sample of jets with $50 < p_T < 250$ GeV that contain a muon selected from simulated muon-enriched multijet events is compared with the data.
- **Dilepton $t\bar{t}$ sample:** In this event sample we expect an enrichment in b jets from top quark decays. At trigger level, events are required to have at least one isolated muon or electron. Offline, as expected for leptonic W boson decays, the leading muon and electron are required to have $p_T > 25$ GeV and to be isolated. Events are further considered if they contain at least two AK4 jets with $p_T > 20$ GeV. In this topology there is also a small contribution from jets from pileup interactions due to the relatively low threshold on jet p_T.
- **Single-lepton $t\bar{t}$ sample:** In this event sample a higher fraction of c jets is expected that comes from the decay of the W boson to quarks. At trigger level the events are selected if there is a presence of at least one isolated electron or muon. Offline, exactly one isolated electron or muon is required that satisfies tight identification criteria. The electron (muon) is required to have a $p_T > 40(30)$ GeV and $|\eta| < 2.4$.

3. Heavy Flavour jets discrimination variables

Heavy-flavour jet identification algorithms use variables associated with the properties of heavy-flavour hadrons that are present in jets. Due to the long lifetime of the b hadron a displacement of the tracks of the order of a few mm to 1 cm can occur, from which a secondary vertex (SV) may be reconstructed. The secondary vertex is directly related to the mass of the heavy-flavour

hadron and thus it is a powerful discriminating variable. The upper right panel of figure 1 shows the corrected secondary vertex mass, for the leading secondary vertex, using jets in an inclusive multijet sample. Another useful discriminating variable is the "massVertexEnergyFraction" which is shown in figure 1 on the bottom right panel. The bottom left panel of figure 1 shows the 3D flight distance significance of the leading secondary vertex using jets from the muon-enriched jet sample. The flight distance significance is defined as the 2D or 3D distance between the primary and secondary vertex positions divided by the uncertainty on the secondary vertex flight . The disagreement between the data and the simulation is related to the sensitivity of this variable to the tracker alignment and the uncertainty in the track parameters and hence on the secondary vertex position.

The impact parameter (IP) is defined as the distance between the primary vertex and the tracks at their points of closest approach and it is used to characterize the displacement of the tracks with respect to the primary vertex (PV). The Impact parameter can be defined in three spatial dimensions (3D) or in the plane transverse to the beam line (2D). The impact parameter significance is defined as the impact parameter value divided by its uncertainty (IP/σ). In figure 1 in the top left panel the 3D impact parameter significance of the tracks is shown for jets in the dilepton $t\bar{t}$ sample. The observed discrepancy around zero is explained by the sensitivity of this variable to the tracker alignment and the uncertainty in the track parameters.

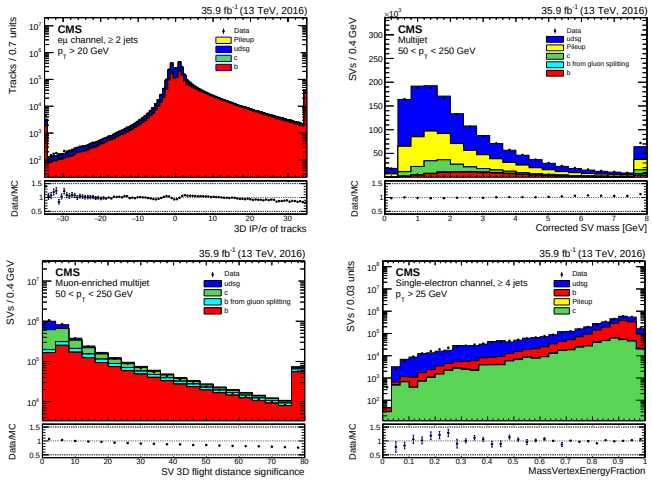

Figure 1: Data-MC comparison for input variables used for heavy flavour tagging for several topologies. The total number of entries is normalized to the observed entries in data

4. Heavy flavour jet identification algorithms

In this section some of the basic heavy flavour jet identification algorithms are discussed.

4.1 The b jet identification

The Jet Probability tagger combines the probabilities from all tracks to originate from the Primary Vertex (PV). This tagger can be used as a reference tagger to taggers that use different input

variables. The CSVv2 tagger combines the information of the displaced tracks with the information of the secondary vertices associated with the jet using multivariate techniques. Compared to CSVv2, the DeepCSV tagger uses a deep neural network with more hidden layers, more nodes per layer, and a simultaneous training in all vertex categories and for all jet flavour. The Soft Lepton (SL) Combined Taggers rely on the presence of a muon or electron from semileptonic b-decays and they can be used as input for a combined tagger such as cMVAv2 . The combined tagger cMVAv2 takes input from other taggers (JP, SL, CSVv2) to perform a MVA training. All the above tagger distributions are shown in figure 2 .The imperfect modeling of the input variables will also have an impact on the modeling of the output discriminator distributions. The upper panel shows the JP and cMVAv2 discriminators using jets in the dilepton $t\bar{t}$ sample whereas the middle panel shows the CSVv2 and DeepCSV discriminators using jets from the muon-enriched sample.

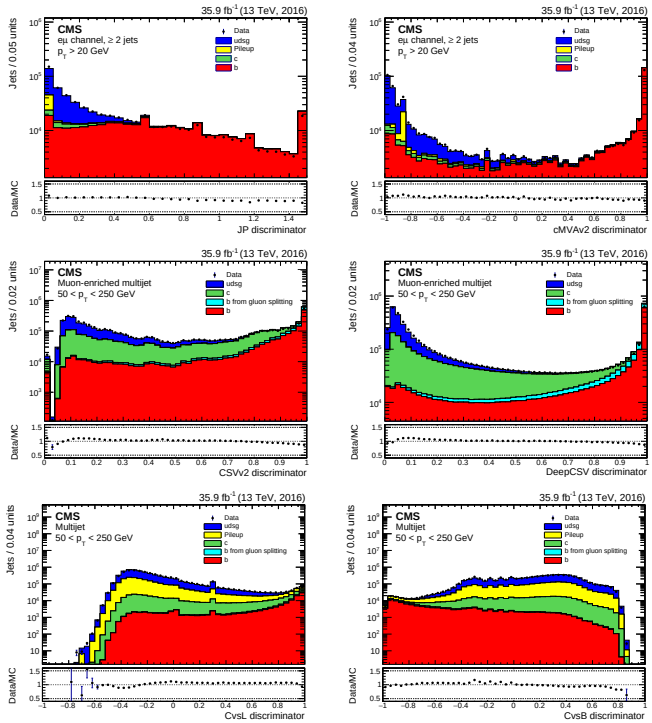

Figure 2: Data-MC comparison of discriminator distributions for several topologies. The total number of entries is normalized to the observed entries in data.

4.2 The c jet identification

The identification of the jets originating from c quarks can be challenging due to fact that the lifetime of the c hadrons is shorter compared to the b hadrons. Same as the b tagging algorithms, c tagging algorithms uses properties related to displaced tracks, secondary vertices, and the presence of soft leptons inside the jets. Two trainings were performed. The first one is for discriminating c jets from light-flavour jets (CvsL) and the other is for discriminating c jets from b jets (CvsB). Both distributions are shown for the inclusive multijet sample in the bottom left and right panel of figure 2 .The discontinuities in both distributions arise from jets for which no tracks pass the track selection criteria.

4.3 Identification of b jets in boosted topologies

Particles decaying to b quarks can be produced with a large Lorentz boost at the high centre-of-mass energy of the LHC. In this case, overlapping jets are reconstructed with a distance parameter of R = 0.8 (AK8) and b tagging techniques can be applied either on the AK8 jet or on its subjets. In both cases the CSVv2 algorithm is used. A new dedicated double-b tagger is developed for boosted jets with double b content. For instance, when a boosted Higgs decays into two b quarks the standard Ak8 b tagging techniques has limitations in identifying $H \to b\bar{b}$ jets . The "*doubleb*" tagger exploits the correlation between the directions of the momenta of the two b hadrons giving rise to a new tagging approach.

To compare data and simulation, two samples were used to collect jets in boosted topologies. The muon-enriched boosted subjets sample, where a muon is required from a combination of single-jet (AK4 and AK8) triggers from multijet muon-enriched samples and the double-muon-tagged boosted jet sample where an extra dijet trigger requirement is added to the previous ones requiring a muon in each of the two jets. In figure 3 in the left panel the CSVv2 discriminator for muon-tagged subjets of AK8 jets with pT > 350 GeV is shown. The agreement is reasonable, with variations of up to 20%. The right part of figure 3 shows the double-b discriminator for double-muon-tagged AK8 jets with pT > 250 GeV.

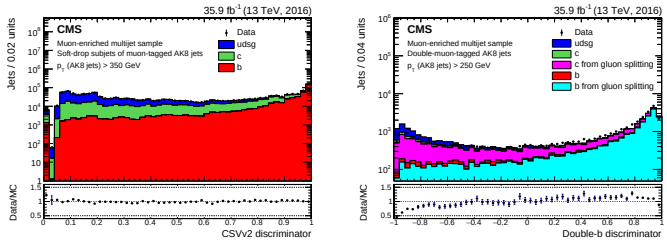

Figure 3: Data-MC comparison for input variables used for heavy flavour tagging for boosted topologies. The total number of entries is normalized to the observed entries in data

5. Conclusions

In CMS experiment a variety of discriminating algorithms and variables are used in order to identify heavy flavour jets in pp collisions at 13 TeV. The performance of these heavy-flavour jet identification algorithms has been validated by comparing different simulated samples with the data that were collected by the CMS detector in 2016, for various event topologies enriched in heavy- or light-flavour jets. All in all, there is a good agreement between data and MC.

References

[1] CMS Collaboration, Identification of heavy-flavour jets with the CMS detector in pp collisions at 13 TeV JINST 13 (2018) no.05, P05011