Available online at www.sciencedirect.com
ScienceDirect
Nuclear Physics B Proceedings Supplement 00 (2014) 1–3
Nuclear Physics B Proceedings Supplement

b jet Identification in CMS

Camille Beluffi

Centre for Cosmology, Particle Physics and Phenomenology (CP3), Universite catholique de Louvain
Chemin du Cyclotron, 2, B-1348, Louvain-la-Neuve, Belgium

Abstract

A large fraction of the CMS physics program relies on the identification (tagging) of jets containing the decay of a B hadron (b jets). 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. An overview of the large variety of b-tagging algorithms and the measurement of their performance with data collected in 2011 and 2012 are presented in this paper. A special focus lies on new methods of b-tagging in jet substructure. Searches for new physics often focus on boosted final states characterized by particles with large transverse momenta, resulting in decay products of heavy particles tending to be collimated and reconstructed as a single jet, known as fat jet. In this case, the reconstruction of the fat jet substructure is necessary to identify the particle initiating the fat jet. The substructure reconstruction can significantly be improved by the identification of b jets.

1. Introduction

Identification of jets arising from bottom-quark hadronisation and decay (b-tagging [1][2]) is used in many physics analyses to perform precise measurements of the standard model (SM) and for new particle searches. New physics signatures with b jets in the final states are expected at high mass, where the b quarks might end up in boosted topologies with overlapping jets from top-quark or Higgs boson decays, making b-tagging more challenging. The Compact Muon Solenoid (CMS)[3] detector recorded proton-proton collisions occurring at the LHC during the 2012 data taking. With its precise charged particle tracking system and robust lepton identification, it is well matched to the task of b-jet identification. In this paper, the different b-tagging algorithms developed and used in CMS are described in section 2. Then the performance measurements are presented in section 3. Finally, b-tagging in boosted topologies is discussed in section 4.

Email address: camille.beluffi@cern.ch (Camille Beluffi)

2. Algorithms and discriminators for b-tagging

The hadronization of a b quark produces a B hadron which propagates a measurable distance before decaying. Such behavior leads to special properties of the arising b jet, like the presence of an inner displaced secondary vertex with a flying distance higher than its resolution. Tracks coming from a secondary vertex have a large impact parameter that can also be used to identify b jets. Besides, in 20% of cases, a b jet will contain a lepton coming from the semi-leptonic decay of the B hadron. These features are used to build taggers, yielding a single discriminator value for each jet. To analyse the 2012 dataset, three taggers were used:

-**Combined Secondary Vertex (CSV)**: secondary vertex and track-based lifetime information are used to build a likelihood discriminator.

-**Jet Probability (JP)**: the jet is assigned a likelihood

estimation that all associated tracks come from the primary vertex.
-**Track Counting High Purity (TCHP)**: it is based on the third track with the highest impact parameter nifficance.
Three operation points, corresponding to a fixed misidentification probability (P) for light partons, are defined: "Loose" (P=10%), "Medium" (P=1%) and Tight (P=0.1%).

3. Performance measurement

In order to use the b-tagging algorithms in physics analyses, the performance of each algorithm has to be calibrated in data. Many methods have been developed in CMS to measure the b-jet tagging efficiency and the misidentification probability to tag a light-parton jet as a b jet. Three samples of events are used: inclusive jet samples, muon-enriched jet samples and enriched $t\bar{t}$ samples.

3.1. b-tagging efficiency measurement

The b-tagging efficiency is measured in data using several methods applied to multijet events and $t\bar{t}$ events. The efficiency ϵ measured in data is compared with the identification efficiency for b jets in the simulation, resulting in a data/MC scale factor: $SF_b = \epsilon_b^{data}/\epsilon_b^{MC}$.

3.1.1. Measurement in multijet events

The PtRel, IP3D and LT methods use a sample of jets enriched in heavy flavour content by requiring a soft muon within the jet (muon-jet). The fraction of b jets in the selected sample is estimated by fitting the data distribution of a discriminant variable (the transverse momentum of the muon relative to the jet axis (PtRel), the 3D impact parameter of the muon track (IP3D), the discriminator distribution of another tagger (LT)). The b-tagging efficiency in data is measured by estimating the number of b jets in the muon-jet sample by the fit, then repeating the fit on the subsample of muon-jets passing the tagging requirement. The efficiency is the ratio between these two values. The System8 method uses two weakly correlated taggers, one of which is the one to be probed. They are tested in two samples with different b-quark enrichement.
Various systematic uncertainties are considered, among them the pileup description, the rate of gluon splitting into b-quark pairs, the muon p_T spectrum and b-fragmentation modelling, and the description of the relative direction of the muon with respect to the jet.

tagger	SF_b in muon-jets	SF_b in $t\bar{t}$ events
JPL	0.982 ± 0.020	0.966 ± 0.015
CSVL	0.983 ± 0.017	0.987± 0.018
JPM	0.947 ± 0.034	0.961± 0.012
CSVM	0.951 ± 0.024	0.953 ± 0.012
TCHPT	0.896 ± 0.035	0.921 ± 0.010
JPT	0.866± 0.036	0.922± 0.017
CSVT	0.916 ± 0.032	0.926 ± 0.036

Table 1: Scale factors SF_b obtained in muon-jet data and $t\bar{t}$ data, averaged over the p_T spectrum of jets from top decays. The overall uncertainties are given.

3.1.2. Measurement in $t\bar{t}$ events

Both lepton+jets and dileptonic final states are used. In the lepton+jets channel, the flavour tag consistency (FTC) method and the bSample method are used. In the dilepton channel, the flavour tag matching (FTM) method is used as well as the LT method, which can be applied on the same events. The FTC method (FTM method) requires consistency between the observed and expected number of tags in the lepton+jets (dilepton) events. A log-likelihood fit is performed with the b jet tagging efficiency as free variable.
In the bSample method, the b-jet tagging efficiency is measured from two subsamples, one enriched in b jets and the other depleted, based on the transverse mass of the muon and jet from the leptonically decaying top. Efficiencies are derived from the difference between the discriminator distributions in the two subsamples.
The main sources of systematic uncertainties are the jet-parton matching, the definition of the renormalization and factorization scales, the choice of the parton distribution function, the pileup description and jet energy scale and jet energy resolution.

3.1.3. Combination of b-tagging efficiency measurements

The combination is based on a weighted mean of the different scale factor measurements, taking into account correlated and uncorrelated uncertainties and evaluating the shared fraction of events between the different methods. Table 1 compares the combined scale factors SF_b measured in multijet and $t\bar{t}$ events, averaged over the p_T spectrum of jets from top decays.

3.2. Misidentification probability measurement

The probability of light-flavour quark and gluon jets being misidentified as b jets is evaluated with negative taggers, which are identical to the default algorithms, except that they use only tracks with negative impact

tagger	SF_{light}
JPL	1.03 ± 0.01 ± 0.07
CSVL	1.10 ± 0.01 ± 0.05
JPM	1.10 ± 0.02 ± 0.20
CSVM	1.17 ± 0.02 ± 0.15
TCHPT	1.27 ± 0.06 ± 0.27
JPT	1.11 ± 0.07 ± 0.31
CSVT	1.26 ± 0.07 ± 0.28

Table 2: Data/MC scale factors SF_{light} for different algorithms and operating points for jet p_T in the range [80-120] GeV/c. Both statistical and systematic uncertainties are quoted.

paramter values or secondary vertices with negative decay lengths. The discriminator values for negative and positive taggers are expected to be symmetric for light-parton jets by resolution effect. We can therefore derive the misidentification probability ϵ^{misid} from the rate of negative-tagged jets ϵ^{-} in inclusive jet data. A correction factor, $R_{light} = \epsilon^{misid}_{MC}/\epsilon^{-}_{MC}$, is evaluated from the simulation in order to correct for second-order asymmetries in the negative and positive tag rates of light-flavour quark and gluon jets, and for the heavy flavour contribution to the negative tags: $\epsilon^{misid}_{data} = \epsilon^{-}_{data} \times R_{light}$. The data/MC scale factors of the misidentification probabilities, $SF_{light} = \epsilon^{misid}_{data}/\epsilon^{misid}_{MC}$, are given in Table 2.

4. b-tagging in boosted topologies

High-mass resonances with a final state containing b quarks are predicted by various models of new physics. They may decay into top-quark pairs or Higgs bosons, and if they have a large enough momentum ("boosted toplogies"), their decay products are very collimated, resulting in a small angular distance ΔR between them, and ending up clustered in a single fat jet. Boosted topologies are usually reconstructed and interpreted using jet substructure reconstruction methods such as top/W/Z-tagging algorithms[4]. Algorithms of b-tagging in the jet substructure can significantly improve the sensitivity of these methods.

4.1. b-tagging in jet substructure

One important reconstruction parameter is the size of the jet, which needs to be optimised to include all decay products, depending on the jet p_T. Two cases have been studied in detail: for top-tagging, the use of the HEPTopTagger[5] algorithm, which is based on Cambridge/Aachen jets of size R = 1.5 (CA15), is investigated. The fat-jet substructure is identified by undoing the CA algorithm clustering. For Higgs-tagging, the focus is on CA jets of size R=0.8 and the jet substructure

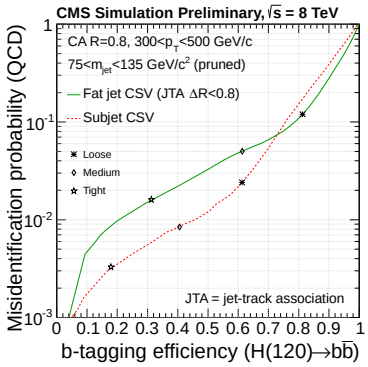

Figure 1: Misidentification probability as a function of b-tagging efficiency for boosted $H \rightarrow b\bar{b}$ jets and inclusive QCD jets for the CSV algorithm applied to fat jets and pruned subjets for fat jets with (300< p_T <500 GeV/c).

is described by pruned jets. Algorithms of b-tagging can then be applied on the fat jet or on its substructure components, the second option giving the best performance (see Fig.1).

4.2. Performance measurement

Measurement of b-tagging efficiency in boosted topologies is challenging, and needs specific treatment since results on standard jets are not necessarily applicable to boosted objects. For Higgs-tagging, efficiency is measured using LT method on different control samples to study the performance of b-tagging both on fat jets and subjets. The agreement found between data and simulation is compatible with what is observed in non boosted topologies. A modified implementation of the FTC method has been developed to measure the b-tagging efficiency in boosted top-quark events and results show that the simulation reproduces the b-tagging efficiencies in data equally well in boosted and in non-boosted top-quark events.

References

[1] CMS Collaboration, "Identification of b-quark jets with the CMS experiment", JINST 8 (2013) P04013, doi:10.1088/1748-0221/8/04/P04013, arXiv:1211.4462

[2] CMS Collaboration, "Performance of b tagging at $\sqrt{s}$ = 8 TeV in multijet, $t\bar{t}$ and boosted topology events", CMS PAS BTV-13-001

[3] CMS Collaboration, "The CMS experiment at the CERN LHC", JINST 3 (2008) S08004, doi:10.1088/1748-0221/3/08/S08004.

[4] A. Altheimer et al., "Jet Substructure at the Tevatron and LHC: New results, new tools, new benchmarks", J. Phys. G 39 (2012) 063001, doi:10.1088/0954-3899/39/6/063001, arXiv:1201.0008.

[5] T. Plehn and M. Spannowsky, "Top Tagging", J. Phys. G 39 (2012) 083001, doi:10.1088/0954-3899/39/8/083001, arXiv:1112.4441.