PROCEEDINGS OF SCIENCE

Commissioning of high-performance b-tagging algorithms in pp collisions at $\sqrt{s} = 7$ TeV with the ATLAS experiment

Nicolas Bousson[*]
CPPM, IN2P3-CNRS/Aix-Marseille-Université, France
E-mail: `nicolas.bousson@cern.ch`

For the ATLAS collaboration.

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. Two robust b-tagging algorithms, JetProb and SV0, taking advantage of the impact parameter of tracks or reconstructing secondary vertices have been swiftly commissioned and used for several analyses of the 2010 data: bottom and top quark production cross-section measurements, searches for supersymmetry etc. Building on this success, several more advanced b-tagging algorithms are commissioned using ~ 330 pb^{-1} of the 2011 data. All these algorithms are based on a likelihood ratio formalism to separate the signal (b-jet) from the background (light or in some cases charm jet) using input distributions from simulated events. The accuracy with which the simulation reproduces the experimental data is detailed, as well as the expected improvement in performance achieved with these new tagging algorithms.

The poster reference is ATL-PHYS-SLIDE-2011-431

The 2011 Europhysics Conference on High Energy Physics-HEP 2011,
July 21-27, 2011
Grenoble, Rhône-Alpes France

---

[*] Speaker.

PoS(EPS-HEP2011)421

1. Introduction

Due to their relatively long lifetime, *b*-hadrons can travel a few millimeters in the detector before decaying. *b*-jets can be identified either inclusively by measuring the impact parameters of the tracks (*i.e.* the distance of closest approach of the track to the collision point) from the *b*-hadron decay products, or explicitly by reconstructing the displaced vertices in jet. The semi-leptonic decays of *b*-hadrons can also be used by tagging the lepton in the jet, but is not detailed here.

2. High-performance *b*-tagging algorithms

The IP3D high-performance tagging algorithm uses a likelihood ratio technique in which input variables are compared to pre-defined smoothed and normalized distributions for both the *b*- and light jet hypotheses, obtained from Monte Carlo simulation. The distributions in this case are two-dimensional histograms of the signed transverse impact parameter significance d_0/σ_{d_0} and longitudinal impact parameter significance z_0/σ_{z_0} of tracks in jets, taking advantage of the correlations between the two variables.

To further increase the discrimination between *b*-jets and light jets, the inclusive vertex formed by the decay products of the *b*-hadron, including the products of the eventual subsequent charm hadron decay, can be sought. The decay length significance $L_{3D}/\sigma_{L_{3D}}$ measured in three dimensions and signed with respect to the jet direction is used as a discriminating variable between *b*-jets and light jets, as well as three of the vertex properties: the invariant mass of all tracks associated to the vertex, the ratio of the sum of the energies of the tracks in the vertex to the sum of the energies of all tracks in the jet, and the number of two-track vertices. These variables are combined using a likelihood ratio technique.

Further discrimination can be achieved with another algorithm, called JetFitter, which exploits the topological structure of weak *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 *b*- and *c*-vertices lie, as well as their position on this line, giving an approximated flight path for the *b*-hadron. With this approach, the *b*- and *c*-hadron vertices are not necessarily 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 as in the SV1 tagging algorithm above, and additional variables such as the flight length significances of the vertices.

Thanks to the likelihood ratio method used for IP3D and SV1, the algorithms can be easily combined: the weights of the individual tagging algorithms are simply summed up. The combination JetFitter+IP3D is based on artificial neural network techniques with Monte Carlo simulated training samples and additional variables describing the topology of the decay chain.

Distributions of the output weight of the combined tagger IP3D+JetFitter is shown in Fig. 1 for data and simulated dijet events. Jets are from a sample of leading jets fulfilling $p_{\mathrm{T}} > 20$ GeV and $|\eta| < 2.5$. Fig. 2 shows, as a function of the jet transverse momentum, the distribution of the tagging rate of the IP3D+SV1 algorithm, which is defined as the fraction of jets that are tagged out of those that could be tagged by the algorithm, for a specific choice of its operating point.

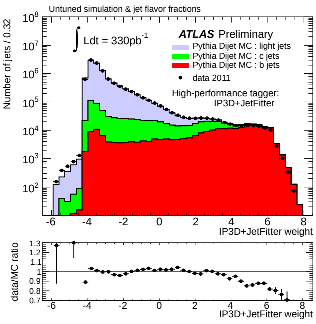

Figure 1: Distribution of the output of the IP3D+JetFitter tagging algorithm for experimental data (solid black points) and for simulated data [1].

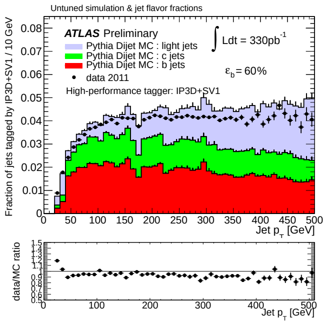

Figure 2: Distribution of the tagging rate for the IP3D+SV1 algorithm at an operating point $\varepsilon_b \approx 60\%$ for data and simulation versus the jet p_{T} [1].

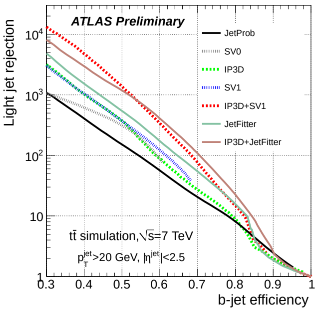

Figure 3: Light-jet rejection as a function of the b-jet tagging efficiency for various tagging algorithms, based on simulated $t\bar{t}$ events [1].

3. Expected performance

Figure 3 shows the light-jet rejection as a function of the b-jet tagging efficiency ε_b for the robust taggers used in 2010 (JetProb and SV0) and for the high-performance b-tagging algorithms. It is obtained by varying continuously the operating point of each tagger, *i.e.* the cut on its output discriminating variable. The jets are from simulated $t\bar{t}$ events. The tagging efficiency is the fraction of jets labeled as b-jets that are tagged, while the rejection is the inverse of the fraction of jets that are labeled as light jets and are actually tagged incorrectly by the algorithm.

4. Conclusions and perspectives

Most of the distributions of the high-performance algorithms input variables as well as their output variables are well reproduced by the simulation, typically at the 10% level. Two important effects which can influence the agreement between data and simulation are the residual discrepancies in modeling the impact parameter resolutions, and the correct modeling of the b, c and light flavour fractions of the jet samples by the Pythia Monte Carlo generator. From the experience with the robust b-tagging algorithms used in 2010, it is known that an agreement between data and simulation at the observed level of 20% or better is sufficient to allow for a successful calibration of the tagging algorithms, and gives confidence that the high-performance tagging algorithms are ready to be used in physics analyses as soon as their performance has been calibrated in data.

These new algorithms are expected to bring a substantial gain for most physics analyses, since the expected mistag rate at a given b-tagging efficiency is projected to be two to five times lower than what is achieved with the robust b-tagging algorithms. In addition, since these algorithms can be operated at a higher b-tagging efficiency ($\geq$ 70%) while keeping an acceptable light jet rejection, they are also very promising for searches of new physics with low production cross sections.

References

[1] ATLAS Collaboration, ATLAS-CONF-2011-102, http://cdsweb.cern.ch/record/1369219 (2011)