PROCEEDINGS OF SCIENCE

b-tagging algorithms and performance in ATLAS

Laurent VACAVANT, on behalf of the ATLAS Collaboration
Centre de Physique des Particules de Marseille
CNRS/IN2P3 et Université de la Méditerranée
163, Avenue de Luminy - Case 902
13288 Marseille Cedex 09, France
E-mail: `vacavant@in2p3.fr`

The ability to identify jets stemming from the fragmentation and hadronization of b quarks is important for the high-p_T physics program of ATLAS: top physics, Higgs boson searches and studies, new phenomena. The algorithms to tag b-jets are described and their anticipated performance discussed. Finally, methods to measure b-tagging performance in the first data and the expected accuracy of those measurements are briefly discussed.

Physics at LHC 2008
29 September - October 4, 2008
Split, Croatia

1. Introduction

The ability to identify jets containing b-hadrons is important for the high-p_T physics program of a general-purpose experiment at the LHC such as ATLAS [1]. This is in particular useful to select very pure top quark samples, to search and/or study Standard Model or supersymmetric (SUSY) Higgs bosons which couple preferably to heavy objects or are produced in association with heavy quarks, to veto the large $t\bar{t}$ background for several physics channels and finally to search for new physics: SUSY, heavy gauge bosons, etc.

The identification of b-jets takes advantage of several of their properties which allow us to distinguish them from jets which contain only lighter quarks. First the fragmentation is hard and the b-hadron retains on average 70% of the original b quark momentum. In addition, the mass of b-hadrons is relatively high (>5 GeV/c^2). Thus, their decay products may have a large transverse momentum with respect to the jet axis and the opening angle of the decay products is large enough to allow separation. The third and most important property is the relatively long lifetime of hadrons containing a b quark, of the order of 1.5 ps ($c\tau \approx 450\mu$m). A b-hadron in a jet with $p_T = 50$ GeV/c will therefore have a significant flight path length $\langle l \rangle = \beta\gamma c\tau$, traveling on average about 3 mm in the transverse plane before decaying. This can be identified either inclusively by measuring the impact parameters of the tracks from the b-hadron decay products or exclusively by reconstructing the displaced vertices. In addition, the semi-leptonic decays of b-hadrons can also be used by tagging the lepton in the jet. This paper discusses the first two approaches. More material can be found in Ref. [2].

2. Impact parameter-based algorithms

The transverse (d_0) and longitudinal (z_0) impact parameters of tracks are computed with respect to the primary vertex and are signed positively if the track crosses the jet axis in front of the primary vertex and negatively otherwise. In ATLAS, a transverse impact parameter resolution of 35 μm is expected for a central track with $p_T = 5$ GeV/c fulfilling the b-tagging quality criteria (mostly requiring a hit on the first pixel layer). To give more weight to well-measured tracks, the impact parameter significance d_0/σ_{d_0} is used for discriminating b- and light jets (cf. Fig. 1).

The simplest tagging algorithm consists in counting tracks with large impact parameter or impact parameter significance. Another algorithm, JetProb, compares for each track its d_0/σ_{d_0} to a resolution function for prompt tracks, measuring the probability that the track originates from the primary vertex. The track probabilities are then combined into a jet probability. The resolution function can be measured in data using the negative side of the signed impact parameter distribution, assuming there is no contribution from heavy-flavor particles.

The most advanced algorithms are using a likelihood ratio approach: the measured value S_i of a discriminating variable (such as d_0/σ_{d_0} for the IP2D algorithm) is compared to pre-defined smoothed and normalized distributions for both the b- and light jet hypotheses, $b(S_i)$ and $u(S_i)$, obtained initially from Monte Carlo. Two-dimensional probability density functions (d_0/σ_{d_0} *vs* z_0/σ_{z_0}) are used for the IP3D algorithm. The ratio of the probabilities $b(S_i)/u(S_i)$ defines the track weight, which can be combined into a jet weight as the sum of the logarithms of the individual track weights.

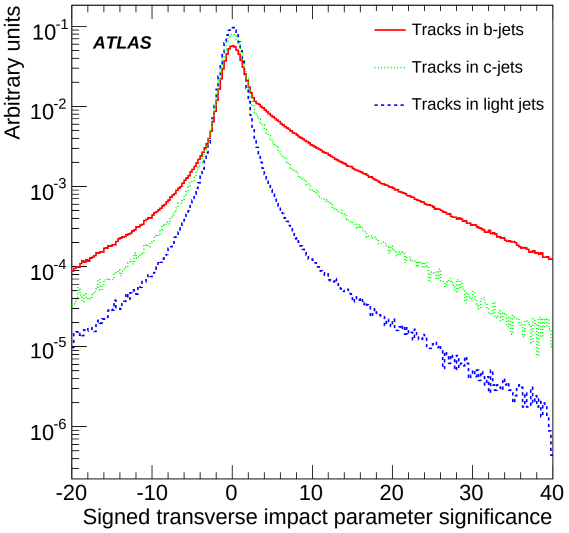

Figure 1: Distribution of d_0/σ_{d_0} for *b*-tagging quality tracks in *b*-jets, *c*-jets and light jets.

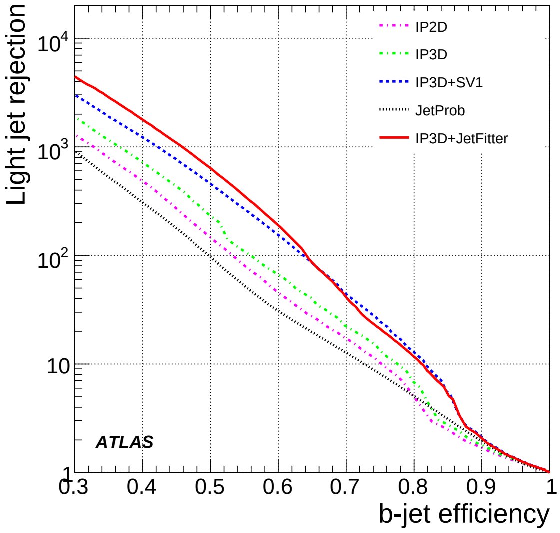

Figure 2: Rejection of light jets versus *b*-jet efficiency in $t\bar{t}$ events with several algorithms.

3. Secondary vertex-based algorithms

To further increase the discrimination between *b*-jets and light jets, the inclusive vertex formed by the decay products of the bottom hadron, including the products of the eventual subsequent charm hadron decay, can be sought. Tracks leading to two-track vertices compatible with a K_s^0, Λ, photon conversion or material interaction are rejected. A simple cut on the distance between the primary vertex and the secondary inclusive vertex can be used as a discriminant. Otherwise, three of the vertex properties are exploited using the likelihood approach described above: the invariant mass of all tracks associated to the vertex, the ratio of the sum of the energies of the tracks participating to the vertex to the sum of the energies of all tracks in the jet and the number of two-track vertices. The resulting weight, SV1, is combined with the IP3D one.

A new algorithm, JetFitter, is also available, which exploits the topological structure of *b*- and *c*-hadron decays inside the jet. A Kalman filter is used to find a common line on which the primary vertex and the beauty and charm vertices lie, as well as their position on this line approximating the *b*-hadron flight path. With this approach, the *b*- and *c*-hadron vertices are not merged, even when only a single track is attached to each of them. The discrimination between *b*-, *c*- and light jets is based on a likelihood using similar variables to the SV1 tagging algorithm above, and additional variables such as the flight length significances of the vertices. The likelihood is split into different categories according to the decay topology.

4. Summary of anticipated *b*-tagging performance

For performance studies, only jets fulfilling $p_T > 15$ GeV/c and $|\eta| < 2.5$ are considered. The expected *b*-tagging performance in $t\bar{t}$ events is shown on Fig. 2 as the rejection power against light jets ($1/\varepsilon_{light}$) versus the *b*-tagging efficiency ε_b. For top studies, $\varepsilon_b = 50\%$ is usually sufficient and a rejection above 500 can be achieved with the most advanced algorithms, while simpler algorithms will provide in early data a rejection of about 80. For processes with a lower cross-section and/or many *b*-jets, a *b*-tagging efficiency of 60% (70%) is more relevant and a rejection of 150 (30) is possible. For reference, the soft muon tagger provides a light jet rejection of 300 for a 10% efficiency on inclusive *b*-jets. The rejection of *c*-jets is naturally limited by the lifetime of charm hadrons: without any specific optimization, a rejection of around 6 is obtained for $\varepsilon_b = 60\%$.

Among the various effects studied in Ref. [2], the impact of residual misalignments in the pixel detector was studied by running the actual ATLAS alignment procedures on a Monte Carlo sample in which the detector elements were slightly shifted and/or rotated accordingly to actual surveys or known fabrication precisions. This is the most realistic case considered so far, and comprises many (but not all) systematic deformations including those caused by the alignment procedure itself. In this case, the light jet rejection is at most 25% lower for the same *b*-tagging efficiency.

It is worth mentioning that the *b*-tagging performance depends strongly on the jet momentum and rapidity. At low p_T, performance is degraded mostly because of larger multiple scattering. This also holds for the high-$|\eta|$ region, where the amount of material in the tracking region increases significantly. In addition the z_0 resolution is degraded at large rapidities because of the pixel detector geometry. In $t\bar{t}$ events and for a given cut on the IP3D+SV1 weight, the tagging efficiency is 60% higher for a 120 GeV/c p_T *b*-jet than for a 20 GeV/c p_T one. For a fixed $\varepsilon_b = 60\%$ in $t\bar{t}$ events, the light jet rejection at $\eta \sim 2.25$ is ten times smaller than at $\eta \sim 0.25$.

Several effects conspire to reduce the *b*-tagging performance as the jet p_T increases above 120 GeV/c. In particular the density of tracks in the core of energetic jets challenges the pattern-recognition ability of the software and of the inner detector itself, and a substantial fraction of very energetic *b*-hadrons decay after the beam-pipe or even after the first pixel layer leading to tracking ambiguities. At very high p_T (above 500 GeV/c), these effects become so critical that dedicated strategies have to be devised for clustering and pattern-recognition. Currently, a light jet rejection of around 10 is obtained for a 40% *b*-tagging efficiency for 1 TeV jets.

5. Measurement of performance in data

While a large effort is put into having a very accurate Monte Carlo simulation, the *b*-tagging performance must be measured in data. Several studies aiming at measuring the *b*-tagging efficiency in di-jet events or in $t\bar{t}$ events have been performed.

The QCD di-jet samples are enriched in heavy flavors by requiring that one of the jets contains a muon. A specific muon+jet trigger has been developed to collect the data, in particular to provide an uniform coverage as a function of the jet p_T. Assuming a rate budget of 1 Hz for this trigger, 100k events are expected for around 30 hours of running time, corresponding to 1 pb^{-1} of data at a luminosity of 10^{31} cm^{-2}s^{-1}. A first method uses Monte Carlo-derived templates of $p_{T,rel}$, the p_T of the muon relative to the jet+muon axis, for *b*-, *c*- and light jets. The *b*-content and the tagging efficiency are extracted simultaneously by fitting the $p_{T,rel}$ distribution of the data with the templates before and after applying a tagging algorithm. The second method employs two samples with different *b*-content and two uncorrelated tagging algorithms, typically the soft muon tagger and one of the track-based or vertex-based methods described in this paper, to construct a system of 8 non-linear equations and 8 unknowns among which are the *b*-tagging efficiency and the *b*-content of the samples. Both methods are working well for jets with $15 < p_T < 80$ GeV/c and can provide *b*-tagging efficiency binned in jet p_T and/or η. A correction, derived from Monte Carlo, is needed in both cases to turn the measured semi-leptonic efficiency into an inclusive one. The total error of both methods is expected to be rapidly dominated by systematic uncertainties. Studies indicates that it should be possible to control the absolute error on ε_b to 6%.

The second approach makes use of the abundant production of $t\bar{t}$ events at LHC and is complementary to the di-jet techniques: a little more data is needed but the tagging efficiency of jets of higher p_T can be measured. One method consists in counting the number of tagged jets in a preselected sample of N $t\bar{t}$ events. Assuming that each event contains exactly two taggable *b*-jets and that there is no mistags, the number of events with one *b*-tagged jet is proportional to $2N\varepsilon_b(1-\varepsilon_b)$ and the number of events with two *b*-tagged jets is proportional to $N\varepsilon_b^2$. Therefore it is possible to extract simultaneously ε_b and the cross-section ($\propto N$). To account for tagging acceptance, *c*-jets, extra jets and non-zero mistag rates, a likelihood is built assuming a value for the light jet efficiency and known fractions of events with the different combinations of *b*-, *c*- and light jets derived from Monte Carlo. Events with three tags are also used. The counting method allows to measure the integrated *b*-tagging efficiency with a relative precision of ±2.7(stat.)±3.4(syst.)% in the lepton+jets channel and ±4.2(stat.)±3.5(syst.)% in the di-lepton channel for 100 pb^{-1} of data at $\varepsilon_b = 60\%$. Another method relies on the identification of a very pure *b*-jet sample by fully reconstructing the $t\bar{t}$ decay chain in the semi-leptonic channel. The *b*-jet on the hadronic side is tagged to improve purity, while the presumed *b*-jet on the leptonic side is unbiased and used as a probe to measure ε_b. Three techniques have been used to identify the probe jet by finding the correct assignment of jets: a topological selection based on reconstructed masses, a likelihood selection using invariant masses, jet and angular information and finally a kinematic fit of all combinations. None of the techniques result in a sample which is 100% pure in *b*-jets, so background subtraction techniques are used. With 200 pb^{-1} of data, the best technique leads to a relative error on ε_b of ±6.4(stat.)±3.4(syst.)% at $\varepsilon_b = 60\%$.

6. Conclusion

A wide spectrum of algorithms has been developed for the identification of *b*-jets in ATLAS. The simpler ones should provide in early data a light jet rejection of 30 for a *b*-tagging efficiency of 60%. Once commissioned, more sophisticated algorithms are expected to achieve a light jet rejection around 150 at the same 60% tagging efficiency. In addition they should allow to operate at a 70% efficiency on *b*-jets while maintaining a reasonable light jet rejection (around 30), which is critical for physics channels with small cross-section and/or high *b*-jet multiplicity. Finally, detailed studies have shown that the *b*-tagging efficiency can be measured directly in data using di-jet or $t\bar{t}$ events. With 100 pb^{-1}, a relative precision of about 5% can be achieved for *b*-jet efficiency. The accuracy with which the light jet rejection can be measured deserves more studies.

References

[1] G. Aad *et al.*, ATLAS Collaboration, *The ATLAS Experiment at the CERN Large Hadron Collider*, JINST 3 S08003 (2008)

[2] G. Aad *et al.*, ATLAS Collaboration, *Expected performance of the ATLAS Detector, Trigger and Physics*, CERN-OPEN-2008-020, arXiv:0901.0512 (2009)