Measurements of the b-tag performance from data in CMS

Joris Maes *(on behalf of the CMS Collaboration)*
Vrije Universiteit Brussel, Inter-university Institute for High Energies, Pleinlaan 2, B-1050 Brussels, Belgium

Top Quark conference 2008, La Biodola (Elba), Italy, 19-23 May 2008
email: jmmaes@vub.ac.be
Reference Physics Analysis Summary pages CMS PAS *BTV_07_001*, *BTV_07_002* and CMS Note *2006/013*.

In order not to rely on the b-tagging performances predicted from simulation studies, several data-driven methods are proposed to estimate the performance of these b-tagging algorithms. A range of methods is presented to estimate the b-tagging efficiency of certain algorithms and their mistagging probability to identify c-quark, uds-quark and gluon jets as b-quark jets. The methods are applied on real data collisions with different final state topologies. The potential of these methods is demonstrated for data sets with an integrated luminosity ranging from 10 pb^{-1} to 10 fb^{-1}.

1 Introduction

The central tracking device of the CMS detector is being designed to allow for the identification of particles originating from a displaced vertex, for example those present in a hadronic jet resulting from the fragmentation of a b quark. Several algortihms are being developed to differentiate jets according to the flavour of the original parton. Simulation studies have shown that by applying these algorithms we can reach b-tagging efficiencies of about 50% for a mistag probability of about 10% for c-quark jets and 1% for uds-quark or gluon jets.

The aim of this presentation is to overview some methods which can estimate the b-tagging performance, being the b-tagging efficiency and the mistagging efficiencies, directly from real collision data without (where possible) relying on simulated event samples. To demonstrate the potential of all methods presented, fully simulated event samples are being used. Wherever relevant and possible backgrounds and systematic uncertainties of any kind have been included in our estimates.

2 Measuring the b-tag efficiency from Top Quark events

The abundantly produced $t\bar{t}$ pairs at the LHC can be used to isolate jet samples with a highly enriched b-jet content, on which the b-jet identification algorithms can be calibrated. Both the semi-leptonic and di-leptonic decay channel are being explored for this method. The event selection requires leptons which are isolated in the tracker and the calorimeter. A minimum number of jets is requested with calibrated $E_T > 25$GeV. A kinematic fit is applied to the event forcing the W boson and top quark mass constraints. To extract a b-enriched jet sample from the selected events in the semi-leptonic channel, only the b-jet coming from the leptonically decaying top quark can be used because one jet was already tagged as a b-jet in the system of the hadronically decaying top quark. Several observables were identified to discriminate between good jet associations and combinatorial background. Examples of these variables are the transverse momenta of the fitted objects, angles between them, the χ^2 probability of the kinematic fit, the mass difference of the hadronically decaying top quark before and after the kinematic fit, etc. The likelihood ratio information, signal likelihood over background likelihood, of all these variables is combined taking into account the correlations between the observables. The jet combination in the event with the highest combined likelihood ratio is chosen and the b-jet is selected as the jet in the leptonic top quark decay. A selection cut on the combined likelihood ratio is used to purify the jet sample in b-quark jets. The optimal cut is chosen in order to minimize the total uncertainty, including for example estimates of the systematic uncertainty due to radiation effects. The b-tag algorithm is applied on the selected jet sample. The formula is used to estimate the b-tag efficiency ϵ_b

$$\epsilon_b = \frac{1}{x_b}\left[x_{tag} - \epsilon_o(1 - x_b)\right]$$

where x_{tag} is the number of jets which are tagged in the selected jet sample, ϵ_o the probability to tag a non-b-jet and x_b the expected fraction of b-flavoured jets in the selected jet sample. The purity of the selected jet sample is therefore to be estimated from simulation, while all other parameters can be obtained from true data collisions. A similar analysis path is followed for the di-lepton channels. The information from all decay channels is combined, resulting in expected uncertainties on the estimated b-tag efficiencies. These potential uncertainties are shown in Figure 1.

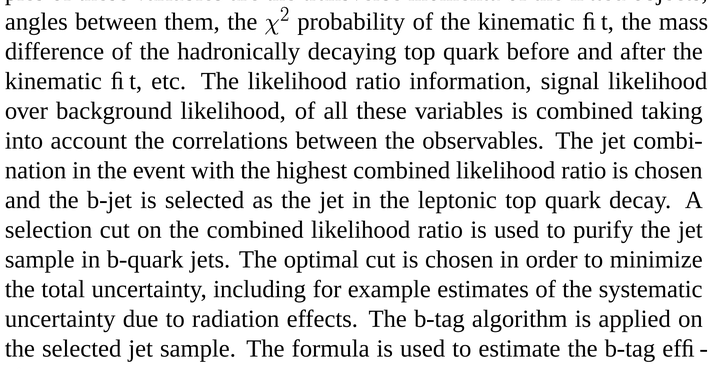

Figure 1: Expected combined uncertainty on the estimated b-tag efficiency in the barrel $|\eta| < 1.5$ (left) and endcap $|\eta| > 1.5$ (right) versus the transverse energy of the jet.

3 Evaluation of mistags for b-tagging using Negative Tags

The Track Counting tagger relies on charged particles tracks with a large 3D impact parameter (IP). Impact parameters can be signed as positive (negative) if the associated tracks are produced downstream (upstream) with respect to the primary interaction vertex, see Figure 2. Tracks with negative impact parameters can be used to evaluate the tagging efficiency from light (uds) quark and gluon jets, this we define as being the mistagging efficiency.

The method is applied on jets with a calibrated $p_T > 20$ GeV from QCD samples. A jet is called taggable if it has at least n associated tracks fulfilling some quality requirements (eg. number of hits, χ^2/ndf, etc.), a value of $n = 1$ is chosen. The taggability is thus simply the ratio between the number of taggable jets and the number of reconstructed jets. For jets of a given flavour, the tagging efficiency is then defined as the number of tagged jets divided by the number of taggable jets. In Figure 2 the distribution of the second highest IP/σ_{IP} is shown for jets in QCD simulated samples.

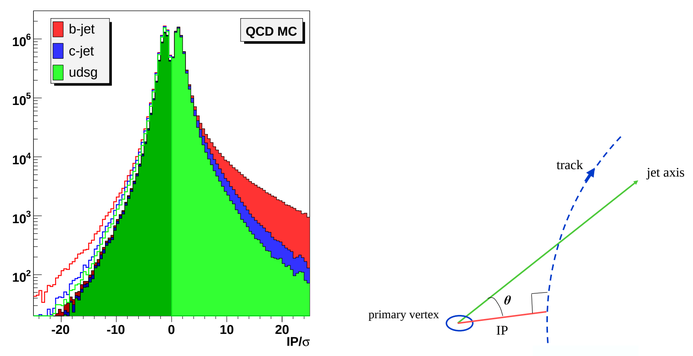

Figure 2: Left: The impact parameter significance of the second highest IP/σ_{IP} track in QCD Monte Carlo jets. Right: Illustration of the sign of the impact parameter of a track. The sign is positive (negative) if the angle θ between the impact parameter direction and the jet axis is smaller (larger) than 90^o.

The mistag efficiency due to light (uds) quark and gluon jets can be evaluated as:

$$\epsilon_{data}^{mistag} = \epsilon_{data}^{-} \cdot R_{light} \quad ,$$

where ϵ_{data}^{-} is the negative tag rate in multi-jet data and $R_{light} = \epsilon_{MC}^{mistag}/\epsilon_{MC}^{-}$ is the ratio between the mistag efficiency of udsg-jets and the negative tag rate of all (udsg+c+b) jets in the simulation. The c and b fractions can be significantly reduced by applying a positive tag veto: the current negative tag jet is rejected if it has any track with IP/$\sigma_{IP} > 4$. Figure 3 shows the values for ϵ_{MC}^{mistag} and ϵ_{MC}^{-} as a function of the jet p_T and $|\eta|$. The ratio R_{light} is about 2.1 for jets with $p_T > 50$ GeV. With these numbers we can obtain the mistag rates on data as

$$\epsilon_{data}^{mistag}(p_T,\eta) = \frac{\epsilon_{data}^{-}(p_T,\eta)}{\epsilon_{MC}^{-}(p_T,\eta)}\epsilon_{MC}^{mistag}(p_T,\eta)$$

For the *medium* working point the total systematic uncertainty on the udsg-mistag rate is estimated to be 8.8%, 7.6% and 5.9% for respectively 10, 100 and 1000 pb^{-1}.

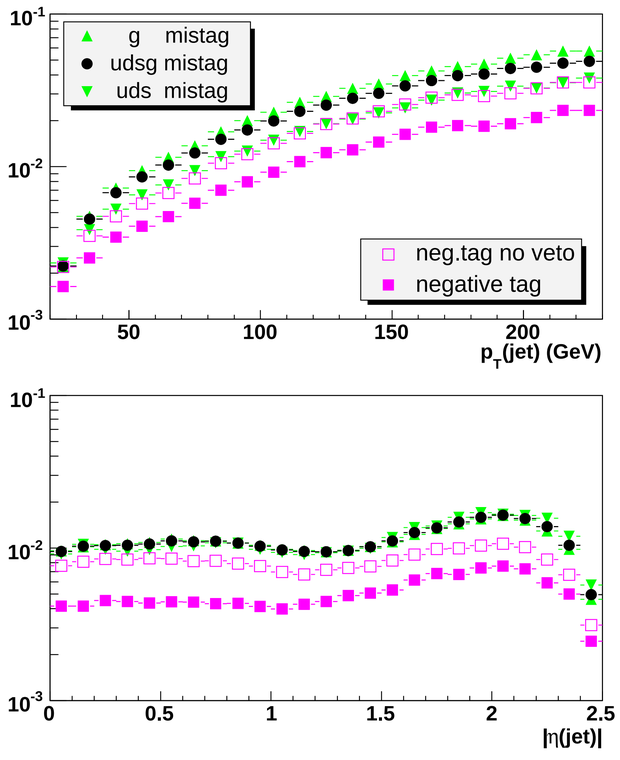

Figure 3: Mistag efficiency and negative tag rate as a function of the jet p_T (upper plot) and of the jet $|\eta|$ (lower plot). The full dots give the udsg mistag efficiency and the full squares give the udsg+b+c negative tag rate. Also shown are the triangles reflecting the tagging efficiency for uds and g-jets separately and the open squares give the negative tag rate if no positive tag veto is applied. Jets from the QCD Monte Carlo are tagged with the Track Counting medium operating point.

4 Measuring b-tagging performance using Jets containing Muons

The analyses are based on samples that have at least two reconstructed jets and a non-isolated muon close to one of the jets with $\Delta R(\mu, jet) <$ 0.4. The variable p_{Trel} is defined as the transverse momentum of the muon relative to the direction of the total muon-jet momentum vector.

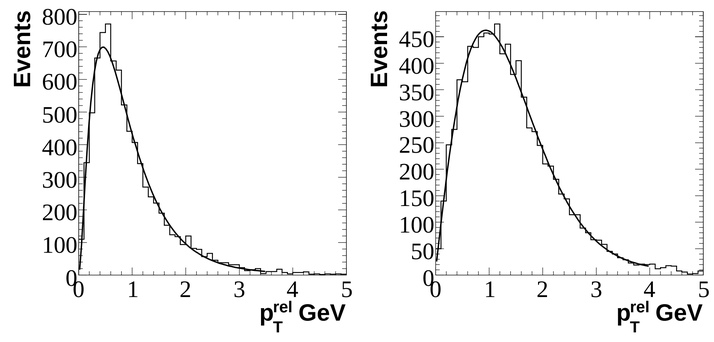

Figure 4: Fitted p_{Trel} templates for c+light jets (left) and b jets (right).

- The *pTrel method* relies directly on a fit to the p_{Trel} distribution of the muon before and after tagging the muon-jet. Templates for the distributions of this variables are shown in Figure 4 where a difference is found between the b jets and the lighter jets. Templates were obtained for different ranges of jet p_T and $|\eta|$. The p_{Trel} distribution of the muons is fitted with a linear combination of the b and c+light jet templates. The process is repeated after tagging the muon-jet. The b-tagging efficiency is calculated as the ratio between the number of b jets after and before tagging, as determined by the p_{Trel} fits.
- The *Counting Method* also relies on the p_{Trel} observable fits but uses additional information derived from data. It assumes that the away-jets in the sample are dominated by light jets, and that the average probability of tagging them can be estimated from light jets data sample with negative impact parameter with respect to the interaction point.
- The third method, so-called *System8 Method* does not rely on the p_{Trel} fits to extract the b jet content of the samples. It consists on solving a system of eight equations constructed from the total number of events in two samples with different b jet content, before and after tagging with two b-tagging algorithms.

The efficiencies are measured as a function of the jet p_T and $|\eta|$ and agree rather well with each other and with the efficiency obtained from the MC truth. This is illustrated in Figure 5. The total uncertainty on the estimated b-tagging efficiency are about equal for all three methods and are equal to 15%, 10% and 5% for respectively 10, 100 and 1000 pb^{-1}. The *System8 Method* performs is slightly less sensitive to systematic uncertainties at start-up for 10 pb^{-1} as it depends less on Monte Carlo simulation, however it produces results in a limited range of p_T.

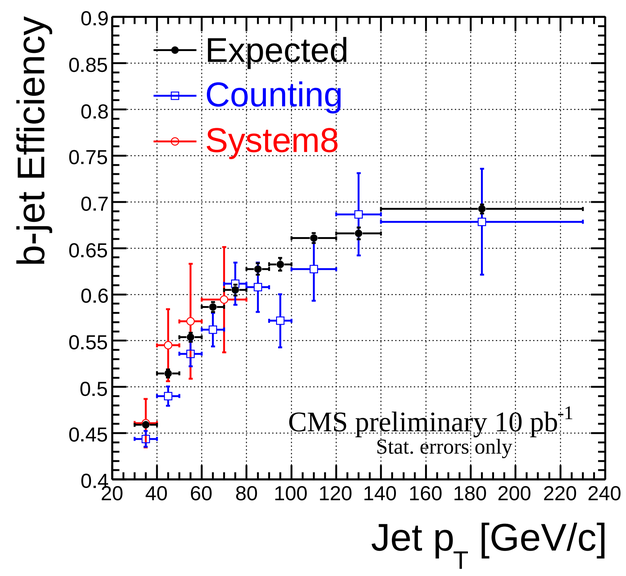

Figure 5: The b-tagging efficiency as a function of the jet p_T as measured with the Counting and System8 methods. The Track Counting tagger with operating point *medium* is used. The measured efficiencies are compared to that obtained from the Monte Carlo truth information. Results are shown with statistical errors only for a corresponding integrated luminosity of about 10 pb^{-1}.

5 Conclusions

A divers range of methods exploiting the properties of jets resulting from heavy flavoured quarks has been presented to estimate the performance of b-tagging algorithms from collision data detected in CMS. Most of them can be applied already in an early start-up scenario with uncertainties around 15% for 10 pb^{-1} of data. With about 1000 pb^{-1} of data these uncertainties can be reduced to about 5% depending on our understanding of the detector performance and the physics in proton collisions at 14 TeV.