Identification of b-quark jets in the CMS experiment

Sudhir Malik[1]
University of Nebraska-Lincoln
Lincoln, NE 68588
E-mail: malik@fnal.gov

The identification of jets arising from the production of b-quarks is an essential tool both for the measurement of standard-model processes and in the search for physics beyond this model at the Large Hadron Collider (LHC). The CMS (Compact Muon Solenoid) experiment has developed a variety of algorithms that use the impact parameters of charged-particle tracks, the properties of reconstructed decay vertices, the presence of a lepton or combinations of these quantities to discriminate between b- and light flavor jets. Proton-proton collisions recorded in 2011 and corresponding to an integrated luminosity of 5.0 fb^{-1} have been used to compare the quality of the reconstruction with expectations from simulation. The performance of the algorithms in terms of efficiency and misidentification probability has been measured from multi jet events and from top-quark pair events.

36th International Conference on High Energy Physics
July 4-11, 2012
Melbourne, Australia

1

PoS(ICHEP2012)311

1. Introduction

The method to identify jets originating from the fragmentation and hadronisation of the b-quarks is called b-tagging. Known particles of the standard model such as the top quark as well as particles predicted by a variety of models beyond the standard model, such as SUSY involve decays into b-quarks. The CMS detector [1] has a precise charged particle tracking and robust lepton identification that is well suited for b-tagging. It has been applied successfully in a large number of physics analyses in CMS and has become an indispensable tool for the reduction background processes. Several unique b-jet properties have been exploited in algorithms [2] for b-tagging at CMS. They use impact parameters (IP) of charged-particle tracks, the properties of reconstructed decay vertices, the presence of a lepton or combinations of these quantities to select samples of jets with different b purities. In this paper we describe the performance of these algorithms and the corresponding discriminators, measurements of b-tagging efficiency and the misidentification probability. The data used in this study were recorded in proton-proton collisions in 2011 at the center of mass energy of 7 TeV for a total integrated luminosity of 5.0 fb^{-1}. CMS has achieved a b-tagging efficiency of 85% for a light-parton misidentification probability of 10%. For analyses requiring higher purity, a misidentification probability of only 1.5% has been achieved, for a 70% b-jet tagging efficiency. The efficiency for b-tagging has been measured in events from multijet and $t\bar{t}$ pair production.

2. Algorithms and discriminators for b-jet identification

All b-tagging algorithms calculate a number called a discriminator for each jet based on its properties. The higher the discriminator value of a jet, the higher is the likelihood of it being a b-jet. The minimum thresholds on these discriminators define Loose ("L"), Medium ("M"), and Tight ("T") operating points with a misidentification probability for light-parton jets of close to 10%, 1%, and 0.1%, respectively, at an average jet p_T of about 80 *GeV/c*. A variety of reconstructed objects like tracks, vertices and identified leptons, are used to build this discriminator that helps to distinguish b-jets from the light-flavor jets. The particles reconstructed using a particle-flow algorithm [3] are clustered into jets using the anti-k_T algorithm [4] with a size parameter of 0.5. Each b-tagging algorithm uses properties of charged particles in a jet, including identified leptons and requires tracks of high purity [5]. The IP (Fig.1), defined as the distance of closest approach between the track and the Primary Vertex (PV), can be used to distinguish between decay products of a B-hadron and prompt tracks. It is calculated in 3-dimensions utilizing the excellent resolution of the pixel detector.

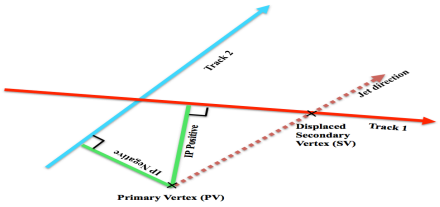

Figure 1: Positive and Negative IP

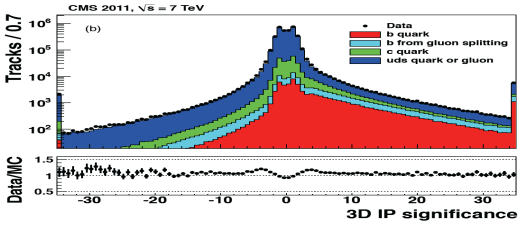

Figure 2: Distribution of 3D IP S_{IP} for all selected tracks

The IP is positive (negative) if the track is produced downstream (upstream) with respect to the PV along the jet direction. Since it depends strongly on the p_T and η of a track and to take into the account the effect of resolution, a ratio of the IP and it's estimated uncertainty, called IP significance (S_{IP}) is used as a discriminating variable in b-tagging algorithms. The distribution of S_{IP} is shown in Fig. 2

2.1 Identification using track impact parameter

The Track Counting (TC) algorithm sorts tracks in a jet by decreasing values of S_{IP}. Although the ranking tends to bias the values for the first track to high positive IP significances, the probability to have several tracks with high positive values is low for light-parton jets. Therefore the two different versions of this algorithm, Track Counting High Efficiency (TCHE) and Track Counting High Purity (TCHP), use the S_{IP} of the second and third ranked track as the discriminator value, respectively. A natural extension of the TC algorithms is the combination of the IP information of several tracks in a jet. The Jet Probability (JP) algorithm uses an estimate of the probability that all tracks associated to the jet come from the PV. The Jet B Probability (JBP) algorithm (discriminator distribution in Fig. 3 (Left)) gives more weight to tracks with the highest S_{IP} (up to four tracks).

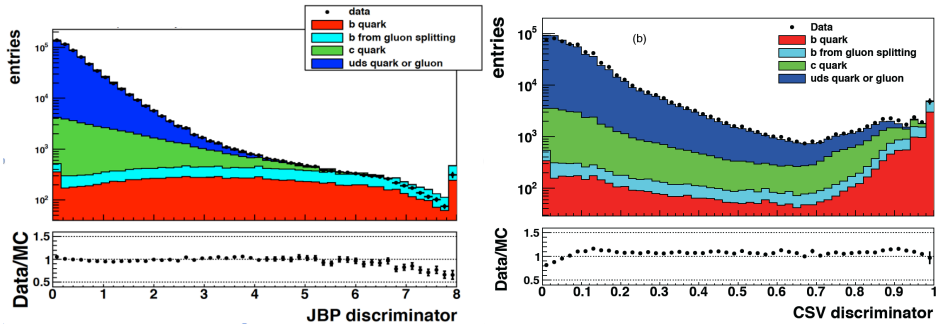

Figure 3: Distribution of (Left) JBP and (Right) CSV discriminator

2.2 Identification using secondary vertices

SV tagging algorithms rely on the reconstruction of at least one Secondary Vertex (SV). The significance of the 3D flight distance is used as a discriminating variable. To enhance the b purity SV candidates must share less than 65% of their associated tracks with the primary vertex and the significance of the radial distance between the two vertices has to exceed 3σ. SV candidates with a radial distance of more than 2.5 cm with respect to the primary vertex, with masses compatible with the mass of K^0 or exceeding 6.5 GeV/c^2 are rejected, reducing the contamination by vertices corresponding to the interactions of particles with the detector material and by decays of long-lived mesons. The flight direction of each candidate also has to be within a cone of $\Delta R < 0.5$ around the jet direction. Simple SV (SSV) algorithms use the significance of the flight distance (ratio of the flight distance to its estimated uncertainty). If several vertices are present, the one with smallest distance error is used. Its two versions are SSV High Efficiency (SSVHE) and SSV High Purity (SSVHP). SSVHE uses vertices with at least two associated tracks whereas SSVHE requires at least three tracks. The Combined SV

(CSV) algorithm combines information from vertices with track-based lifetime (LT) information. In many cases, tracks with an $S_{IP} > 2$ can be combined in a "pseudo vertex", allowing for the computation of a subset of secondary-vertex-based quantities even without an actual vertex fit. When even this is not possible, a "no vertex" category reverts to track-based variables that are combined in a way similar to that of the JP algorithm. This has the advantage that the algorithm works even in case no vertices are reconstructed. Fig 3 (Right) shows the distribution of the CSV discriminator. The performance of b-tagging algorithms is summarized in Fig. 4

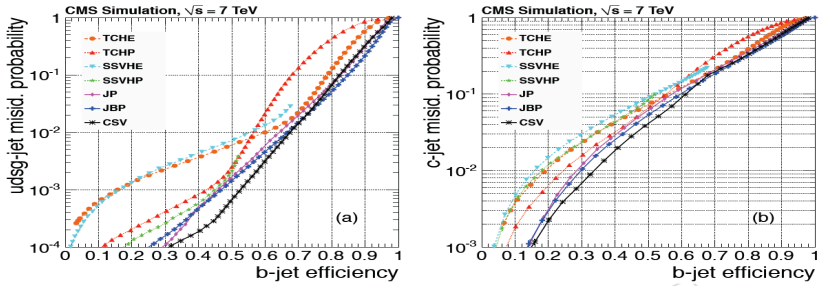

Figure 4: Performance curves obtained from simulation for (Left) light-jets and (Right) c-jet misidentification probabilities as a function of the b-jet efficiency.

3. Efficiency measurements

Despite a good agreement between Monte Carlo (MC) simulation and data, it is essential to measure the b-tagging efficiency in situ. If MC and data match well, one can use simulation for a wide range of topologies after applying corrections determined from specific data samples. A scale factor, SF_b , defined as the ratio of efficiency measured with collision data to the efficiency found in equivalent simulated samples, can be used to apply corrections to simulated events. The efficiency for tagging b-quark jets has been measured in events from multijet [6] and t-quark pair production [7]. Some efficiency measurements are performed using samples that include a jet with a muon within $\Delta R = 0.4$ from the jet axis (a "muon jet"). Muons are identified very efficiently in the CMS detector, making it straightforward to collect samples of jets with at least one muon.

Due to the large b-quark mass, the momentum component of the muon transverse to the jet axis, p_T^{rel} , is larger for muons from b-hadron decays than for muons in light-parton jets or from charm hadrons. This component is used as the discriminant for the "PtRel" method. In addition, the impact parameter of the muon track, calculated in three dimensions, is also larger for b- hadrons than for other hadrons. This parameter is used as the discriminant for the "IP3D" method. Both of these variables can thus be used for measuring the b-flavor content in a given sample. A dijet sample with high b-jet purity is obtained by requiring that events have exactly two reconstructed jets: the muon jet as defined above and another jet fulfilling the TCHP (third ranked track) b-tagging criterion at "Medium" b-tagging efficiency working point (TCHPM). Muon jets are separated into a tagged (N_{data}^{tag}) and untagged (N_{data}^{untag}) subsample by a discriminator working point whose efficiency is to be measured. For the two subsamples

separately, the spectra of muon jets p_T^{rel} or IP3D (Fig 5) is fitted using templates of b, c and udsg jets derived from simulation or inclusive jet data. From each fit the fractions of b jets (tagged and untagged) is extracted from the data. The following equation is used to calculate the efficiency.

$$\varepsilon_b^{tag} = \frac{f_b^{tag} . N_{data}^{tag}}{f_b^{tag} . N_{data}^{tag} + f_b^{untag} . N_{data}^{untag}}$$

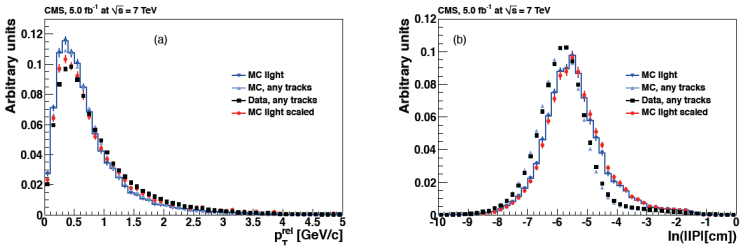

Figure 5: Comparison of distributions of (Left) muon p_T^{rel} for jets with p_T between 80 and 120 GeV/c and (Right) ln(|IP|[cm]) for jets with p_T between 160 and 320 GeV/c

The System8 method is applied to events with a muon jet and at least one other, "away-tag", jet. The muon jet is used as a probe. The reference lifetime tagger and a supplementary p_T^{rel} based selection are tested on this jet. The away-tag jet is tested with a separate lifetime tagger TCHP with "Loose" working point (TCHPL). There are eight quantities that can be counted from the full data sample. The quantities depend on the number of passing or failing tags. A set of equations correlates these eight quantities with the tagging efficiencies.

The method of using JP tagger as a reference can give the fraction of b jets in a data sample and also in a subsample that has been selected by an independent tagging algorithm. To calculate the efficiency one acquires the fraction (C_b) of b jets with JP information (before tagging) in the MC and fits the JP distribution by templates for b,c and udsg jets to get the fraction of b jets (f_b^{tag}). The efficiency is the ratio of $C_b . f_b^{tag} . N_{data}^{tag}$ and $f_b^{beforetag} . N_{data}^{beforetag}$.

Several systematic uncertainties affect the measurement of the b-jet tagging efficiency. While some are common to all the above methods, some are common to a subset and some unique to a particular method. For "PtRel" specific method, since the p_T^{rel} distribution in data is fitted with a sum of templates for b jets and for c+udsg jets, uncertainties on the ratio between light-parton and charmed jets in the simulation must be considered. For the System8 method one uncertainty comes from the selection on the muon $p_T^{rel} > 0.8$. The common uncertainties are pile-up, gluon splitting and p_T^{μ}. To calculate pileup uncertainty the average value of the pile-up in data is varied by $\pm$ 10%.

The top quark decays to W boson and a b-quark 99.8% of the time. The measurement of the heavy-flavor content of $t\bar{t}$ events can provide either a direct measurement of the branching fraction of the decay of the top quark to a W boson and a b quark, B(t $\rightarrow$ Wb), or, assuming B(t $\rightarrow$ Wb) = 1, the b-jet tagging efficiency.

4. Misidentification measurements with negative taggers

The measurement of the misidentification probability for light-parton jets relies on the definition of inverted tagging algorithms, selecting non-b jets using the same variables and tech niques as the standard versions. These "negative taggers" can be used in the same way as the regular b-jet tagging algorithms both in data and in the simulation. As the negative-tagged jets are enriched in light flavours, the misidentification probability can be measured from data, with the simulation used to extract a correction factor. The misidentification probability is evaluated from tracks with a negative impact parameter or from secondary vertices with a negative decay length. When a negative tagger is applied to jets of any flavor, the corresponding tagging efficiency is denoted "negative tag rate". The misidentification probability is calculated as $\varepsilon_{data}^{misid} = \varepsilon_{data}^{-} . R_{light}$ where ε_{data}^{-} is the negative tag rate measured in jet data and R_{light} is a correction factor equal to $\varepsilon_{MC}^{misid} / \varepsilon_{MC}^{-}$ (ratio of light flavor misidentification probability rate and negative tag rate of correction factor taken from simulation). The negative and positive b-jet tagging discriminator distributions in data are compared to simulation in Fig.6. To compare measured misidentification probability to that

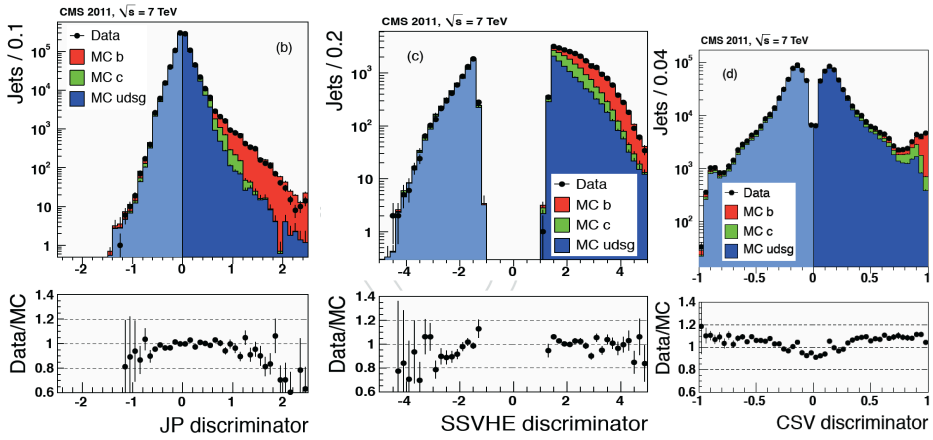

Figure 6: Negative and positive b-jet tagging discriminator distributions in data compared to simulation for light and b-jets

predicted by the simulation, a scale factor SF_{light} is defined as $SF_{light} = \varepsilon_{data}^{misid} / \varepsilon_{MC}^{misid}$.

Systematic uncertainties coming from b and c fractions, gluon fraction, long lived K_s^0 and Λ decays, photon conversion and nuclear interactions, mismeasured tracks, the ratio of the number of negative over positive tagged jets, pile-up, and event sample, are taken into consideration. Some are explained below

- b and c fractions: The fraction of b-flavour jets has been measured in CMS to agree with the simulation within a $\pm 20\%$ uncertainty. A $\pm 20\%$ uncertainty is conservatively estimated for the overall fraction of b and c jets.

- Gluon fraction: This affects both the misidentification probability in simulation and the overall negative tag rates. An uncertainty of ±20% is extracted from the comparison of simulation with data.
- Long lived K_s^0 and Λ decays, photon conversion and nuclear interactions, mismeasured tracks, the ratio of the number of negative over positive tagged jets, pile-up, and event sample, are taken into consideration.
- Pile-up: The misidentification probability depends on the pileup model used in the simulation. The simulated events are reweighted in order to match the pileup rate in the data.
- Event sample: Physics analyses use jets from different event topologies. For a given jet p_T, the misidentification probability is different for the leading jet or if there are other jets with higher p_T values in the same event. Measured misidentification scale factors for leading and subleading jets have a dispersion of about 7%. In addition, misidentification scale factors vary by 2–7%, depending on the tagger, for different running periods. These two uncertainties are added in quadrature to account for an uncertainty due to sample dependence. This is the dominant contribution to the overall systematic uncertainty on the misidentification probability.

5. Results

The combined and parameterized measurements of the ratio of b-tagging efficiencies of the data to that in simulation are shown in Fig. 7 for CSVM and JPL taggers. The "PtRel" and the System8 methods provide precise measurements for the lower part while IP3D and LT methods are designed for higher part of the p_T spectrum. Table 1 (Left) shows data/MC SF_b for the lower p_T range 80 to 120 *GeV/c*. The p_T dependent SF_b measured in multijet events are compared to SF_b from $t\bar{t}$ events in Table 1(Right). It shows that SF_b for muon jets and inclusive jets are compatible.

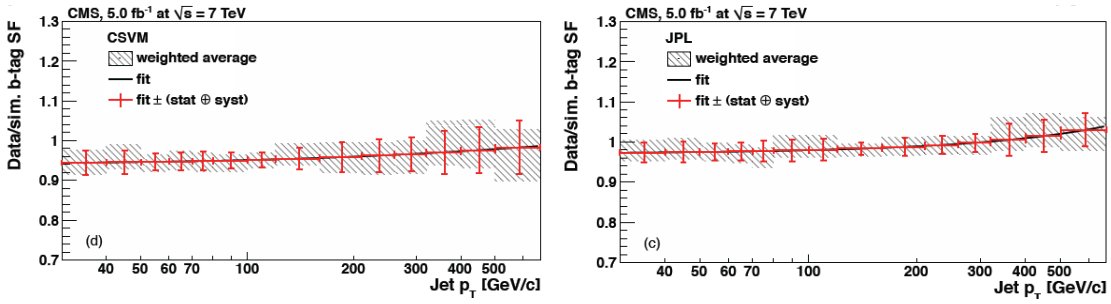

Figure 7: Ratio (Data/MC) of b-tagging efficiencies for CSVM and JPL taggers

b tagger	SF_b (PtRel)	SF_b (System8)	SF_b (LT)	SF_b (comb.)
JPM	$0.90 \pm 0.01 \pm 0.03$	$0.93 \pm 0.03 \pm 0.06$	$0.99 \pm 0.01 \pm 0.05$	0.92 ± 0.03
JBPM	$0.92 \pm 0.01 \pm 0.02$	$0.96 \pm 0.03 \pm 0.08$	$0.99 \pm 0.01 \pm 0.05$	0.91 ± 0.03
TCHEM	$0.94 \pm 0.01 \pm 0.03$	$0.99 \pm 0.03 \pm 0.07$	$0.98 \pm 0.01 \pm 0.03$	0.95 ± 0.02
TCHPM	$0.95 \pm 0.01 \pm 0.03$	$0.94 \pm 0.02 \pm 0.09$	$0.97 \pm 0.01 \pm 0.02$	0.96 ± 0.02
SSVHEM	$0.92 \pm 0.01 \pm 0.02$	$0.92 \pm 0.03 \pm 0.05$	$0.97 \pm 0.01 \pm 0.02$	0.95 ± 0.02
CSVM	$0.93 \pm 0.01 \pm 0.02$	$0.97 \pm 0.03 \pm 0.06$	$0.97 \pm 0.01 \pm 0.03$	0.95 ± 0.02

b tagger	SF_b in multijet events	SF_b in $t\bar{t}$ events
JPM	0.92 ± 0.03	0.95 ± 0.03
JBPM	0.92 ± 0.03	0.94 ± 0.03
TCHEM	0.95 ± 0.03	0.96 ± 0.03
TCHPM	0.94 ± 0.03	0.93 ± 0.03
SSVHEM	0.95 ± 0.03	0.96 ± 0.03
CSVM	0.95 ± 0.03	0.97 ± 0.03

Table 1: (Left) SF_b for $80 < p_T < 120$ GeV/c and (Right) SF_b in multijet and $t\bar{t}$ events

All the methods [2] that use $t\bar{t}$ events, for example profile likelihood ratio, flavor tag consistency method, give efficiency values compatible with "PtRel" and System8 methods and are also consistent with each other.

Table 2 shows the misidentification probabilities and the corresponding data/MC scale factors SF_{light} for different algorithms and for the Medium "M" working point for jet p_T in the range 80 - 120 *GeV/c*. The statistical uncertainties are quoted for the misidentification probabilities, while both the statistical and the systematic uncertainties are given for the scale factors. The first error is statistical and the second error is systematic.

tagger	misidentification probability	SF_{light}
JPM	0.0107 ± 0.0001	$1.03 \pm 0.01 \pm 0.17$
JBPM	0.0110 ± 0.0001	$0.95 \pm 0.01 \pm 0.13$
TCHEM	0.0282 ± 0.0003	$1.21 \pm 0.01 \pm 0.15$
TCHPM	0.0304 ± 0.0003	$1.24 \pm 0.01 \pm 0.13$
SSVHEM	0.0208 ± 0.0002	$0.94 \pm 0.01 \pm 0.08$
CSVM	0.0151 ± 0.0002	$1.11 \pm 0.01 \pm 0.12$

Table 2: Misidentification probabilities and SF_{light} for different algorithms for the Medium "M" working point.

References

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

[2] CMS Collaboration, *Identification of b-quark jets in the CMS experiment*, CERN preprint: CERN-PH-EP-2012-262

[3] CMS Collaboration, *Particle-Flow Event Reconstruction in CMS and Performance for Jets, Taus, and Missing ET*, CMS-PAS-PFT-09-001, (2009).

[4] M. Cacciari, G. P. Salam, and G. Soyez, *The Anti-k(t) jet clustering algorithm*, JHEP 04 1149 (2008) 063, doi:10.1088/1126-6708/2008/04/063, arXiv:0802.1189

[5] CMS Collaboration, *Description and performance of CMS track reconstruction*, CMS-PAS-TRK-11-001, (2011)

[6] CMS Collaboration, *b-Jet Identification in the CMS Experiment*, CMS-PAS-BTV-11-004, (2011)

[7] CMS Collaboration, *Measurement of the b-tagging efficiency using $t\bar{t}$ events*, CMS-PAS-BTV-11-003, (2011)