IL NUOVO CIMENTO Vol. 37 C, N. 6 Novembre-Dicembre 2014
DOI 10.1393/ncc/i2015-11849-x

Colloquia: LaThuile14

Performance of b-jet identification in ATLAS

J. Jovicevic on behalf of the ATLAS Collaboration
KTH, Royal Institute of Technology - Stockholm, Sweden

ricevuto il 31 Luglio 2014

Summary. — A performant identification of the jets originating from b-quarks is one of the key ingredients which allows for the diverse physics program of the ATLAS experiment. Algorithms for the b-jet identification in ATLAS are exploiting the long lifetime and the high mass of the b-hadrons, as well as the information on the tracks associated with jets. The performance of these algorithms is measured in data to enable reliable usage in the physics analyses. In this article, the performance of the multivariate technique MV1 for b-jet identification, which as its inputs takes information from the other b-jet identification algorithms in ATLAS, is presented. The efficiency of the MV1 algorithm is measured using dileptonic top pair events and is based on a likelihood approach. This approach allows to exploit per-event flavour and momentum correlations between the two jets. Correction factors which take into account differences in the b-jet identification efficiencies in simulation and data are derived as a function of jet transverse momentum and pseudo-rapidity. All the results are derived using the proton-proton collision dataset at centre of mass energy of $\sqrt{s} = 8\,\text{TeV}$ corresponding to an integrated luminosity of $\mathcal{L} \approx 20.3\,\text{fb}^{-1}$.

PACS 29.85.Fj – Data analysis.

1. – Introduction

Identification of the jets originating from b-quarks is an important part of the Large Hadron Collider (LHC) physics program. In proton-proton collisions in ATLAS [1], b-quarks can be produced either directly or in decays of heavier particles. In the direct parton interactions, b-quarks can be produced through flavour creation, flavour excitation and gluon splitting mechanisms. As a decay product, they can originate from top quarks, Higgs bosons, but also from particles which are predicted by some beyond Standard Model (SM) theories such as Supersymmetry. The branching ratio of a top quark decaying to a b-quark is close to unity. The SM Higgs boson decays to $b\bar{b}$ pair 56% of the time. Using b-tagging is crucial not only for searches of signal which contains b-quarks. It is equally important for suppression of background processes which contain b-jets.

2. – Identification of the b-jets in ATLAS

Once b-quarks are produced, they hadronize forming jets with B hadrons. Some of the main properties of B hadrons are

- Large mass (typically 5-6 GeV);
- Long lifetime (around 1.5 ps) and therefore long decay length, typically around 5 mm for a particle of 10 GeV energy. The consequence of the long decay length is a presence of a secondary vertex in such events in the detector;
- The secondary vertices generate displaced tracks which have a large impact parameter. The impact parameter is defined as the closest distance between line associated to the track and the primary vetex;
- B hadrons have a chance of leptonic decay, therefore a soft lepton may be expected nearby a b-jet.

All these characteristics of B hadrons and their decays can be exploited in ATLAS due to its excellent tracking and vertexing performance.

There are 3 main types of b-jet identification algorithms in ATLAS: impact parameter (IP) based, secondary vertex based and decay chain reconstruction based. One of the IP based algorithms is the IP3D. It takes into account both transverse and longitudinal impact parameter significance to discriminate between b and non-b jets. SV0 and SV1 are secondary vertex based algorithms and aim at reconstructing a displaced vertex. They exploit track based invariant mass of the vertex and a flight length significance. The secondary vertex based algorithms have a low mistag rate, but limited efficiency. JetFitter is a decay chain reconstruction algorithm which aims at reconstructing the full hadron decay chain (from b and c quarks). It takes the most performant track and vertex variables as inputs for a neural network and provides the likelihood for a given jet to be b, c and light flavour (u, d, s or gluon).

In order to simultaneously achieve higher rejection of light jets and cover a wider range of b-tagging efficiencies, the individual algorithms are combined. MV1 is a multivariate technique which takes into account inputs from IP3D, SV1 and the combination of IP3D and JetFitter, and properly treats the input correlations. The efficiency of the MV1 b-tagging algorithm depends on a cut on the output weight distribution (w). Figure 1 shows the rejection of the light flavour jets as a function of b-jet tagging efficiency in a sample of simulated $t\bar{t}$ events for various b-jet identification algorithms. It can be seen that the MV1 algorithm shows the best performance. In a following text, MV1 will be the only tagging algorithm of interest.

3. – Calibration techniques

To account for the differences between simulation and real data, the b-tagging algorithms need to be calibrated to data. Several methods have been developed to measure b-jet efficiency, c-jet efficiency and mistag efficiency. The calibration results are presented in a form of data/simulation efficiency scale factors $SF = \frac{\epsilon_b^{data}}{\epsilon_b^{simulation}}$.

The sample in which b-jet tagging efficiency is measured has to be enriched with b-quarks. This can be obtained by selecting a QCD dijet sample with muons or a sample enriched with $t\bar{t}$ events. A high purity of $t\bar{t}$ events can be obtained by requiring two

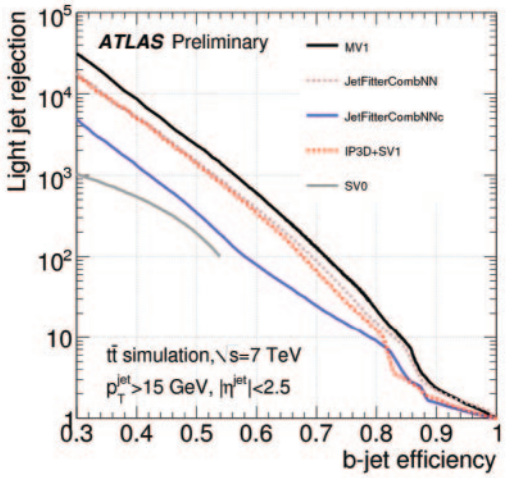

Fig. 1. – The rejection of light flavour jets as a function of b-jet tagging efficiency in a sample of simulated $t\bar{t}$ events for various b-jet identification algorithms [2].

oppositely charged leptons and 2 or 3 jets in the final state. In this case, both W bosons from top quarks ($t \to Wb$) decay leptonically. Such a selection is used for the PDF (probability density function) based calibration method and more details can be found in [3]. Figure 2. shows the p_T spectra of jets in the $e\mu + 2$ jet final state. The excellent agreement between data and simulation can be observed.

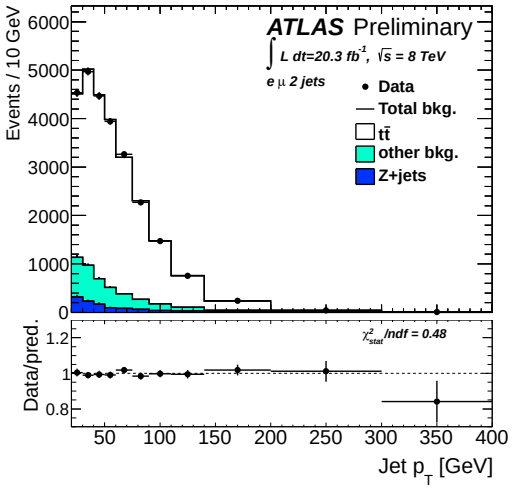

Fig. 2. – Jet p_T distribution for the $e\mu + 2$ jet selection. The "other" background component contains contributions from single top, diboson and lepton fakes. All processes are normalised using dedicated control regions in data. The uncertainties are statistical only. Excellent data and background agreement within uncertainties can be observed [3].

4. – PDF-based calibration method

The PDF based calibration method is performed using the dileptonic $t\bar{t}$ events. Compared to previously used methods, large gain in precision is obtained by considering the correlations between jets.

In previously used methods we assume that the fraction of tagged events in data f_{tagged} is given by

$$f_{\text{tagged}} = f_b \epsilon_b + (1 - f_b)\epsilon_j. \tag{1}$$

In eq. (1) f_b and $(1 - f_b)$ are fractions of real b jets and fraction of non-b jets. ϵ_j is mistag efficiency. The b-tagging efficiency ϵ_b can be determined by measuring f_{tagged} in data, and taking other parameters from simulation. In this approach jets are considered as uncorrelated objects.

The PDF based calibration method aims to exploit the per event jet correlations. For example, in the case of events with 2 jets, we can measure a fraction of tagged events in data where only one jet is tagged and when both jets are tagged.

$$\begin{aligned} f_{2\text{ tags}} &= f_{bb}\epsilon_b^2 + f_{bj}\epsilon_j\epsilon_b + (1 - f_{bb} - f_{bj})\,\epsilon_j^2, \\ f_{1\text{ tag}} &= 2f_{bb}\epsilon_b\,(1 - \epsilon_b) + f_{bj}\left[\epsilon_j\,(1 - \epsilon_b) + (1 - \epsilon_j)\,\epsilon_b\right] + (1 - f_{bb} - f_{bj})\,2\epsilon_j\,(1 - \epsilon_j)\,. \end{aligned} \tag{2}$$

In this way additional information is added which can be used to achieve a higher precision of the efficiency measurement. However, a calibration in N bins of some kinematic variable would lead to $2 \times N^2$ coupled equations. Therefore, a likelihood formalism is used to exploit the per event jet flavour correlation. This takes into account the jet kinematic dependence of the b jet identification algorithm allowing the b-tagging efficiency to be measured to a high precision.

Using the probability density functions (PDFs) $\mathcal{P}$, the likelihood for the 2 jets in the event to have a transverse momentum $p_{T,1}$ and $p_{T,2}$ and MV1 weight output w_1 and w_2 is defined as:

$$\begin{aligned} \mathcal{L}\,(p_{\text{T},1}, p_{\text{T},2}, w_1, w_2) = \big[\; & f_{bb}\mathcal{P}_{bb}\,(p_{\text{T},1}, p_{\text{T},2})\,\mathcal{P}_b\,(w_1|p_{\text{T},1})\,\mathcal{P}_b\,(w_2|p_{\text{T},2}) \\ & + f_{bj}\mathcal{P}_{bj}\,(p_{\text{T},1}, p_{\text{T},2})\,\mathcal{P}_b\,(w_1|p_{\text{T},1})\,\mathcal{P}_j\,(w_2|p_{\text{T},2}) \\ & + f_{jj}\mathcal{P}_{jj}\,(p_{\text{T},1}, p_{\text{T},2})\,\mathcal{P}_j\,(w_1|p_{\text{T},1})\,\mathcal{P}_j\,(w_2|p_{\text{T},2}) \\ & + 1 \leftrightarrow 2\;\big]/2, \end{aligned} \tag{3}$$

where $\mathcal{P}_{f_1 f_2}(p_{T,1}, p_{T,2})$ is 2D PDF for jets of flavour f_1 and f_2 to have momentum $p_{T,1}$ and $p_{T,2}$, $\mathcal{P}_f(w, p_T)$ is a PDF for the b-tagging weight for a jet of flavour f at given p_t. f_{bb}, f_{bj}, $f_{jj} = 1 - f_{bb} - f_{bj}$ are possible flavour fractions in the 2 jet case. All PDFs are determined from simulation, except for the b-jet weight PDF, which contains information to be extracted from data. The efficiency of the b jet tagging corresponding to the MV1 weight cut of w_{cut} as given by

$$\epsilon_b(p_T) = \int_{w_{cut}}^{\infty} \mathrm{d}w' \mathcal{P}_b(w', p_T). \tag{4}$$

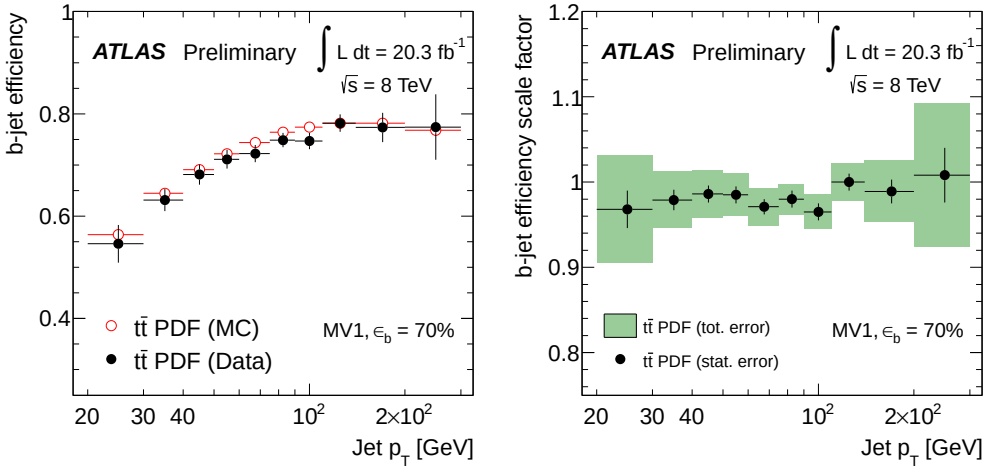

Fig. 3. – Left: Measured b-jet efficiencies in data and simulation as a function of jet p_T The error bars correspond to the total statistical and systematic uncertainties. Right: Data to simulation scale factors as a function of jet p_T. Both statistical only (black lines) and total errors (green shaded region) are shown.

5. – Results

The b-jet tagging efficiency of MV1 is measured using the dataset ATLAS collected at 8 TeV center of mass energy corresponding to an integrated luminosity of $\mathcal{L} = 20\,\text{fb}^{-1}$. The results shown correspond to an overall 70% efficiency of MV1 for b-jets tagging, as evaluated on samples with simulated $t\bar{t}$ events. Figure 3 shows the measured efficiency in data and simulation (left) and data to simulation scale factors (right) as a function of jet p_T.

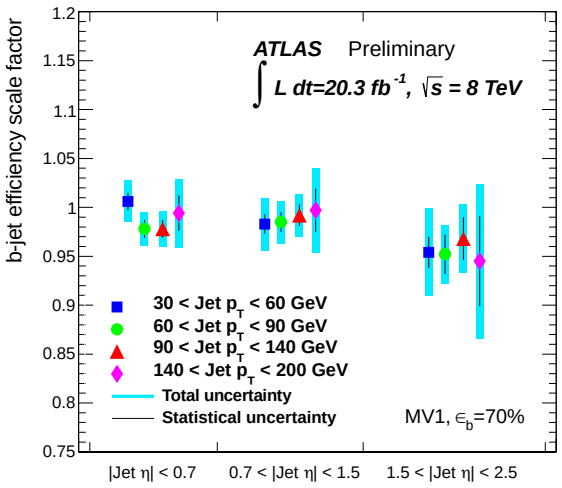

Fig. 4. – The η dependence of the scale factors in different jet p_T bins. The scale factors are computed as a function of p_T in 3 η regions: $0 < |\eta| < 0.7$, $0.7 < |\eta| < 1.5$ and $1.5 < |\eta| < 2.5$. Both statistical and total uncertainties are shown [3].

TABLE I. – *Data to simulation scale factors for the MV1 b-tagging algorithm at 70% b-jet efficiency working point. The statistical, systematics and total uncertainties are shown separately in bins of jet p_T.*

p_T range (GeV)	Combined SF	Stat. error	Syst. error	Total error
20–30	0.968	0.022	0.059	0.063
30–40	0.979	0.012	0.030	0.033
40–50	0.986	0.010	0.027	0.028
50–60	0.985	0.010	0.023	0.025
60–75	0.971	0.009	0.020	0.022
75–90	0.980	0.010	0.015	0.018
90–110	0.965	0.010	0.018	0.020
110–140	1.000	0.010	0.020	0.022
140–200	0.989	0.014	0.033	0.036
200–300	1.008	0.032	0.077	0.084

Table I shows scale factors, statistical, systematic and total uncertainties separately for each p_T bin. All data to simulation scale factors are consistent with unity within their uncertainties. Depending on the p_T bin, the uncertainties are ranging between 2% and 8%. For the $60 < p_T < 140\,\text{GeV}$, data to simulation scale factors are measured with a precision of 2%. Dominant sources of systematic uncertainties are from theory modelling of $t\bar{t}$ and main background processes and of experimental kind due to the jet energy resolution, jet energy scale and mistag rate.

The method allows for binning in any jet kinematic quantity. With such high precision it was possible to check if calibration scale factors exhibit any dependance on jet rapidity η. Therefore, in addition, calibration is performed in 4 bins of jet p_T in range $30 < p_T < 200\,\text{GeV}$ in three $|\eta|$ regions: $0 < |\eta| < 0.7$, $0.7 < |\eta| < 1.5$ and $1.5 < |\eta| < 2.5$. The resulting scale factors in different $|\eta|$ regions are shown in fig. 4. They are tested further as a function of $|\eta|$, inclusively in p_T and no significant dependence is observed.

The PDF based calibration method reduces significantly by 20–50% the overall uncertainty with respect to the previously used methods. Thanks to the more efficient use of data the reduction in statistical only uncertainty is on average 50–60%.

REFERENCES

[1] ATLAS COLLABORATION, *JINST*, **3** (2009) S08003.
[2] ATLAS COLLABORATION, *ATLAS-CONF-2012-043*.
[3] ATLAS COLLABORATION, *ATLAS-CONF-2014-004*.