Measurement of *b*-tagging efficiency and mis-tagging rates with CSIP method.

R.Demina, A.Khanov, F.Rizatdinova, E.Shabalina

April 8, 2004

Abstract

We present a summary of the CSIP (Counting Signed Impact Parameter) algorithm optimization and study of its performance on p14 data and Monte Carlo. The *b*-tagging efficiency on data is measured by different methods for four working points. Good agreement within statistical errors is observed between all measured values. Cross-check of methods used for *b*-tagging efficiency measurement on data is also performed. The scale factor between data and Monte Carlo is found to be flat versus jet E_T and η. It varies from (0.76 ± 0.01) to (0.73 ± 0.01) depending on the chosen working point. The light jet (jet from u,d,s quarks) tagging rate function (LTRF) is measured on data by two different techniques. Inclusive LTRF depends on the E_T and η spectra of jets in the particular data set. LTRF measured on the EMqcd data varies from 0.2% to 1.2% depending on the working point.

Contents

1 Introduction

Detailed description of Counting Signed Impact Parameter (CSIP) algorithm is given in [1]. In short, the method is based on the fact that tracks from *b* and *c*-decays tend to have larger impact parameters compared to tracks from primary vertex. In addition to that, *B*-products have positive projection of their impact parameters on the jet axis. CSIP method uses signed impact parameter significance which provides good separation between tracks from primary vertex and *b* or *c*-decays.

The present Note is an update of the studies of the CSIP performance for the p14 version of the DØ reconstruction software. This reconstruction version has a new tracking code (AA+HTF) which is in particular characterized by a significantly lower fake rate. The CSIP algorithm has been optimized accordingly to profit from the better tracking performance.

2 Data samples

The following data sets were used to determine the CSIP performance:

- For the measurement of *b*-tagging efficiency: muon-in-jet sample from the Common Sample Group (4.7M events). [1] A meduim quality muon with p_T >8 GeV inside a ΔR <0.5 cone around the jet axis is required.
- For the mis-tag rate measurement: jet triggers skim (1.2M events) [2] and part of EMqcd skim (1.2M events). [3] The jettrig sample is based on several purely calorimetric triggers (JT_8TT, JT_15TT, JT_25TT_NG, JT_45TT, JT_65TT, JT_95TT, JT_125TT). There is no further selection requirements, so this sample is basically unbiased. The EMqcd skim requires one EM object (p_T >15 GeV) and at least one jet. For both samples, an additional requirement on transverse missing energy $ME_T < 10$ GeV has been applied.

All samples were reconstructed with d0reco version p14.03.02, applying jet energy scale corrections version 5.1 and T42 calorimeter noise suppression algorithm [2]. For the primary vertex finding, the 2-pass probabilistic algorithm implemented in d0root (d0root_analysis version 9.36 and d0root_btag version 9.44) was used. The samples were converted to the top_tree root tuples using top_analyze version Nefertiti. [4]

For the Monte Carlo studies, the following samples were used:

- *b*-tagging efficiency: $t\bar{t} \to l$+jets, [5] $Z \to b\bar{b}$ where one *b* is forced to decay semileptonically, [6] and $W + b\bar{b}$. [7]
- *c*-tagging efficiency: $Z \to c\bar{c}$ where one *c* is forced to decay semileptonically,[8] and $W + c\bar{c}$.[9]
- mis-tagging rate, flavor composition of QCD: QCD from JES group (p_T^q >20 GeV, 40 GeV, and 80 GeV), [10] QCD from B-id group (p_T^q >40 GeV, and 80 GeV), [11] and W+light jets. [12]

[1] Location: /prj_root/1001/top_write/top_analyzed_data/Nefertiti/MUJETSBID

[2] Location: /work/polk-clued0/elis/Nefertiti/JETTRIG_BID

[3] Location: /prj_root/1001/top_write/top_analyzed_data/Nefertiti/EMQCD/p14.03.02

[4] The complete list of package versions can be found in
http://www-d0.fnal.gov/Run2Physics/top/d0_private/wg/top_analyze/Nefertiti/instructions_nefertiti.html

[5] Location: /rooms/library/top/top_analyzed_mc/Parsifal_Updated/p14.05.01/ttbar/ljets/175

[6] Location: /rooms/attic/work/mcskims/zbbmu

[7] Location: /rooms/library/top/top_analyzed_mc/Parsifal_Updated/p14.02.00/wbb

[8] Location: /rooms/attic/work/mcskims/zccmu

[9] Location: /rooms/library/top/top_analyzed_mc/Parsifal_Updated/p14.02.00/wcc

[10] Location: /rooms/library/top/top_analyzed_mc/Parsifal_Updated/p14.05.01/qcd

[11] Location: /rooms/library/top/top_analyzed_mc/Parsifal_Updated/p14.05.01/qcd

[12] Location: /rooms/library/top/top_analyzed_mc/Parsifal_Updated/p14.03.02/wjjjj

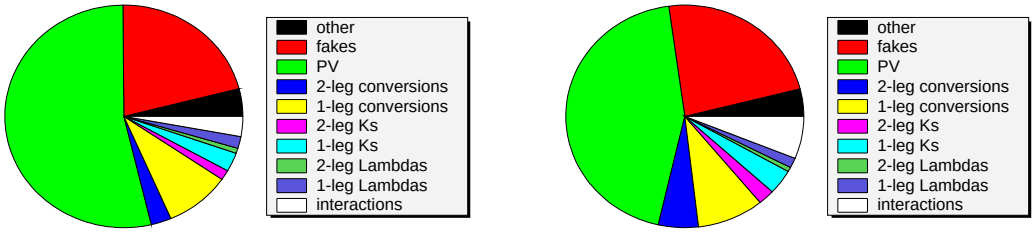

Figure 1: Origin of tracks contributing to the negative (left) and positive (right) mistags in $W + jjjj$ events.

These samples were reconstructed with d0reco versions p14.02.00 through p14.05.01, applying jet energy corrections for Monte Carlo v5.0 , and the same primary vertex algorithm as for the data. The Monte Carlo samples were converted to the top_tree root tuples using top_analyze version Parsifal-Updated. [13]

3 CSIP optimization in p14

The algorithm did not change significantly with respect to the p13 version. A few modifications have been introduced to improve the performance. All of these modifications can be reset to make the algorithm to perform in exactly the same way it was in p13.

3.1 The origin of mistags in Monte Carlo

A study has been performed to figure out the origin of tracks contributing to the negative and positive tag rate in light jets. This study was performed on the W+light jets sample, where events without any b- or c quarks were preselected. The mistag is found to be dominated by the tracks from the primary vertex. The second largest contribution comes from fakes which still constitute about 5% of all tracks. Here a fake is determined as a reconstructed track which cannot be matched to any monte carlo track within 10 σ of all five parameters at the dca point. The sketch of contributions from the different sources of mistags is shown in Fig. 1.

3.2 Treatment of tracks close to the jet axis

In case a track is almost parallel to the jet axis in $x-y$ plane, it is no longer possible to determine the sign of its dca projection onto the jet axis. In an earlier version of the algorithm this problem was ignored, hence the sign was taken arbitrarily. In the new version, a track is counted as a tagging one if either it has correct dca projection sign (plus for positive tags and minus for negative tags), or the angle between track and the jet axis in the $x-y$ plane is below $\Delta\psi$ (these tracks are counted for both positive and begative tags). The two ultimate cases are $\Delta\psi$=0 (reproducing the old behavior) and $\Delta\psi$=π (the sign of the dca projection is ignored). Provided that the tag rate is dominated by 2 tracks $\times$ 3σ mode, and mistag comes mostly from tracks originating from the primary vertex and therefore having symmetric signed dca significance distribution (which is not exactly the case), in the latter ultimate case the mistag rate increases by a factor of 4. The reasonable value of $\Delta\psi$ is expected to be around the jet direction resolution. By looking at the performance curve versus different values of $\Delta\psi$, the default value $\Delta\psi$=0.02 is chosen. In order to further improve the performance, at least one track with $\Delta\psi$ above the cut is required.

[13] The complete list of package versions can be found in
http://www-d0.fnal.gov/Run2Physics/top/d0_private/wg/top_analyze/Parsifal-updated/instructions_parsifal.html

index	0	1	2	3	4	5
i_{χ^2}	$\chi^2 < 3$	$3 < \chi^2 < 9$	$\chi^2 > 9$			
i_{CFT}	$n_{CFT}=0$	$1 \leq n_{CFT} \leq 10$	$11 \leq n_{CFT} \leq 12$	$13 \leq n_{CFT} \leq 14$	$n_{CFT} \geq 15$	
i_{SMT}	$n_{SMT}=0$	$n_{SMT}=1$	$n_{SMT}=2$	$n_{SMT}=3$	$n_{SMT}=4$	$n_{SMT} \geq 5$

Table 1: Types of tracks and their indices.

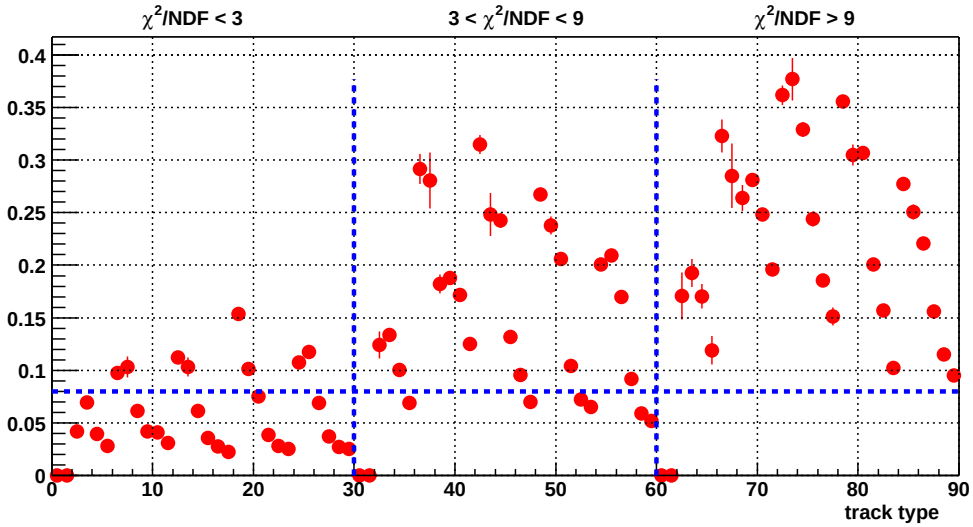

Figure 2: Fraction of tracks with signed dca signifi cance below -3 for tracks of different types as explained in the text.

3.3 Selection of tagging tracks

In order to optimize the choice of tagging tracks, all tracks were divided into cathegories according to their fit χ^2/NDF, number of CFT hits, and number of SMT hits. For all these cathegories, the fake rate was estimated as the fraction of tracks with signed dca significance less than -3. The results of this estimation obtained on the EMqcd data sample are shown in Fig. 2. The *x* axis of the plot represents the track type index calculated as $i_{tr} = 30i_{\chi^2} + 6i_{CFT} + i_{SMT}$, according to Table 1. The above fraction was then required to not exceed 0.08. This left tracks with the following characteristics:

- for tracks with $\chi^2/NDF < 3$: all tracks with at least 2 SMT hits;
- for tracks with $3 < \chi^2/NDF < 9$: tracks with at least 4 SMT hits and more than 12 CFT hits, or tracks with at least 5 SMT hits and either no CFT hits or more than 10 CFT hits;
- tracks with $\chi^2/NDF > 9$ are not allowed.

3.4 Optimization of the track p_T cut and dca signifi cance scale factor

It was found that the variation of the track p_T cut provides the biggest improvement in efficiency for the same increase in the mistag rate. Therefore, this parameter was chosen as the one for the varying the working

point. The four working points for which the results will be discussed are $p_T > 0.5$ GeV, $p_T > 1$ GeV, $p_T > 1.5$ GeV, and $p_T > 2$ GeV.

The dca significance scale factor *a* was chosen to be 1.2 which roughly reflects the average value of track dca pulls. An attempt was made to select different values of *a* for different track cathegories, as this provides more uniform cuts on the track dca significance. However, no significant improvement has been observed.

4 Taggable efficiency

4.1 Taggability on data

The standard b-identification group jet taggability definition is as follows.

- Good calorimeter jet should be matched to a track jet ($\Delta R < 0.5$);
- Track jet consists of at least two tracks; distance between two tracks $\Delta R < 0.5$;
- Tracks used to form the track-jet should have at least one SMT hit and at least one of the tracks requires to have a $p_T > 1$ GeV.

Taggable efficiency was studied using EMqcd, jettrig and μ-in-jets samples. Taggability as function of the E_T and η of the jet obtained on all three samples is shown at Fig. 3. It is found to be sample dependent, although its behaviour versus E_T and η is similar. The difference in the taggability can be explained by different kinematics of events in the selected samples. The highest taggability is observed for EMqcd sample. It rapidly increases with jet E_T and reaches the plateau of $\sim$ 85% at jet $E_T \sim$70 GeV.

Taking into account the noticable difference in average taggability on different samples as well as some discrepancy between the shapes of the distributions one has to obtain the taggability parameterization using signal samples.

4.2 Taggability on Monte Carlo

Taggability on the Monte Carlo events is shown on Fig. 4, left, for jets of different flavor. The highest taggability is observed for *b*-jets as expected, since the average track multiplicity in heavy flavor jets is larger than in the light ones. Fig. 4, right, demonstrates the ratio of the *b*- to light and *c*- to light jet taggability. The largest difference of about 10% is observed between low-E_T *b* and *l* jets, which corresponds to the case of jets with low track multiplicity.

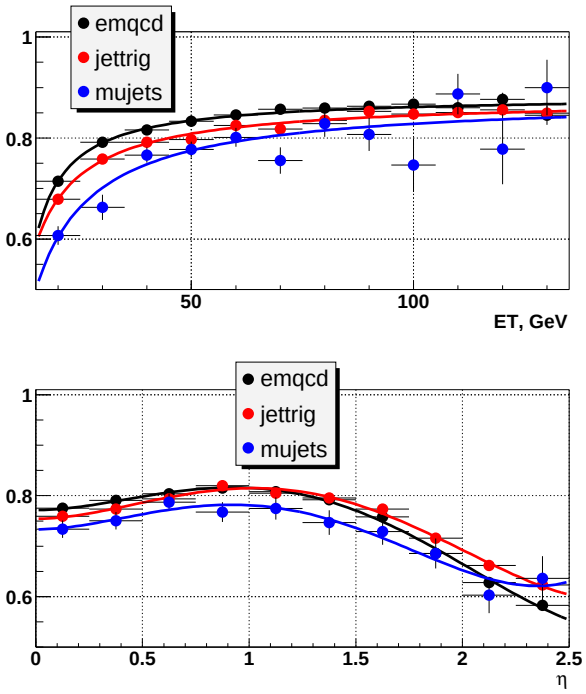

Figure 3: Jet taggability as a function of jet E_T and η for EMqcd, jettrig and mujets samples.

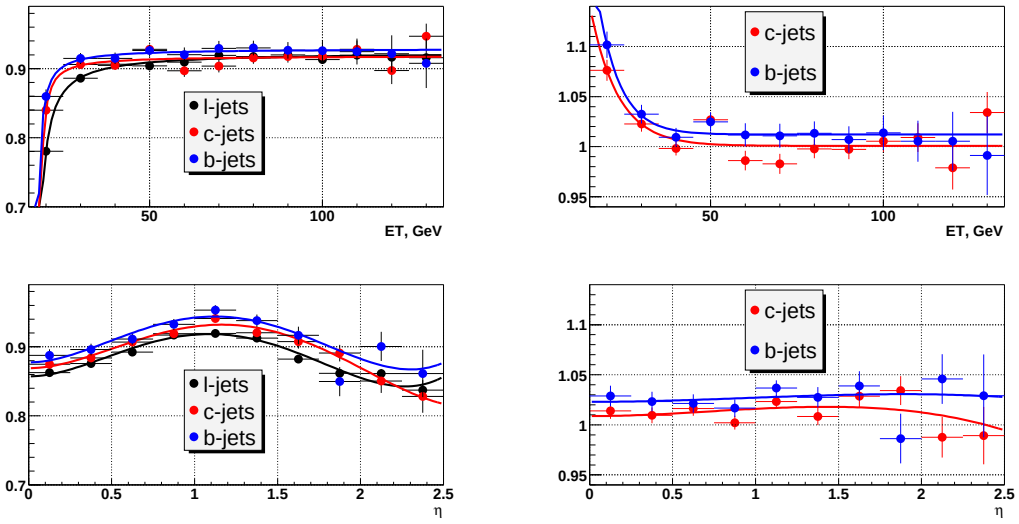

Figure 4: Jet taggability as a function of jet E_T and η for *b*-,*c*- and light jets in Monte Carlo (left) and the ratio of the *b*- to light and *c*- to light jet taggability (right).

5 Tagging efficiency and scale factor

5.1 Methods of *b*-tagging efficiency measurements on data

b-tagging efficiency has been measured by three different methods for four working points:

- Single tags vs no tags (ST vs NT). This method is based on the p_{Trel} fits to the data before tagging (no tags) and to the single tagged muonic jets (ST). The details of the efficiency determination can be found in [3].
- Double tags vs single tags (DT vs ST). This method is variation of the first one. The p_{Trel} fits are performed to the single tagged sample and double tagged sample correspondingly to extract the *b*-tagging efficiency. For more detailed information, see [3].
- System8. This method provides the independent measurement of the *b*-tagging efficiency. The detailed description of this method is done in [4].

A known problem with ST vs NT and DT vs ST methods is the difficulty to fit the p_{Trel} distributions to sum of three templates (*b* for muons from *b*-jets including cascade decays $b \to c \to \mu$, *c* for muons from *c*-jets and *l* for muons from π, K decays in flight) because of the minor differences in *c* and *l* template shapes. We have studied the sensitivity of the p_{Trel} fit based methods to the fixed value of ratio $l/(c+l)$ and to the ratio of efficiencies for *l* jets to *c* jets $\varepsilon_l/\varepsilon_c$. This has been done by looking at the measured *b*-tagging efficiency by two methods, ST vs NT and DT vs ST, as a function of variation of the ratios $l/(c+l)$ and $\varepsilon_l/\varepsilon_c$. Fig. 5 (left) shows the dependence of ST vs NT and DT vs ST methods on the ratio $l/(c+l)$. Both methods agree with each other at $l/(c+l) = 0.7$. In p13, this ratio used to be 0.64. Fig. 5 (right) shows the results of variation of the ratio of efficiencies, $\varepsilon_l/\varepsilon_c$. No sensitivity to this parameter is observed on data.

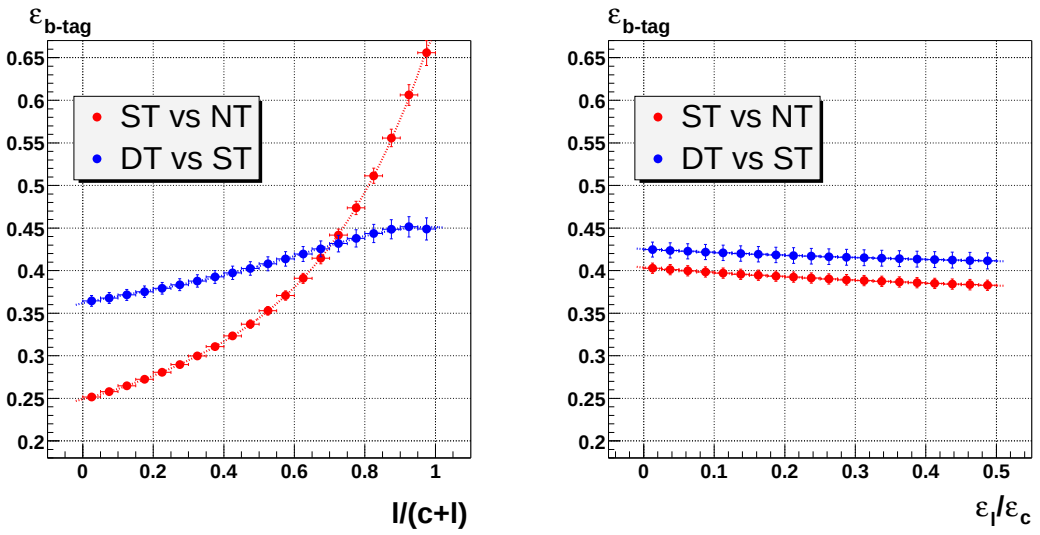

Figure 5: *b*-tagging efficiency as function of $l/(c+l)$ (left) and $\varepsilon_l/\varepsilon_c$ (right) ratios measured on μ+jets data.

System8 has difficulty to provide a measurement of the *b*-tagging efficiency for high E_T jets. By looking at the tagging efficiency of the soft lepton tag for the c/l- and *b*-jets (Fig. 6) reported by System8, it can be seen that the soft lepton tag looses its separation power for jets of E_T above $\sim$70 GeV, and therefore System8 no longer can be resolved.

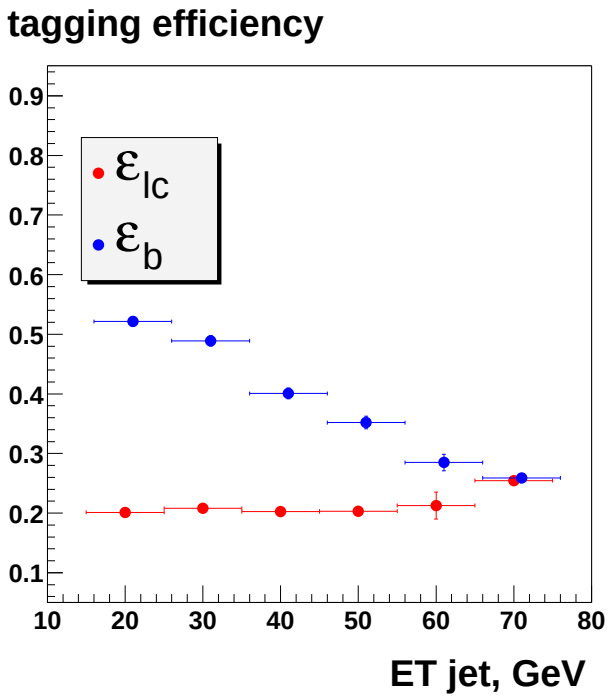

Figure 6: *b*-tagging and c/l tagging effi ciency for the soft lepton tag as a function of the jet $\not{E}_t$.

5.2 *b*-tagging effi ciency measured on data

Average *b*-tagging efficiency for muonic jet measured by these three methods is presented in the Table 2. Good agreement in *b*-tagging efficiency is observed between all three methods, although ST vs NT always gives lower values.

	$p_T > 0.5$	$p_T > 1$	$p_T > 1.5$	$p_T > 2$
System 8	0.443±0.005	0.395±0.005	0.352±0.005	0.315±0.005
ST vs NT	0.407±0.006	0.361±0.006	0.312±0.005	0.268±0.004
DT vs ST	0.440±0.007	0.395±0.007	0.347±0.006	0.308±0.006

Table 2: Average semileptonic *b*-tagging effi ciency measured on data.

b-tagging efficiency as function of $E_{T,jet}$ and η_{jet} measured by three different methods is shown in Fig. 7 for one working point (p_T^{track} >1 GeV).

We choose System8 to compute the final *b*-tagging efficiency numbers and to study the E_T and η dependencies of the *b*-tagging efficiency for the different working points. The other methods are used for the cross-check. The *b*-tagging efficiency measured on the muon-in-jet data as a function of jet E_T and η for all working points is shown on Fig. 8.

The systematic error on the measured b-tagging efficiency is estimated by variation of parameters of System8. The major contribution comes from variation of two parameters, κ_b (decorrelation factor between soft muon tagger and CSIP tagger) and β (increase in b-tagging efficieny if an opposite jet is tagged), while

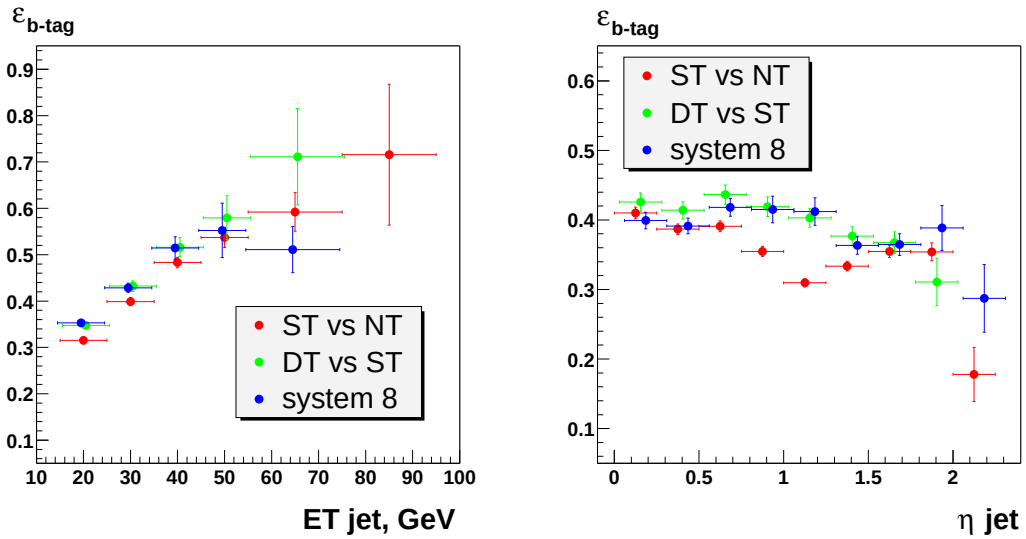

Figure 7: Semileptonic b-tagging effi ciency as function of $E_{T,jet}$ and η_{jet} measured on μ+jets data.

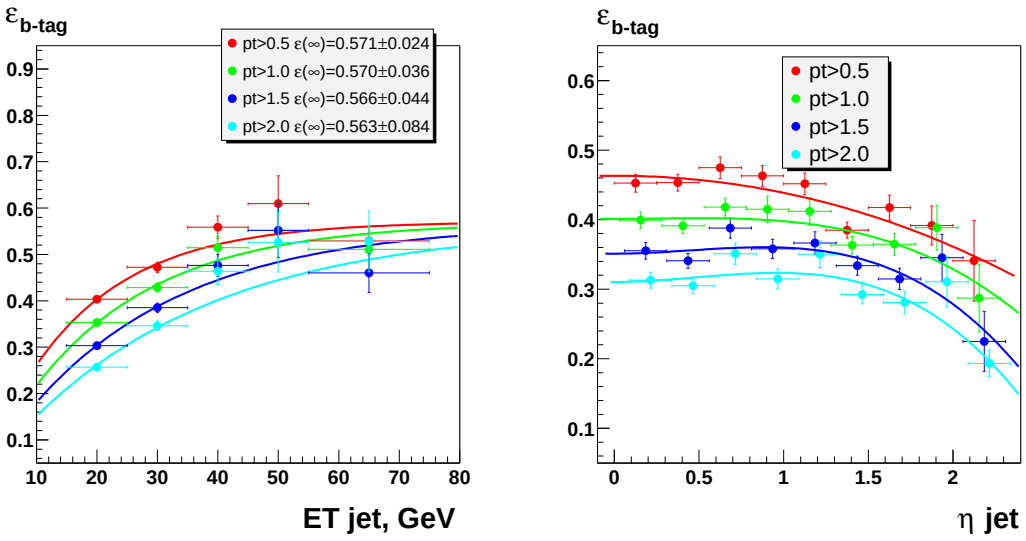

Figure 8: Semileptonic b-tagging effi ciency as a function of jet E_T and η measured on μ-in-jets data for the four working points.

contribution from other sources is negligible. Both κ_b and β were measured on MC $Z \to b\bar{b}$ and found to be $\kappa_b = 0.98 \pm 0.02$ and $\beta = 1.02 \pm 0.02$ in the whole jet p_T region (see Fig. 9). Systematic errors due to each parameter are obtained by $\pm 1\sigma$ variation of the parameter, and are shown on Fig. 10 as functions of jet E_T.

The total systematic error (absolute) on the b-tagging efficiency is shown in Fig. 11. No η_{jet} dependence was found for the total systematic error. Average value of the systematic error on the b-tagging efficiency is

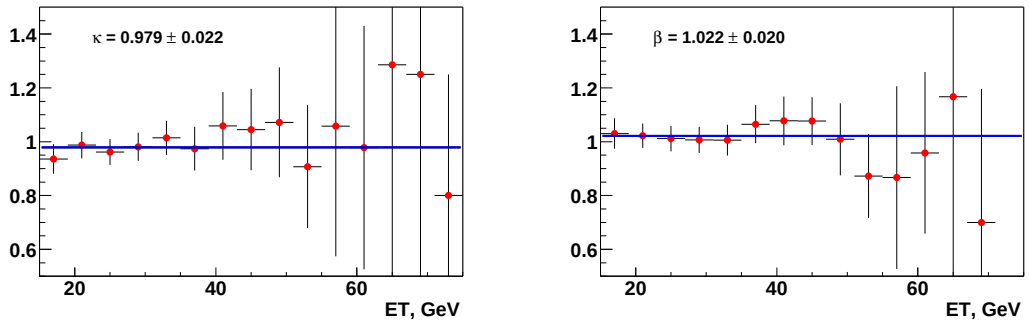

Figure 9: k_b (left) and β (right) as functions of jet p_T.

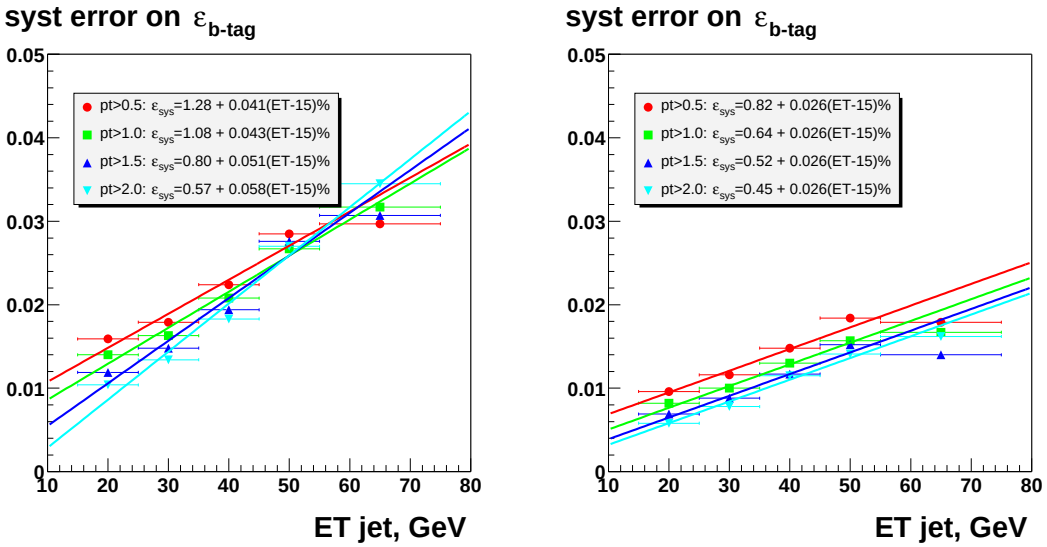

Figure 10: Systematic errors on b-tagging effi ciency due to k_b (left) and β (right).

1.6 % over whole η_{jet} region for the working point $p_T > 1.5$ GeV.

One of the questions about the reliabilty of *b*-tagging measurements by any particular method on data at high E_T is if this method is stable on low statistics. We tried to estimate the stability of System8 and ST vs NT methods with respect to statistics by removing every second event in the available sample or quarter of the statistics and looking at the *b*-tagging efficiency behaviout vs jet E_T. The results are shown in Fig. 12 (left plot for System8 and right for ST vs NT). Although the fluctuations of *b*-tagging efficiency obtained on low statistics are pretty big, there is tendency to increase *b*-tagging efficiency measured with System8 on low statistics samples. ST vs NT method does not demonstrate noticeable dependence on the statistics.

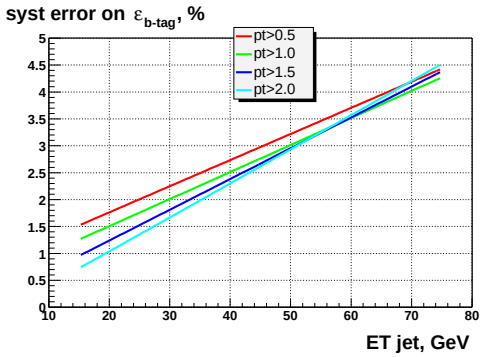

Figure 11: Total systematic error on the b-tagging effi ciency

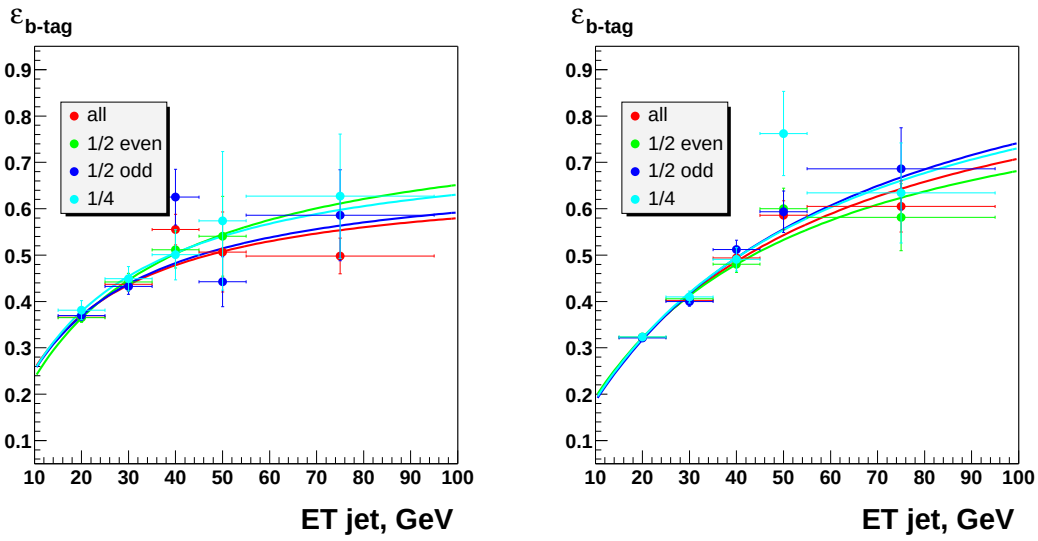

Figure 12: Semileptonic b-tagging effi ciency as function of jet E_T measured by System8 (left) and ST vs NT (right) on the different portions of the μ-in-jet sample. Effi ciency was measured with $p_T^{track} > 1$ GeV.

5.3 The b- and c-tagging effi ciency in Monte Carlo

The jet flavor was determined by matching the direction of the reconstructed calorimeter jet to the quark direction within the cone $\Delta R = \sqrt{\Delta\eta^2_{jet,q} + \Delta\phi^2_{jet,q}} < 0.3$. If there is more than one quark found within the cone, the jet is considered to be a b-jet if the cone contains at least one b-quark; otherwise, it is called a c-jet if there is at least one c-quark in the cone. Light jets are required to have no b- or c- quarks within the jet cone of $\Delta R < 0.3$.

Tagging efficiency in the Monte Carlo is defined as a ratio of the number of tagged jets to the total number of taggable jets of particular flavor within the acceptance. The b- and c-tagging efficiencies are obtained as a function of jet E_T and η. A combined 2D parameterization was derived assuming that E_T and η dependencies are not correlated and can be factorized.

The b-tagging efficiency obtained on $Wb\bar{b}$, $Z \to b\bar{b}$ and $t\bar{t}$ samples is shown in Fig. 13 as a function of jet E_T and η. Significant difference in one-dimensional b-tagging efficiencies obtained on different samples is explained by different E_T and η spectra of b-jets as demonstrated in Fig. 13 (left upper plot). b-jets in $t\bar{t}$ events are much more energetic and more central compared to b-jets in $Wb\bar{b}$ events resulting in higher efficiency both versus jet E_T and η.

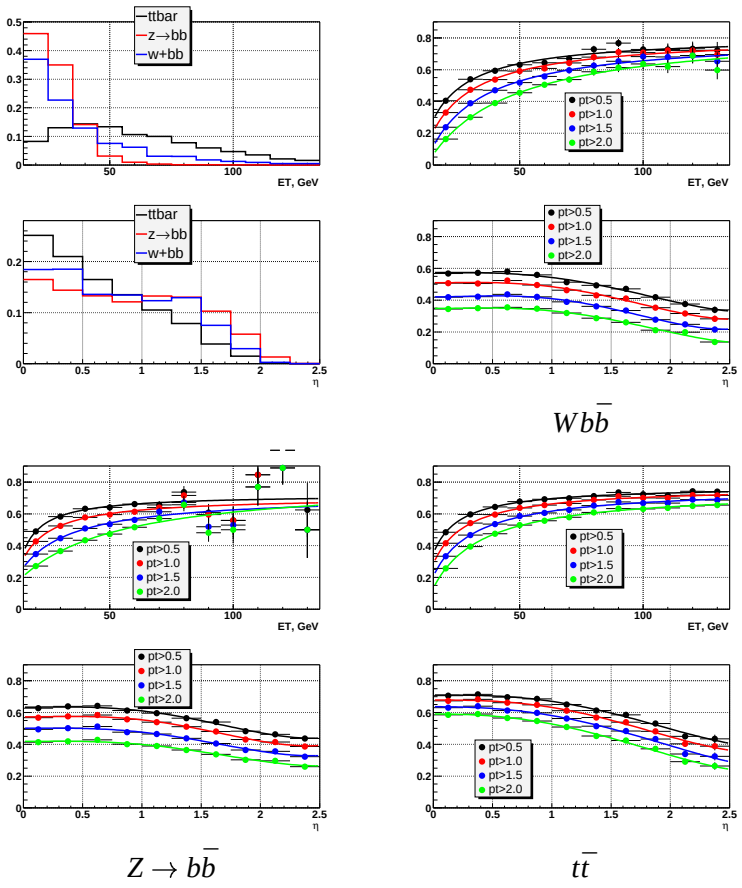

Figure 13: E_T and η distributions of b-jets in $Wb\bar{b}$, $Z \to b\bar{b}$ and $t\bar{t}$ events, and corresponding b-tagging effi cien-cies.

The two-dimensional parameterization of the b-tagging efficiency is done in the following way. b-tagging efficiency is parameterized by two-dimensional function which is the product of two one-dimensional distributions. For two-dimensional histogram of non-tagged jets in the MC sample the number of jets in each (E_T, η) bin is multiplied by the average tagging efficiency in this bin (known from the 2-d parameterization). The total number of predicted tagged events summed over all (E_T, η) bins is then normalized to the actula number of tagged jets in the sample. The resulting 2-d parameterization of b-tagging efficiency obtained on $Wb\bar{b}$ sample is shown in Fig. 14.

We use $t\bar{t}$ sample to cross-check the assumption about independence of b-tagging efficiency versus E_T and η. It was done it by calculating the ratio of the two 2-d parameterizations, one of which was obtained on $Wb\bar{b}$ and another one on $t\bar{t}$ sample. Ratios of the 2-d parameterizations are shown in Fig. 15.

The measured b-tagging efficiency on data is efficiency to tag a *semileptonic* b-jet. To obtain the

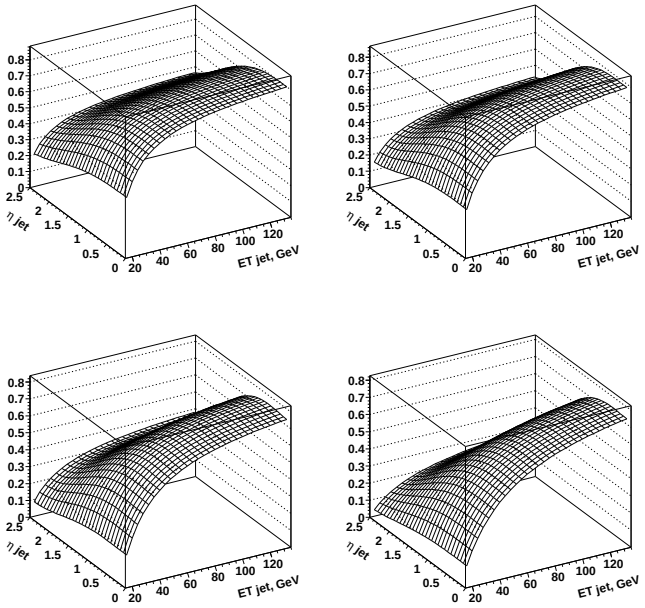

Figure 14: The 2-d parameterization of b-tagging effi ciency in $W\bar{b}$ events for four working points.

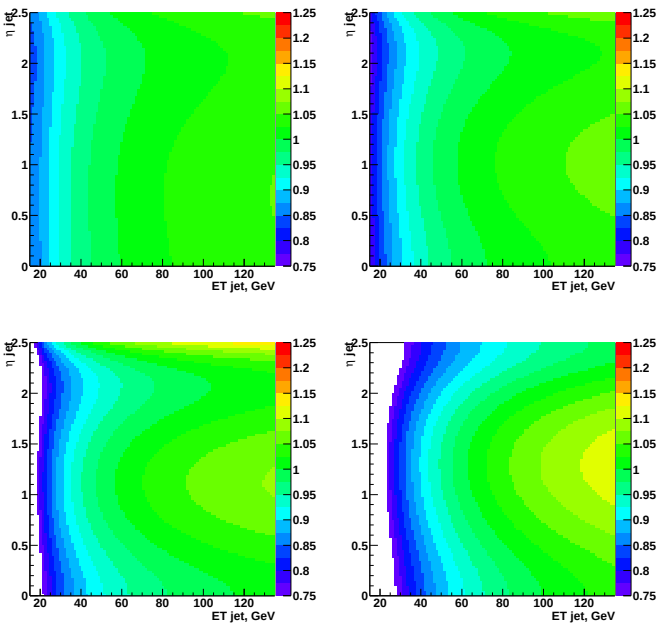

Figure 15: The ratio of b-tagging effi ciency in MC ($W\bar{b}$) to $t\bar{t}$ for four working points.

hadronic efficiency one needs to know the ratio between semileptonic and hadronic b-tagging efficiencies. We calculated the ratio of 2-dimensional parameterizations for semileptonic to hadronic efficiencies using $Z \to b\bar{b}$ MC sample. It is shown in Fig. 16 for four working points. There is practically no difference between semileptonic and hadronic b-tagging efficiencies for working points with low $p_T^{track} \leq 1$ GeV. The semileptonic b-tagging efficiency is systematically higher compared to hadronic b-tagging efficiency at low jet E_T and high jet η for working points $p_T^{track} \geq 1$ GeV. This should be taken into account by introducing the correction factors from Monte Carlo studies.

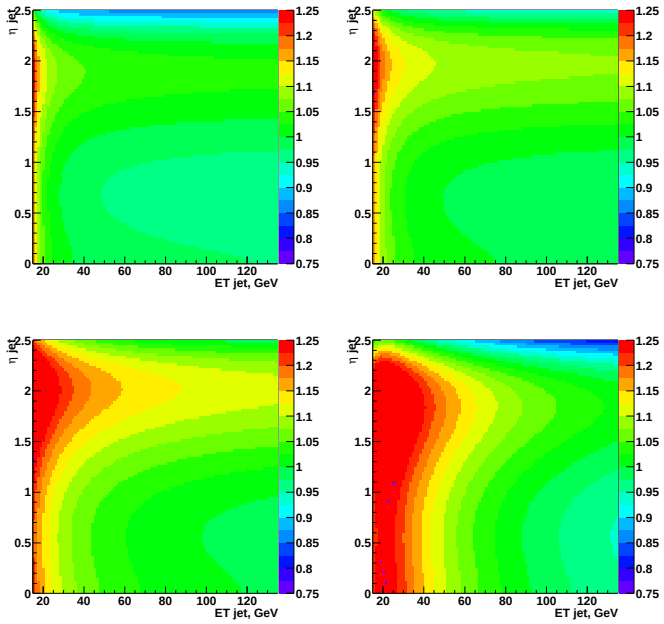

Figure 16: The ratio of semileptonic to hadronic b-tagging effi ciencies in MC ($Z \to \boldsymbol{b\bar{b}}$ for four working points.

c-tagging efficiency was measured on $Wc\bar{c}$ and $Z \to c\bar{c}$ MC samples. Results are shown in Fig. 17. Two-dimensional parameterization obtained on $Wc\bar{c}$ sample is shown in Fig. 18.

5.4 Scale factor

Differences in the track finding efficiency in data and MC as well as in the impact parameter resolutions lead to the differences in the b-tagging efficiency in data and Monte Carlo. Correction factor SF is necessary to relate the b-tagging efficiency in data and Monte Carlo:

$$\varepsilon_{data} = \varepsilon_{MC} \times SF \tag{1}$$

Here both ε_{data} and ε_{MC} were measured on semileptoni b-jets. The MC semileptonic b-tagging efficiency was measured on $Z \to b\bar{b}$ and $t\bar{t}$ samples, where we required b-jets to have a muon inside ($\Delta R(b,\mu) < 0.5$). For the data we have used the b-tagging efficiency obtained with System8. Fig 19 shows the ratios of semileptonic b-tagging efficiency on data and $t\bar{t}$ (left) and $Z \to b\bar{b}$ (right) as a functions of jet E_T and η. Scale factors are fairly flat in jet E_T, η. Constant straight line fits versus jet E_T and η look consistent for all working points and for both MC samples. The values of the scale factor obtained from jet E_T and η fits are brought together in Table 3.

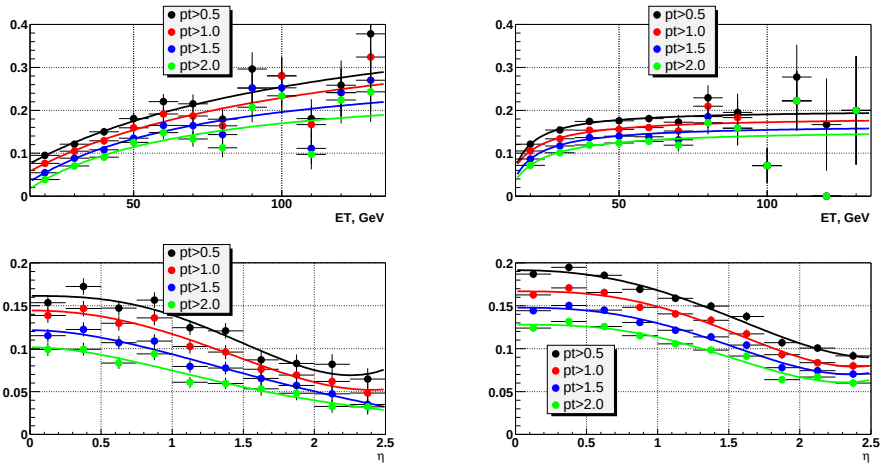

Figure 17: The c-tagging efficiency in MC $Wc\bar{c}$ (left) and $Z \to c\bar{c}$ (right) as a function of jet E_T and η.

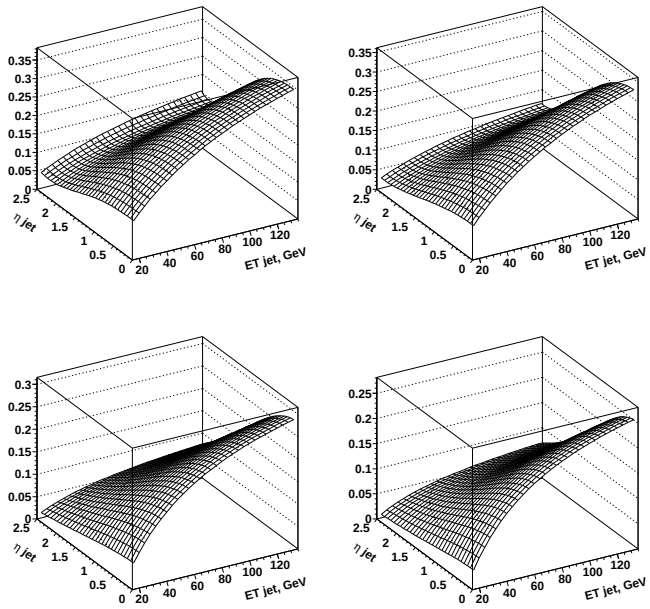

Figure 18: The two-dimensional parameterization of c-tagging efficiency obtained on $Wc\bar{c}$ sample for four working points.

	$p_T > 0.5$	$p_T > 1$	$p_T > 1.5$	$p_T > 2$
fit E_T	0.761±0.009	0.740±0.009	0.731±0.010	0.735±0.012
fit η	0.760±0.009	0.736±0.010	0.728±0.010	0.734±0.012

Table 3: The scale factor values obtained from a constant straight line fits.

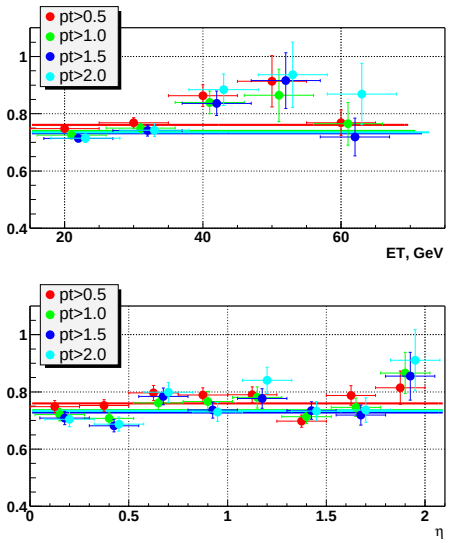

Figure 19: The scale factor as a function of jet E_T and η calculated using semileptonic b-tagging efficiency obtained on $Z \to b\bar{b}$ Monte Carlo sample.

6 Tag rate functions determination

Light jet tagging efficiency can be determined in two ways:

1. We measure negative TRF on data, and apply correction factors for heavy flavor jets and difference in positive and negative tagging rates for light jets;
2. We measure positive TRF on data and subtract fractions of *b* and *c* jets known from QCD Monte Carlo with corresponding efficiencies measured on data (for *b*-jets; as for *c*-jets, we scale it using MC information about the *b*-to-*c* tagging efficiencies).

We used the first method to measure the light jet tagging efficiency and study its dependencies on jet E_T and η. Second method was used for the estimation of the systematic error.

6.1 Method 1

Light jet tagging rate can be obtained using negative tagging rate measured on data: $\varepsilon^+_{light} = \varepsilon^- \times SF_{ll} \times SF_{HF}$. Correction factors SF_{ll} and SF_{HF} are measured on QCD Monte Carlo. SF_{ll} is determined as ratio of positively tagged light jets to the negatively tagged light jets. SF_{HF} is caclulated as ratio of negatively tagged light jets to the negatively tagged inlclusive jets (including heavy flavor). Finally, total correction factor is $SF_l = SF_{ll} \times SF_{HF}$. The average numbers of scale factors SF_{ll}, SF_{HF} and SF_l are listed in the table 6.1 for four working points.

	$p_T > 0.5$	$p_T > 1$	$p_T > 1.5$	$p_T > 2$
SF_{ll}	1.550±0.022	1.537±0.027	1.575±0.034	1.622±0.043
SF_{HF}	0.713±0.010	0.661±0.012	0.612±0.013	0.568±0.015
SF_l	1.104±0.014	1.016±0.016	0.964±0.018	0.920±0.021

Table 4: Correction factors for difference in positive and negative tagging of light jets SF_{ll}, for the presence of heavy flavor in negative tagging rate SF_{HF} and the total correction factor SF_l for light jets measured for four working points.

The total corrections between negative and light tagging rates are about 10% for all working points.

The E_T and η dependence of ε^+_{light} is shown for the four different working points in Fig. 20 obtained on jettrig and EMqcd data.

Comparison of one-dimensional distributions shows a big difference in the light tagging rate for the different samples. This discrepancy originates from the different E_T and η spectra of chosen data, as shown in Fig. 21. The two-dimensional parameterization of the two-dimensional distribution has been done for both samples and compared to each other. The example of the 2-d parameterization obtained on jettrig data is shown in Fig. 22. The ratio of jettrig to EMqcd is shown in Fig. 23. The differential light tag rate efficiency is systematically higher in EMqcd compared to jettrig. This is related to the higher average number of tracks per jet in EMqcd, as it is shown in Fig. 24 for the central detector region, where the η dependence is small.

We also performed self-consistency check, comparing the negative tag rate prediction from the NTRF to the actual number of tagged events in E_T, η bins. The results of the test are shown in Fig. 25. The discrepancy is believed to be due to the fact that the E_T and η dependencies are in fact not completely factorizable. The comparison of the predicted and observed number of negatively tagged jets in EMqcd and jettrig samples is shown in Figs. 26 and 27.

6.2 Method 2

The light tagging rate can be calculated by using positive tagging rate. Positive tagging rate can be written as $\varepsilon^+_{incl} = \varepsilon^+_{light} \times f_{light} + \varepsilon_c \times f_c + \varepsilon_b \times f_b$. Here

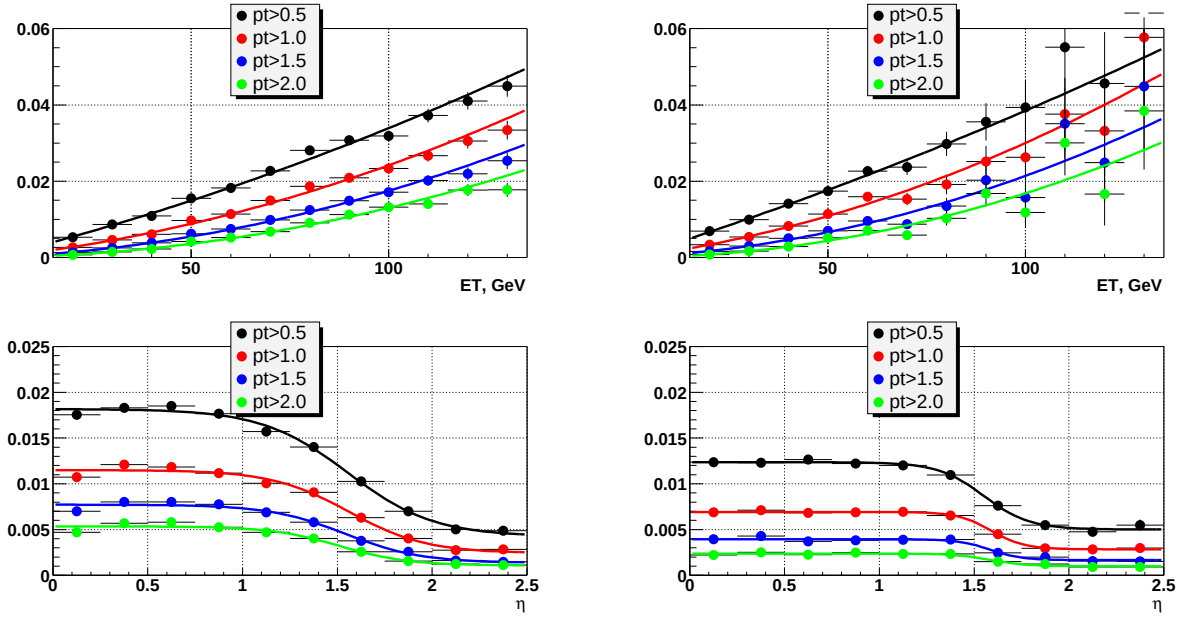

Figure 20: Light tagging rate as a function of jet E_T and η measured on jettrig data (left) and EMqcd data(right).

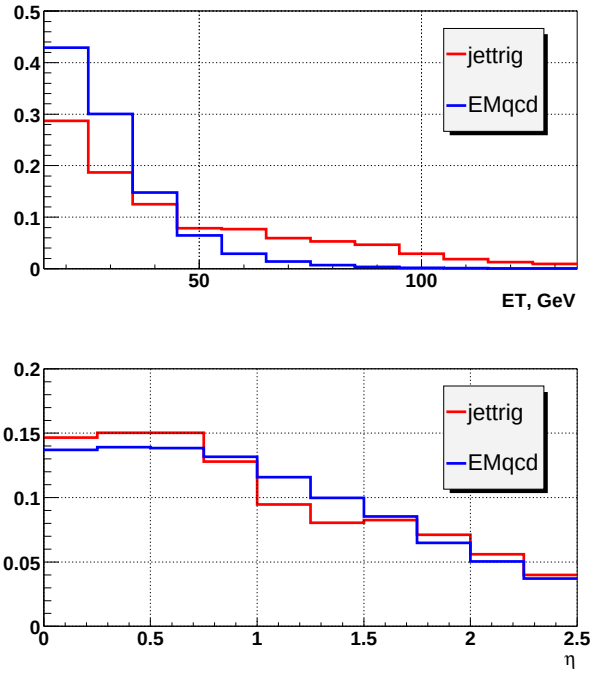

Figure 21: The kinematics of jets in EMqcd and jettrig.

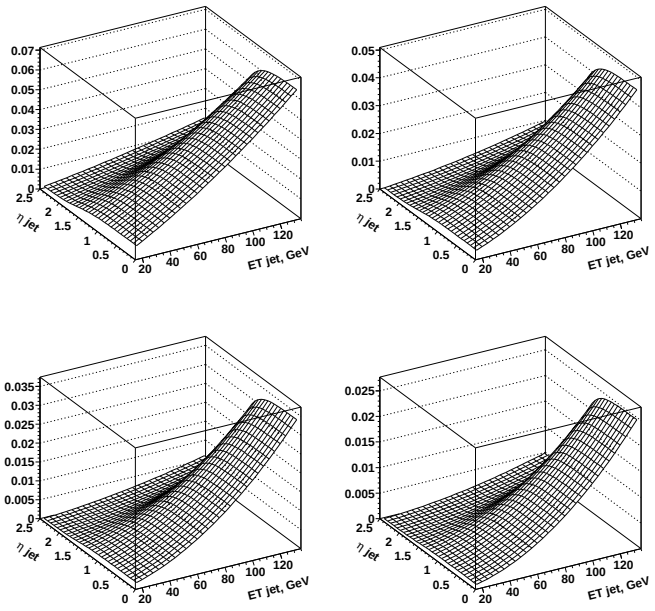

Figure 22: 2-d parameterization of LTRF obtained on jettrig data for four worknig points.

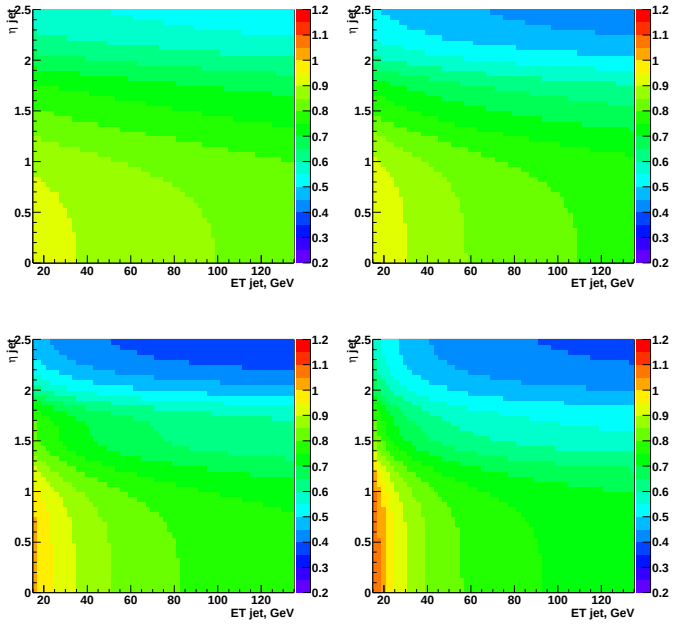

Figure 23: Ratio of the LTRF obtained on jettrig data to the LTRF obtained on EMqcd data.

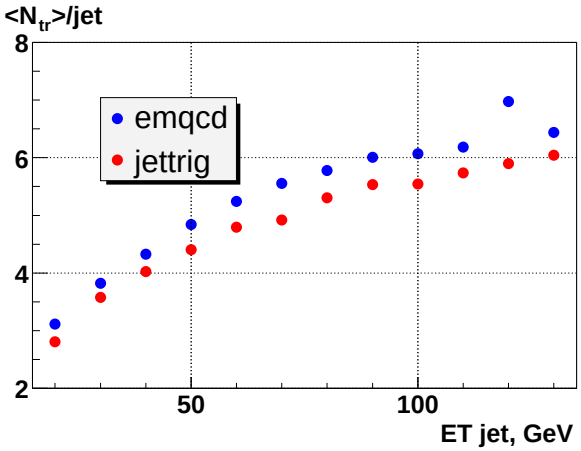

Figure 24: The average number of tracks per jet in EMqcd and jettrig samples in the central region ($|\eta^{det}| < 0.8$).

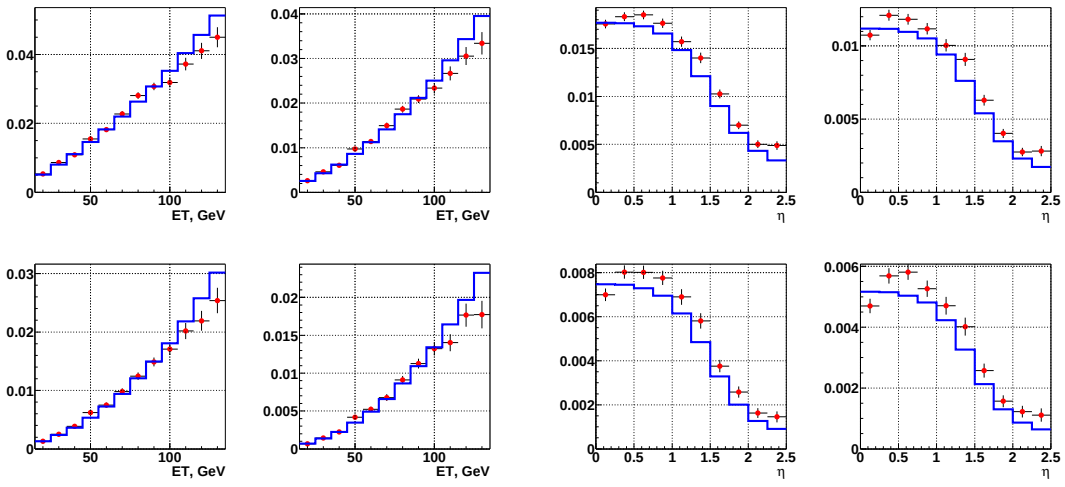

Figure 25: Negative tag rate as predicted by NTRF (blue lines) and actual negative tags (red points) vs jet E_T (left) and η (right).

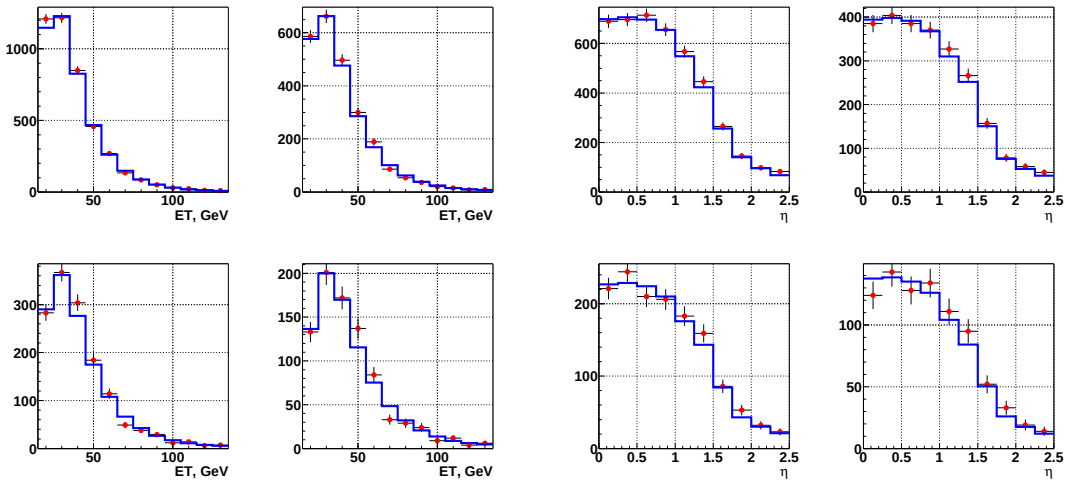

Figure 26: The number of negatively tagged events in EMqcd predicted by NTRF (blue lines) and actual negative tags (red points) vs jet E_T (left) and η (right).

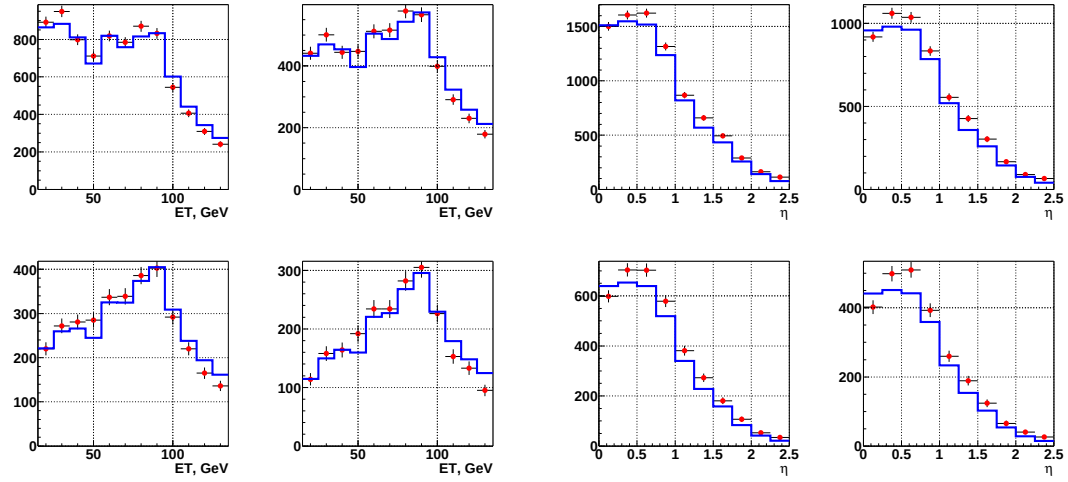

Figure 27: The number of negatively tagged events in jettrig predicted by NTRF (blue lines) and actual negative tags (red points) vs jet E_T (left) and η (right).

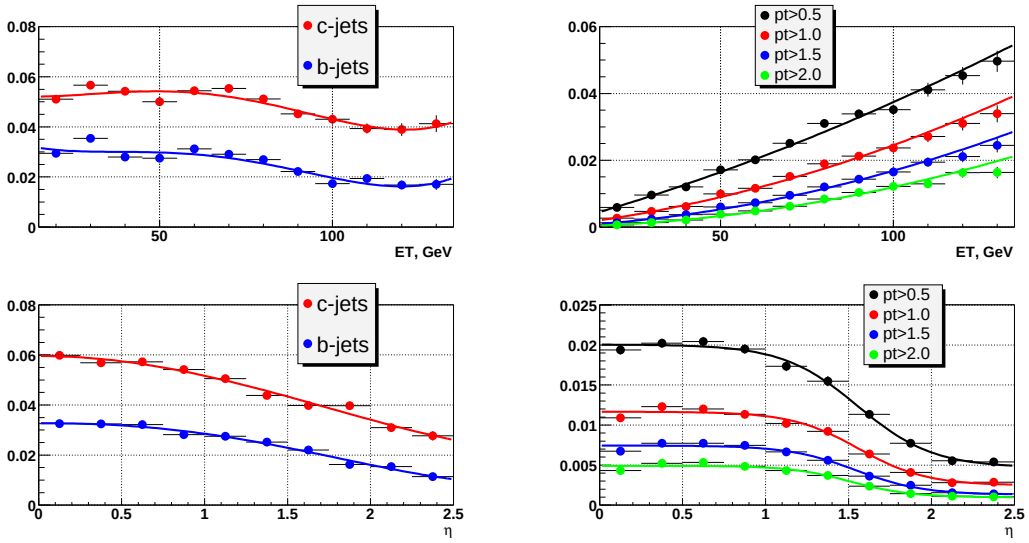

Figure 28: Heavy flavor fractions in QCD MC (left) and ε^+ vs jet ET determined from positive (right) tag rate.

- f_{light}, f_c, f_b - fractions of light, b and c-jets in the data sample. Since we can not measure these fractions in data we assume that they are the same as in QCD MC sample.
- ε_b is b-tagging efficiency measured on data.
- ε_c is c-tagging efficiency measured on Monte Carlo corrected to the data by multiplying it by $SF_b = \varepsilon_b^{data}/\varepsilon_b^{MC}$.

One can derive the $\varepsilon^+_{light} = (\varepsilon^+_{incl} - \varepsilon_c \times f_c - \varepsilon_b \times f_b)/(1 - f_c - f_b)$.

The weak point of this method is use of f_{light}, f_c, f_b from Monte Carlo. We know from Run I CDF measurements [5] that b-jet cross-subsection is larger compared to the one predicted by Herwig. One can see from ε^+_{light} definition that increase of heavy flavor fractions will lead to decrease of the light tagging rate determined in that way. Since we study the mistag rate as a functions of jet E_T and η, we assume that spectra f_c, f_b in data are the same as in the Monte Carlo, which is not obvious as well. A priori, fractions f_c, f_b depend on data preselection to a large extent.

Fractions of b and c-jets measured on QCD MC sample as functions of E_T and η of jet are shown in Fig.28 (left). Both fractions decrease with E_T. The calculated light tagging rate is shown in Fig.28 (right).

6.3 Consistency check of LTRF obtained by two methods

The results for LTRF obtained by the two methods on the jettrig sample are shown in Fig. 29. The relative systematic errors determined as half the difference between parameterizations from two methods divided by LTRF from Method 1, are shown in Fig. 30. One can see that LTRF agree within 20%.

6.4 Systematic error on LTRF

We consider the following sources of systematics:

- Difference between negative tagging rates obtained on EMqcd and jettrig data. It is shown in Fig. 31 (left) and is constant over all jet E_T region. We assign it to be 9% (half of the difference between EMqcd and jettrig) for all working points.

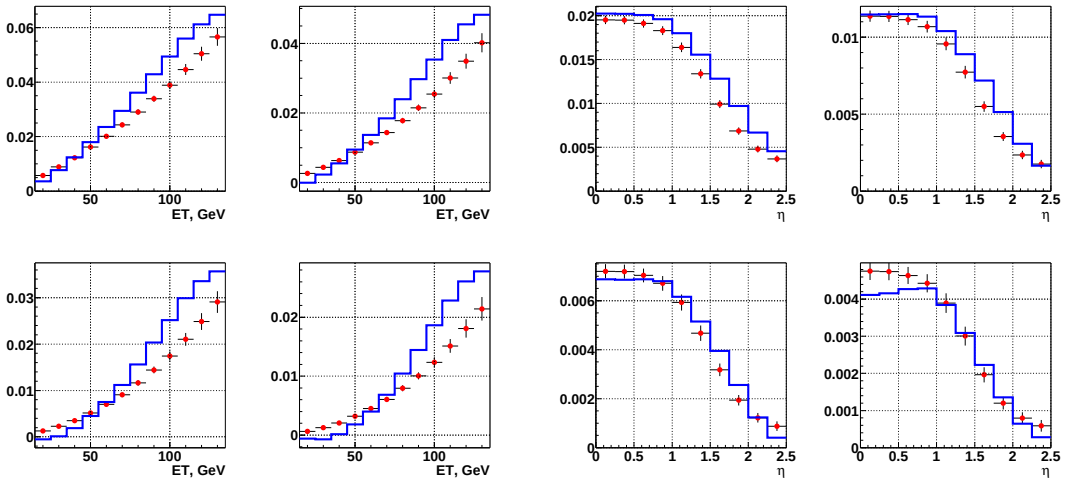

Figure 29: ε^{+}_{light} obtained on the jettrig sample using Method 1 (red points) and Method 2 (blue lines) vs jet E_T (left) and η (right).

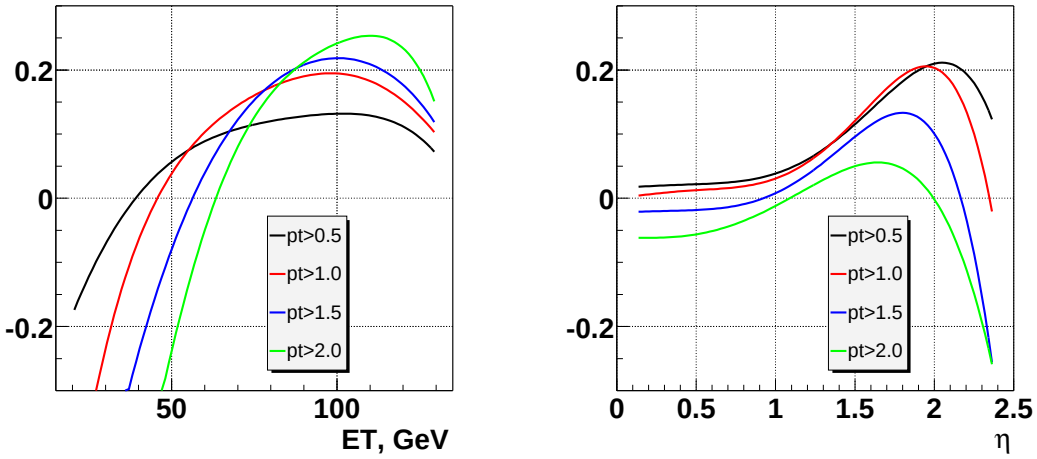

Figure 30: The relative ε^{+}_{light} error as a function of jet ET (left) and η (right).

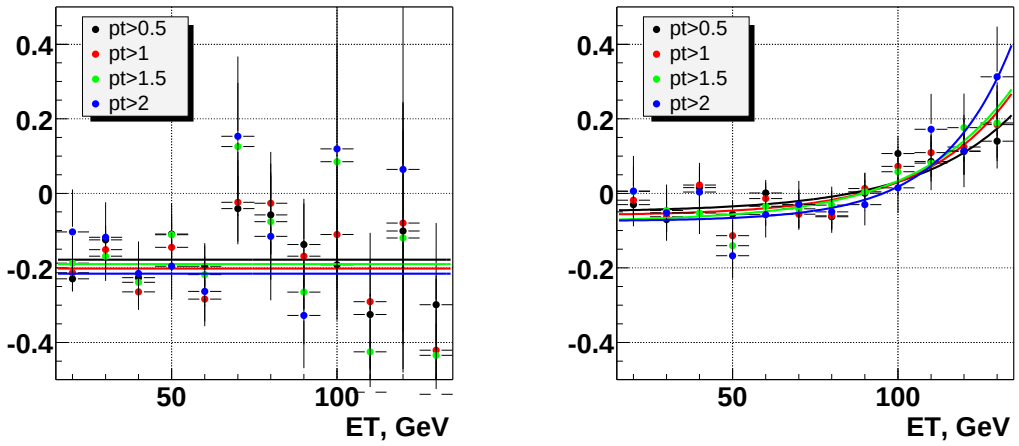

Figure 31: The relative ε^{+}_{light} error as a function of jet ET (left) and η (right).

- Difference between number of tagged jets and number of predicted tagged jets using 2-dimensional parameterization. We calculated it using jettrig data closure test. Results are shown in Fig. 31 (right). We observe inconsistency up to 20% at high jet E_T in jettrig data. EMqcd sample has small statistics in this region and therefore can not be used for evaluation of this kind of systematic error.
- Uncertainty in SF_{HF} due to uncertainty of the fraction of heavy flavor in QCD MC: $\Delta SF_{HF}/SF_{HF} = \Delta f_h/f_h \times (1 - SF_{HF})$ (see Appendix B). We assume that error due to this factor is relatively small compared to errors from other sources.

We add first two errors in quadrature to obtain the total relative systematic error on LTRF, which is of the order of 12% at low jet E_T and rises up to 17% at $E_T \sim$120 GeV. The total systematic error at high E_T is in agreement with the error obtained from comparison of LTRFs derived by methods 1 and 2 (see Fig. 30).

7 Summary

Studies of taggability on different samples show strong dependence of this characteristic on selection criteria and jet parameters.

Light tagging rate function (mis-tagging rate) is estimated using two different methods. Two-dimensional parameterization of the LTRF is obtained using method based on the negative tagging rate correction. We found that *b*-tagging efficiency measurements made by three different methods are consistent with each other within statistical errors. The average *b*-tagging efficiency obtained on μ-in-jets data varies from (44.0 ± 0.7)% to $(31.1 \pm 0.9$% at the mis-tagging rate in the range from 1.2 % to 0.2%.

The systematic error on the *b*-tagging efficiency measured by System8 method is found to be independent on jet η, but E_T dependent. The average absolute systematic error is 1.6 % over whole η_{jet} region for the working point $p_T > 1.5$ GeV.

The scale factor between *b*-tagging efficiencies measured on data and MC is flat versus E_T and η for all working points and should be used as a number (slightly different for different working point).

Appendix A Functional form of fits

The following fits were used for parameterizations:

- *b*- and *c*-tagging efficiency in MC:
  $f(E_T) = p_0 \frac{E_T - p_1}{E_T + p_2}$
  $f(\eta) = p_0 + p_1\eta^2 + p_2|\eta|^3 + p_3\eta^4$
- light tagging efficiency in MC and LTRF in data:
  $f(E_T) = p_0 E_T + p_1 E_T^2$
  $f(\eta) = p_0 + p_1 \tanh(\frac{p_2 - |\eta|}{p_3})$
- *b*-tagging efficiency in data (muon-in-jets):
  $f(E_T) = p_0(1 - \exp(-E_T/p_1))$
  $f(\eta) = p_0 + p_1\eta^2 + p_2|\eta|^3$

All 2d parameterizations are assumed to factorize in (E_T,η), that is, they can be represented as $f(E_T,\eta) = Cf(E_T)f(\eta)$.

Appendix B Uncertainty in SF_{HF} due to uncertainty in HF fraction

SF_{HF} is defined as the ratio of negatively tagged light jets to the negatively tagged inlclusive jets:

$$SF_{HF} = \frac{\varepsilon_l^-}{\varepsilon^-}$$

with jet tagging probabilities defined as

$$\varepsilon_l^- = \frac{N_l^{tag}}{N_l} \qquad \varepsilon_h^- = \frac{N_h^{tag}}{N_h} \qquad \varepsilon^- = \frac{N_l^{tag} + N_h^{tag}}{N_l + N_h}$$

here N_l, N_h and N_l^{tag}, N_h^{tag} are numbers of light and heavy flavor jets before and after tagging, respectively. The fraction of heavy flavor jets

$$f_h = \frac{N_h}{N_l + N_h}$$

Therefore we derive

$$\frac{1}{SF_{HF}} = 1 + f_h\left(\frac{\varepsilon_h^-}{\varepsilon_l^-} - 1\right) \qquad \text{and} \qquad \frac{\Delta SF_{HF}}{SF_{HF}} = \frac{\Delta f_h}{f_h}(1 - SF_{HF})$$

References

[1] R.Demina, A.Khanov, F.Rizatdinova. DØ Note 4049;
R.Demina, A.Khanov, F.Rizatdinova. DØ Note 4060.

[2] J.-R. Vlimant et al, DØ Note 4146.

[3] R.Demina et al, DØ Note 4133.

[4] B.Clement, D.Bloch, D.Gele, S.Greder, I.Ripp-Baudot, D0 Note 4159.

[5] T. Affolder et al., The CDF Collaboration, Phys. Rev. D64, 032002 (2001)