Optimization of track-to-jet association for b-tagging algorithms

JUAN RAMÓN CASTIÑEIRAS DE SAA

Introduction

In this report I will briefly describe my project, which is based on the study of the performance of b-tagging algorithms. The specific aim has been the modification of the track-to-jet association criteria, in order to achieve an improvement under challenging conditions, such as high p_T of the jets or high pile-up, i.e. the number of primary interactions in the event. For that purpose, simulated samples have been used.

Jets that arise from bottom-quark hadronization (b-jets) are present in many relevant physical processes. The ability to accurately identify b jets is crucial in reducing the otherwise overwhelming background from processes involving jets from gluons and light-flavour quarks (u, d, s), as well as c-quark fragmentation. The properties of the bottom hadrons can be used to identify the hadronic jets into which b quarks fragment. These hadrons have relatively large masses, long lifetimes and daughter particles with hard momentum spectra. Several algorithms have been developed and used in the CMS experiment, and exploit these properties to achieve the identification of b quarks [1]. The description of the CMS detector can be found elsewhere [2].

Description of the algorithms for b-jet identification

Several reconstructed objects, such as tracks, vertices or identified leptons, can be used to build observables that discriminate between b and light-parton jets. Simple algorithms use only one of these observables; while more complex ones can use several of them to achieve a higher discriminant power. Each of these algorithms yields a single discriminator value for each jet. In the following paragraphs, a brief explanation of the selection of the objects used in b-jet identification will be given, and then the track counting algorithms will be described as they are the ones which will be studied in this project.

Track selection

Jets are clustered from objects reconstructed by the particle-flow (PF) algorithm [3], which combine the information from all subdetectors to reconstruct particles. Then the particles are clustered into jets using the anti-k_T clustering algorithm [4]. Primary vertex candidates are selected by clustering reconstructed tracks based on the z coordinate of the closest approach to the beam line. Among the primary vertices found this way, the one with the highest $\sum\left(p_T^{track}\right)^2$ is selected as candidate for the origin of the hard interaction.

The b-jet tagging algorithms require a sample of well-reconstructed tracks of high purity. Tracks are associated to the jet if they lie in a cone of $\Delta R < 0.5$ around the jet axis, where ΔR is defined as $\sqrt{(\Delta\phi)^2 + (\Delta\eta)^2}$. Besides, extra requirements are imposed in addition to the selection applied in the tracking step. Those criteria are the following:

- $p_T^{track} > 1\ GeV$.
- $\chi^2/n.d.o.f. < 5$.
- At least 8 hits in the total detector associated to the track.
- At least 2 hits in the pixel system.

- Loose selection on the impact parameters d_{xy} and d_z (defined as the transverse and longitudinal distance to the primary vertex at the point of closes approach in the transverse plane)
    - $d_{xy} < 0.2\ cm.$
    - $d_z < 17\ cm.$
- Distance of closest approach of the track to the axis less than $700\ \mu m$.
- Point of closest approach within 5 cm of the primary vertex.

Identification using track impact parameters

The IP of a track with respect to the primary vertex, which is calculated in three dimensions, can be used to distinguish the decay products of a b hadron from prompt tracks. This quantity has the same sign as the scalar product of the vector pointing form the primary vertex to the point of closest approach with the jet direction. Track originating from the decay of particles travelling along the jet axis will tend to have positive IP values; on the other hand, the IP of prompt tracks can be positive or negative. The impact parameter significance S_{IP}, defined as the ratio of the IP to its estimated uncertainty, is used as an observable. The **Track Counting** (TC) algorithm sorts tracks in a jet by decreasing values of the IP significance. 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. Hence, there are two versions of the algorithm that use the S_{IP} of the second and third ranked track as the discriminator value, and are called *Track Counting High Efficiency* (TCHE) and *Track Counting High Purity* (THCP), respectively.

Track-to-jet association

In order to achieve a better performance of the current b-tagging algorithms for jets of high p_T and events with high pile-up or a large number of jets in the event, one of the features susceptible to be modified is the way the tracks are associated to the jets.

Two Monte Carlo (MC) samples of multijet events generated by Pythia have been used. One of them has been generated by requiring the p_T of the outcoming partons from the primary interaction ($\hat{p}_T$) to range between 80 and 120 GeV. In the other one, the $\hat{p}_T$ ranges between 300 and 470 GeV. At analysis level, we require the p_T of the reconstructed jets to be in the same interval as the $\hat{p}_T$ of the corresponding MC sample.

Explicit association

By the default algorithm, tracks are associated to a certain jet in a cone of $\Delta R < 0.5$ around the jet axis. However, an “explicit association” can be obtained by considering the tracks that have been used in the jet clustering by the PF algorithm.

The performance of the algorithms can be described by plotting the probability of misidentifying jets of other flavours (i.e. the probability to misidentify charm, or light jets composed of u, s, d quarks and gluons as b-jets) as a function of the corresponding efficiency to identify a b-jet, as the cut on the discriminator is varied. The performance respect to light jets misidentification has a special interest due to the dominant presence of this background in many relevant processes. Each algorithm has its own characteristic performance curve; some of them offer a better performance in the region of high efficiency (i.e. TCHE), meanwhile others are better suited for the region of high purity -low misidentification rate-, such as the TCHP algorithm.

The comparison between the performance curves for the default and the explicit association is shown in **Figure 1**. As it can be seen, high p_T performance is worse than the low p_T one. However, there is a gain in the performance below an efficiency of 0.45 aproximately, while for higher efficiencies the performance is worsened. In the high p_T scenario, the deterioration at high efficiency does not have a large impact compared to the upturn at high purity. In view of this result, we can assert that the usage of the explicit association delivers a better performance for large values of p_T of the jets at high purity.

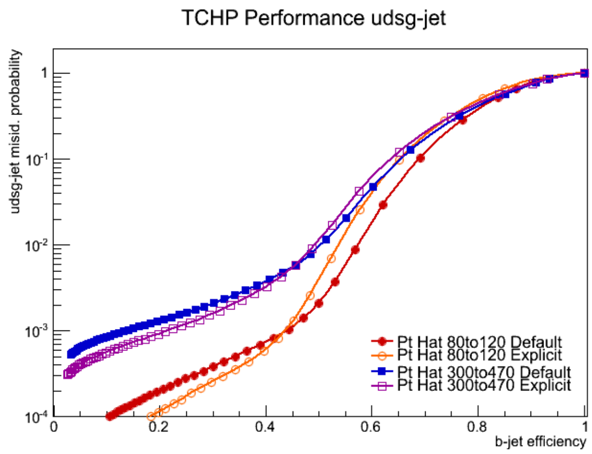

Figure 1. *Performance of the TCHP algorithm with respect to the light jet misidentification probability. The default and the explicit association are compared, also depending of the range of p_T of the jets.*

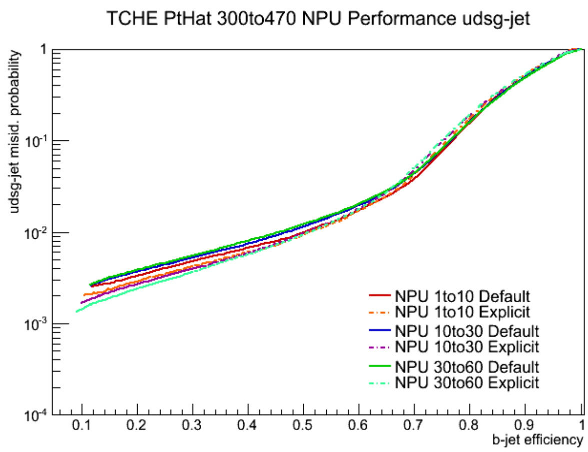

Figure 2. *Performance of the TCHE algorithm with respect to the ligh jet misidentification probability for high p_T jets, regarding several ranges of pile-up (NPU). The default (solid line) and the explicit association (dotted line) are compared.*

It is also possible to study the dependence of the performance with respect to the pile-up. In Figure 2 the performance with respect to light jets misidentification is shown for several ranges of pile-up, for jets of high p_T. As it can be seen, the behaviour of the tagging algorithms worsens with the increase of the pile up, but when using the explicit association an improvement on the performance at high purity is achieved.

p_T^{Rel} *based association*

Another possible modification which we considered in order to improve the performance of the b-tagging algorithms is to optimize track selection requirements for jets with high p_T. As a first step, it is necessary to study the behaviour of different track variables, in order to optimize the selection cuts. Moreover, some of the variables are correlated, so it is interesting to create two-dimensional plots in which one variable is plotted versus the other one. Regarding this, the study of the p_T of the tracks as a function of ΔR respect to the jet axis has a special relevance, because tracks that come from the decay of heavy B-hadrons tend to have greater values of p_T and smaller ΔR. Another interesting variable is the p_T^{Rel}, which is defined as the component of the p_T of the track respect to the jet axis. This variable is widely correlated with the p_T, and it has also been noted to have greater values for b-jets. An example of one of those two-dimensional plots is shown in Figure 3.

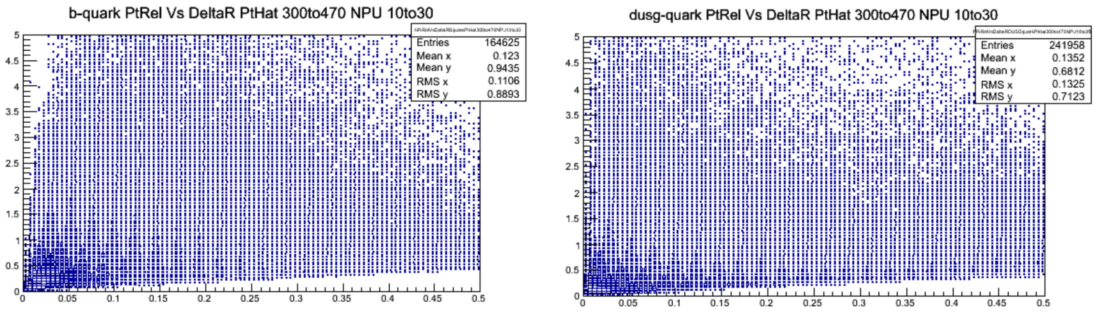

Figure 3. p_T^{Rel} versus ΔR for b jets (left) and light-flavoured jets (right), for high p_T and a medium value of pile-up.

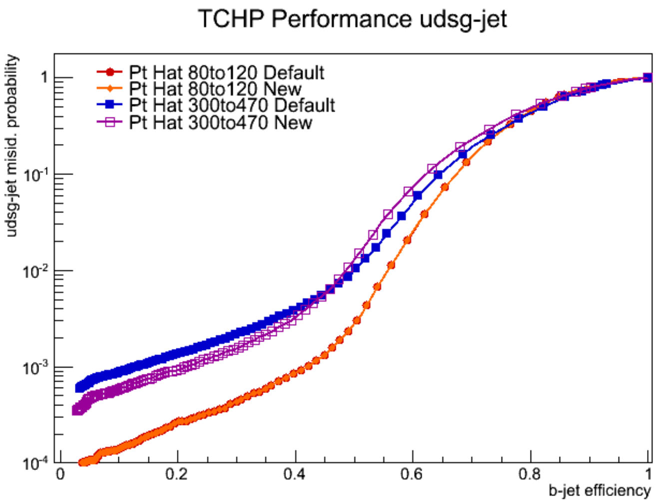

Figure 4. *Performance of the TCHP algorithm with respect to the light jet misidentification probability. The default and the new selection are compared, also depending of the range of* p_T *of the jets (the new selection is only applied to the higher range of* p_T*).*

Selecting several intervals of ΔR and studying the values of the p_T^{Rel} cut that maximize the ratio of the efficiency of b-quarks over the square root of the efficiency of light-quarks, it is possible to find the most appropriate values for the selection cuts of the variables in each range. The modifications are only studied on the high p_T region as it is the one we are interested in. A comparison is done between the performance calculated with these *new* criteria and the *default* ones. The result is shown in Figure 4. It is visible that the improvements achieved are very similar with the ones obtained using the explicit association, so both studies are comparable.

In addition, as it was done before, we can also study the dependence with the pile-up. Figure 5 shows how the performance improves in the high purity region when applying the new selection.

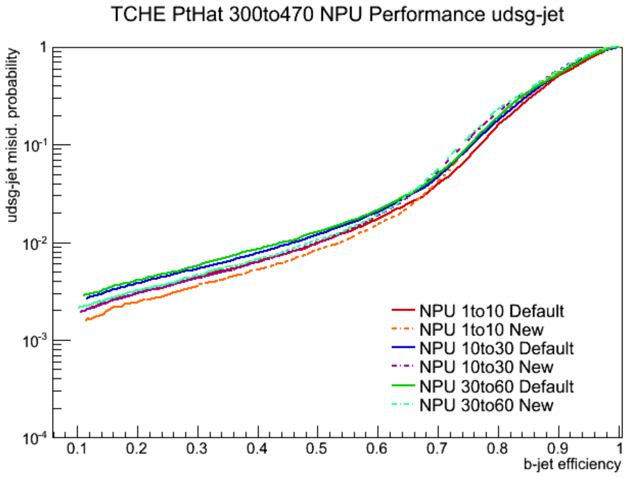

Figure 5. *Performance of the TCHE algorithm with respect to the light jet misidentification probability for high p_T jets, regarding several ranges of pile-up (NPU). The default (solid line) and the new selection (dotted line) are compared (the new selection is only applied to the higher range of p_T).*

Conclusions

Several modifications to the way tracks are associated to the jets have been implemented in order to improve the performance of the b-tagging algorithms for jets of high p_T and events with high pile-up or a large number of jets in the event. In first place, we used the tracks explicitly associated to the jet by the PF algorithm. In second place, we optimized the track selection for high p_T exploiting the higher p_T^{Rel} values of tracks coming from b-hadrons. In both cases, a similar improvement of the performance is achieved for high b-tagging purity, in the conditions we were interested in.

Basic references

[1] CMS Collaboration, “Identification of b-quark jet with the CMS experiment”, JINST 8(2013) P04013

[2] CMS Collaboration, “The CMS experiment at the CERN LHC”, JINST 3(2008) S08004.

[3] CMS Collaboration, “Particle-Flow event reconstruction in CMS”, PAS PFT-10-003.

[4] M. Cacciari, G. P. Salam, and G. Soyez, “The anti-κ_T jet clustering algorithm”, JHEP 04(2008)