Identification of b-quark jets in the CMS experiment

Ivan Marchesini on behalf of the CMS Collaboration[*]

[†]*University of Hamburg, Germany*

E-mail: `ivan.marchesini@desy.de`

The CMS experiment investigates high-energetic proton-proton collisions at the LHC, in order to broaden the knowledge of the Standard Model of particle physics and to discover possible new physics beyond the SM.

A large fraction of the CMS physics program relies on the identification of jets containing the decay of a B hadron. The b-jets can be discriminated from jets produced by the hadronization of light quarks based on characteristic properties of B hadrons, such as the long lifetime or the presence of soft leptons produced during their decay.

The CMS detector, with its excellent tracking system, robust lepton identification and finely segmented calorimeters, is well suited to the task of identification of b-jets. An overview of the large variety of b-jet identification algorithms and the measurement of their performance with data collected in 2012 is presented. The algorithms described are based either on the identification of tracks displaced from the primary proton-proton collision, the reconstruction of secondary vertices or the presence of soft leptons inside jets. Some algorithms combine this information using multivariate techniques.

---

[*]Speaker.

[†]A footnote may follow.

1. Introduction

A wide range of physics processes investigated at the LHC is characterized by jets arising from the hadronization of bottom quarks (b-jets) and the ability to identify b-jets (b-tagging) is a fundamental prerequisite for several analyses. Examples include the decay of top quarks, the Higgs boson and various particles predicted by new theories such as Supersymmetry.

Several B hadron properties can be exploited for b-tagging [2]: the long life-time ($\sim$1.5 ps), the high mass ($\sim$5.2 GeV) and the large charged decay multiplicity ($\sim$5). The decays are characterized by hard momentum spectra and by a large semileptonic branching ratio, with the production of electrons and muons in approximately 40% of decays.

The CMS detector is well understood and a good agreement between data and simulations is observed. However, b-tagging is a complex tool and it is mandatory to measure the b-tagging performance in data.

2. Algorithms

The inputs to b-tagging are jets, tracks and primary vertices. The jets are reconstructed using the anti-k_T clustering algorithm [3], with a distance parameter of 0.5. The track reconstruction is based on a Kalman Filter [4]. Tracks are associated to jets if they are closer than $\Delta R < 0.5$ or $\Delta R < 0.3$ from the jet axis, depending on the algorithm. The primary vertex (PV) is reconstructed from the tracks compatible with the beam spot using the Adaptive Vertex Fitter algorithm [5]. Using these observables, the b-tagging algorithms calculate single discriminator values for each jet. Loose (L), medium (M) and tight (T) operating points are defined, corresponding to thresholds on the discriminators yielding a misidentification probability for light-parton jets of about 10%, 1% and 0.1%, respectively, with an average jet-p_T of about 80 GeV/c.

The main observables for b-tagging are the impact parameter (IP) of tracks and the secondary vertex (SV) [1]. The IP is defined as the distance between the track and the PV at the point of closest approach. As for the PV, an Adaptive Vertex Fitter is used also to identify secondary vertices from B hadron decays.

The Track Counting (TC) algorithm sorts the jet-tracks by decreasing values of the IP significance, IP/σ_{IP}. In the High Efficiency (TCHE) version of the tagger the discriminator is given by the IP significance of the second ranked track, while the third one is used by the High Purity version (TCHP). The Jet Probability (JP) algorithm uses an estimate of the likelihood that all jet-tracks come from the PV. The Jet B Probability (JBP) algorithm is defined in a similar way, but more weight is given to the tracks with the highest IP significance up to a maximum of four such tracks.

The Simple Secondary Vertex (SSV) algorithm uses as discriminator the significance of the SV flight distance with respect to the PV. The High Efficiency version (SSVHE) considers only vertices with at least two associated tracks, while the High Purity version (SSVHP) requires three associated tracks. The overall best performing tagger is the Combined Secondary Vertex (CSV), which combines SV observables with track-based lifetime information, providing discrimination even when no SV is found. The performance of the taggers is summarized in Fig. 1 (left).
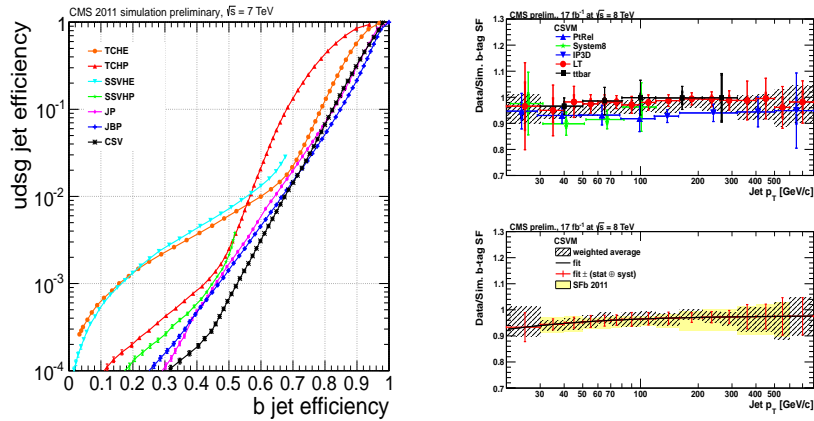
Figure 1: Left: Light-flavor jets misidentification probability as a function of the b-tagging efficiency, for several b-tagging algorithms in simulation. Right: Individual (top) and combined (bottom) measurements of the b-tagging scale factors for the medium operating point of the CSV tagger.

3. Performance Measurements

Several methods have been developed to measure the b-tagging performance in data, based on different samples. An excellent channel is given by $t\bar{t}$ events, with at least one leptonic top-quark decay. This sample is b-enriched, due to the almost exclusive top-quark decay into bW, and the isolated lepton from the W decay allows an efficient background rejection. Inclusive jet samples are also used, triggered by the presence of one or two high-p_T jets, which can be enriched in b-jets requiring a jet containing a muon. An additional b-tagged jet in the event can also be required, further enriching the sample in $b\bar{b}$. The misidentification rates are determined using negative taggers, based on tracks with a negative impact parameter or secondary vertices with a negative decay length. The ratio between the efficiencies in data and simulation is called Scale Factor (SF). Scale factors are close to unity, as shown for the CSV tagger in Fig. 1 (right), with maximum deviations of 10% (25%) for b-tagging efficiencies (misidentification probabilities).

4. Conclusions

The CMS collaboration has developed a variety of b-tagging algorithms, whose performance has been quantified using data collected in 2012 with an integrated luminosity of 19.8 fb^{-1} at $\sqrt{s} = 8$ TeV. Scale factor corrections to simulations are determined, which are close to unity.

References

[1] S. Chatrchyan *et al.* [CMS Collaboration], JINST **8**, P04013 (2013) [arXiv:1211.4462 [hep-ex]].

[2] J. Beringer *et al.* [Particle Data Group Collaboration], Phys. Rev. D **86**, 010001 (2012).

[3] M. Cacciari, G. P. Salam and G. Soyez, JHEP **0804**, 063 (2008) [arXiv:0802.1189 [hep-ph]].

[4] V. Khachatryan *et al.* [CMS Collaboration], Eur. Phys. J. C **70**, 1165 (2010) [arXiv:1007.1988 [physics.ins-det]].

[5] R. Fruhwirth, W. Waltenberger and P. Vanlaer, J. Phys. G **34**, N343 (2007).