Boosted heavy particles and jet substructure with the CMS detector

Ivan Marchesini on behalf of the CMS Collaboration

University of Hamburg, Germany

In the last years, the understanding of jets and jet substructure has become increasingly important, in particular in the context of new physics searches. Many new physics models involve highly boosted hadronically-decaying particles, which result in jet-like objects with large masses and an intrinsic substructure. Discrimination of these heavy jets from ordinary quark and gluon jets is possible through a plethora of new techniques. The understanding of jets can be exploited also for the identification of pileup jets and for the discrimination between quark jets and gluon jets. A sampling of these techniques is discussed together with their validation on collider data recorded in proton-proton collisions at $\sqrt{s} = 8\,\text{TeV}$ with the CMS detector in the year 2012. The commissioning in the boosted regime of algorithms used to identify jets originating from bottom quarks is also discussed. Many studies have highlighted the potential of using jet substructure techniques to improve the sensitivity in physics searches. An overview of recent CMS results employing these techniques is presented.

1. Introduction

The LHC has crossed new energy frontiers in particle physics, where searches for new physics beyond the Standard Model (SM) typically involve objects with very large transverse momenta (p_T). In this regime the resulting decay products for hadronic decays of heavy particles tend to be collimated and can fall within a single jet ("fat-jet"). In this case, selections based on multiple jet searches cannot be applied and jet substructure is necessary to identify ("tag") the particle initiating the jet.

Jets are reconstructed at CMS [1] by clustering the objects ("candidates") reconstructed using a particle flow (PF) approach [2, 3]. The list of neutral and charged particle candidates produced by the PF reconstruction are typically clustered using an anti-kT algorithm of radius R=0.5 (AK5) [4]. For some studies, jets are reconstructed with the Cambridge-Aachen algorithm, either of radius R=0.8 (CA8) or R=1.5 (CA15) [5].

The performance of jet substructure observables used to identify merged hadronic decays of W bosons (W-jets) has been extensively studied at

CMS [6]. Section 2 discusses the results achieved. An algorithm developed to reconstruct highly boosted, hadronically-decaying top quarks [7] is described in Section 3. A wide range of physics processes is characterized by jets arising from the hadronization of bottom quarks (b-jets) and the ability to identify b-jets (b-tagging) is a fundamental prerequisite for several analyses. Section 4 summarizes a first study at CMS, dedicated to the commissioning of b-tagging algorithms in boosted topologies [8]. Two benchmark topologies are considered, with boosted tops and with boosted Higgs decaying to $b\bar{b}$. Section 5 discusses the use of jet shape information to reduce the incidence of jets from pileup (PU) [9]. A likelihood discriminator based on a similar concept, capable of distinguishing between jets originating from quarks and from gluons, is presented in Section 6. Finally, several of the presented tools have already been used in searches for physics beyond the SM, as shown in Section 7.

2. Identification of hadronically decaying W bosons

To study the discrimination of W-jets from gluon- and quark-initiated jets (referred to as QCD jets), a number of topologies are considered. A semileptonic $t\bar{t}$-enriched sample provides a source of W-jets in data. To study the misidentification of W-jets two topologies are analyzed, namely dijet and W+jet, where the W decays leptonically [10, 11].

The main observable to identify W-jets is the CA8 jet mass, which can be improved by grooming methods such as pruning [12, 13]. A good W-tagging performance is achieved selecting pruned jet masses between $60\,\mathrm{GeV}/c^2$ and $100\,\mathrm{GeV}/c^2$. Possible improvements can be achieved by exploiting additional information from jet substructure, such as the mass drop μ [14] or the N-subjettiness τ_N [15]. The performance of various substructure observables combined with the pruned jet mass requirement is shown in Fig. 1 (left). The most performing variable is the N-subjettiness τ_2/τ_1. A combination of the observables in a Likelihood and a Multi-layer Perceptron Neural Network (MLP) multi-variate discriminant is also shown.

A general good agreement between data and Monte Carlo is observed for the substructure variables considered. Small discrepancies in the W-tagging performance between data and simulation can be taken into account applying to simulation scale factors (SF). The SF extraction is done estimating the W-tagging selection efficiency in data and simulation, based on a $t\bar{t}$ control sample. For a W-tagger based on a τ_2/τ_1 requirement and on a pruned jet mass selection the computed scale factor is 0.905 ± 0.08. The pruned jet mass distribution is shown in Fig. 1 (right).

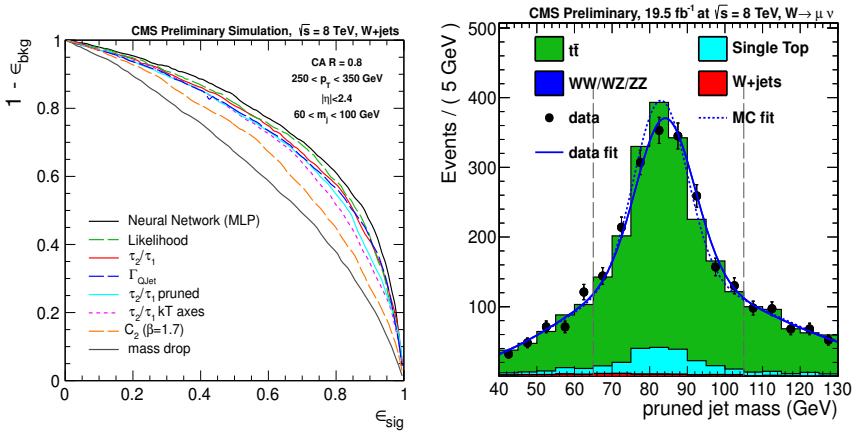

Fig. 1. Left: W-tagging performance for various discriminant observables in a low jet p_T region, 250-350 GeV/c. Right: Pruned jet mass distribution in a semileptonic $t\bar{t}$ sample, for jets satisfying a τ_2/τ_1 requirement.

3. Boosted top jet tagging

The CMS top-tagger is based on the algorithm developed by Kaplan et al. [16] and uses CA8 jets. The algorithm seeks the subjets of the top fat-jet reversing the clustering sequence. With a first primary decomposition the algorithm attempts to split the jet into two subclusters. A following secondary decomposition attempts to split the clusters found by the primary decomposition. The three highest p_T subjets found are examined pairwise and the invariant mass of each pair is calculated. The jet is identified as top if the jet mass is close to the top quark mass, at least three subjets are found and the minimum pairwise mass is greater than $50\,\mathrm{GeV}/c^2$. Good performances are achieved for jets with $p_T > 400\,\mathrm{GeV}/c$, when the decay products are collimated enough to be clustered in a single jet (Fig. 2, left).

The performance of the CMS top-tagger is evaluated in [17], using a semileptonic $t\bar{t}$ control sample and obtaining a data-to-simulation SF= 0.926 ± 0.03. The misidentification probability is measured using an anti-tag and probe method. Events with two or more jets are selected, with the two leading jets having $p_T > 400$ GeV/c. One jet is required to fulfill all the top-tagging requirements, except from asking the minimum pairwise mass to be lower than $30\,\mathrm{GeV}/c^2$, enriching the sample in QCD events. The top-tagging efficiency on the second jet with $p_T > 400$ GeV/c gives a measurement of the misidentification probability (Fig. 2, right).

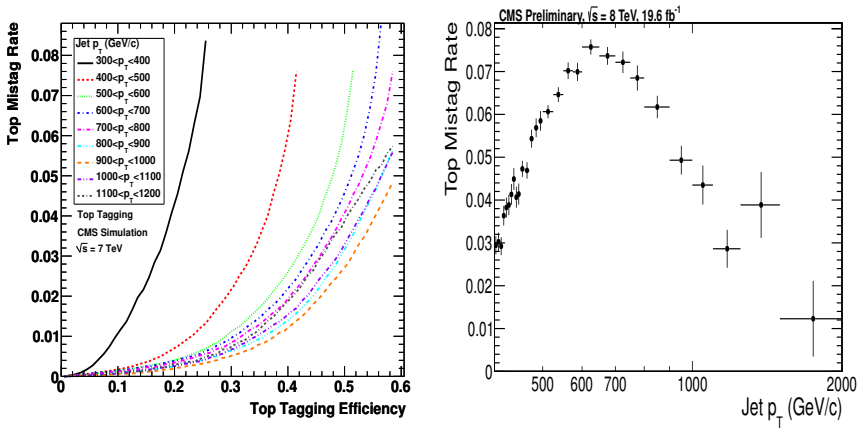

Fig. 2. Left: Simulated mistag rate versus efficiency for the CMS-top-tagger. The efficiency is calculated on seven $Z' \to t\bar{t}$ samples with Z′ masses between 750 GeV/c and 4 TeV/c. The mistag rate is calculated on a QCD dijet sample. Right: Mistag rate of the CMS-top-tagger as a function of jet p_T, measured in data using an anti-tag and probe method.

4. B-tagging in boosted topologies

The b-tagging performance in event topologies with boosted top quarks is studied in samples of simulated $T'\overline{T}' \to tH\overline{T}H$ events with a T′ mass of 1 TeV/c^2. Merged hadronic decays of top quarks are selected using the `HEPTopTagger` algorithm [18], which is based on CA15 jets and produces three subjets. Event topologies with boosted Higgs bosons are studied in samples of simulated $B'\overline{B}' \to bH\bar{b}H$ events with a B′ mass of 1 TeV/c^2 and 1.5 TeV/c^2. Smaller CA8 jets are used in this case and two subjets are clustered using the pruning technique. For both channels the Combined Secondary Vertex (CSV) b-tagging algorithm is adopted [8].

Two b-tagging approaches are considered: (i) application of b-tagging to fat-jets, (ii) application of b-tagging to subjets, which are reconstructed within fat-jets. As exemplified in Fig. 3 (left) for the top channel, subjet b-tagging overall outperforms the fat-jet tagging. Dedicated studies using suitably defined control samples have been performed to validate b-tagging in the boosted environment. The level of agreement present in the boosted regime is found to be as good as in the non-boosted regime for isolated AK5 jets. The SF for the non-boosted and for the boosted regimes are found to be in perfect agreement.

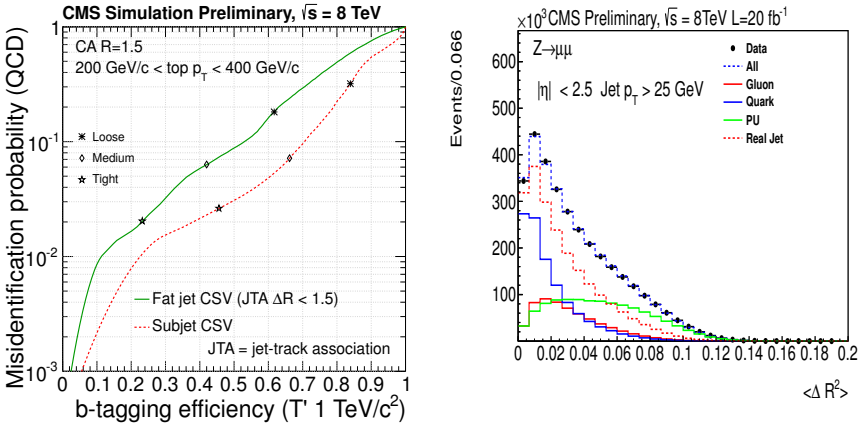

Fig. 3. Left: Performance of the CSV algorithm in simulation on CA15 fat-jets and on subjets of the same jets. The misidentification probability on inclusive QCD jets is shown versus the b-tagging efficiency on top quarks from a pair-produced T′ →tH sample. Right: $\langle \Delta R^2 \rangle$ for jets with $p_T > 25\,\mathrm{GeV}/c$ and $\eta < 2.5$.

5. Pileup jet identification

Identification of PU jets is performed in two ways at CMS, either using vertex information or through the use of jet shape information. As some fraction of charged particles in PU jets is typically not pointing to the vertex of the primary proton-proton interaction, the removal of PU based on vertex information is highly efficient. However, it can only be applied in the central region of the detector, where tracking is available. Jet shape information can be exploited to extend the identification of PU jets beyond the tracker acceptance. The most discriminating variable is shown in Fig. 3 (right), given by the radial extension of the jet, with respect to the jet axis: $\langle \Delta R^2 \rangle = \sum_i \Delta R_i^2 p_{Ti}^2 / \sum_i p_{Ti}^2$, where i runs over the jet PF-candidates.

Shape and tracking information are combined using a boosted decision tree, known as PU-jet multivariate analysis (MVA). The performance of the MVA is evaluated in simulated $Z \to \mu\mu$ events and on data using a control sample of $Z(\to \mu\mu)$+jets, where the jet recoiling against the Z is used as probe. For central jets the performance is excellent and signal efficiencies up to 99% can be achieved for PU rejections of 90-95% (85%) for $30\,\mathrm{GeV} < p_T < 50\,\mathrm{GeV}$ ($20\,\mathrm{GeV} < p_T < 30\,\mathrm{GeV}$). The performance degrades for large $|\eta|$ values, but still the fraction of PU jets can be significantly reduced. The agreement between data and simulation is generally good, with maximum discrepancies up to ∼20% in the forward region.

6. Quark-gluon discrimination

Hadronic jets initiated by gluons exhibit a different behaviour with respect to jets from light-flavor quarks. They are characterized by a higher charged particle multiplicity, by a softer fragmentation function and are less collimated. Observables sensitive to these differences can be combined in a multivariate analysis to develop a quark-/gluon-jet discriminator. This is useful for analyses reconstructing hadronic final states with a specific number of jets from light-quarks or to reduce combinatorial backgrounds in the mass reconstruction of heavy particles decaying into distinct jets.

The first discriminating variable is the multiplicity of the charged PF candidates. The jet width is quantified by the minor axis of the ellipse approximating the $\eta - \phi$ jet shape. Finally, a variable sensitive to the fragmentation function is defined as: $\sqrt{\sum_i p_{T,i}^2 / \sum_i p_{T,i}}$, where the index i runs over the PF jet candidates. The performance of the likelihood-product discriminator of these observables is shown in Fig. 4 (left).

The validation on data is performed on two samples: Z+jets events, which are quark-enriched, and dijet events, which are gluon-enriched. While the quark efficiency is simulated with a 5% precision, the discriminating performance of gluons is worse in data by up to 15%.

7. Searches employing substructure

Several searches at CMS have highlighted the potential of substructure. The CMS top-tagger has been used in searches for $t\bar{t}$ resonances, manifesting themselves in an enhancement of the invariant mass distribution $m_{t\bar{t}}$ of the $t\bar{t}$ system [19]. Several extensions of the SM suggest the existence of such resonances, for instance Kaluza-Klein excitations of particles or additional heavy gauge bosons, referred to as Z′, decaying predominantly to $t\bar{t}$. The fully hadronic final state is selected requiring two top-tagged jets with large p_T. No excess of events above the yield expected from the SM is observed and limits on the production cross section times branching fraction are set (Fig. 4, right). Depending on the specific model, non-SM resonances with masses below 2.1-2.7 TeV/c^2 are excluded at the 95% CL.

In explaining the features of electroweak symmetry breaking for scenarios beyond the SM, boosted final states with vector-like heavy quarks are typically present. In [20] an inclusive search for the pair production of vector-like bottom quark partners B′ that decay into tW, bZ or bH final states is performed. To select highly boosted W, Z or Higgs bosons that are merged into a single jet, a tagger is employed based on the pruned jet mass and on the mass drop observable. No significant excess of events is observed with respect to the SM expectations. A 95% CL limits between

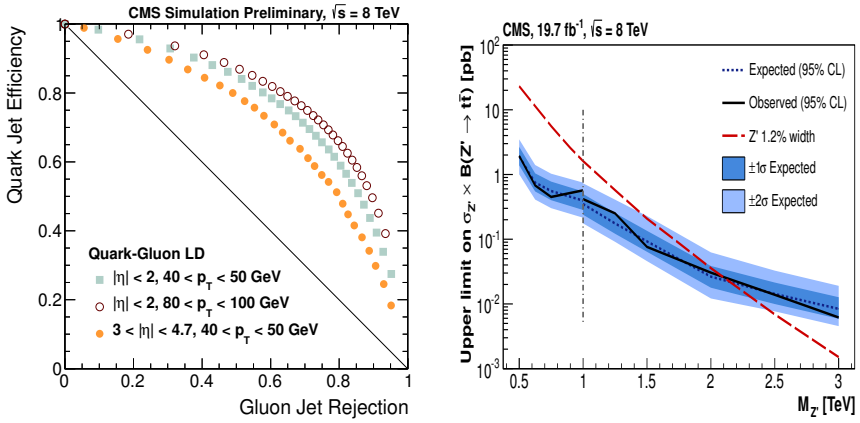

Fig. 4. Left: discrimination performance curves of the quark-gluon tagger, for different p_T and η regions. Right: The 95% CL upper limits on the production cross section times branching fraction as a function of the $t\bar{t}$ invariant mass for Z′ resonances with $\Gamma(Z')/M(Z')$=1.2% compared to theory predictions.

582 GeV/c^2 and 732 GeV/c^2 are set on the B′ mass for various decay branching ratios. In [21] a search for the production of heavy partners of the top quark with charge 5/3 is performed, assuming 100% braching ratio to tW. Both top-tagging and W-tagging are exploited by this study, which sets a lower limit on the mass of the heavy quark of 770 GeV/c^2 at the 95% CL. A search for an heavy partner of the top quark with charge 2/3 has been also performed [22], scanning all the possible branching ratios between three assumed decay modes: bW, tZ, and tH. The search in a final state with a single lepton has been performed using a multivariate analysis, exploiting both W- and top-tagging. Limits between 687 GeV/c^2 and 782 GeV/c^2 at the 95% CL are quoted for the heavy quark mass.

N-subjettiness and the pruned jet mass substructure variables are employed in [23, 24] to select final states with boosted hadronic decays of W and Z bosons, predicted by several models of physics beyond the SM. For instance, a Randall-Sundrum graviton decaying to WW or ZZ or a heavy partner of the SM W boson W′ which decays to WZ. Limits are set on the mass of the heavy particle, depending on the model. A heavy W′ decaying to WZ is excluded for masses up to 1.73 TeV/c^2 at the 95% CL.

8. Conclusions

Jet substructure techniques are discussed, developed for the identification of boosted hadronically decaying particles and for the discrimination

of jets with different flavors or not coming from the primary proton-proton collision. These tools are tested against data collected at the CMS experiment, observing extremely good performances, in particular in the context of new physics searches.

References

[1] S. Chatrchyan *et al.* (CMS Collaboration), JINST **3**, S08004 (2008).

[2] CMS Collaboration, CMS-PAS-PFT-09-001.

[3] CMS Collaboration, CMS-PAS-PFT-10-002.

[4] M. Cacciari, G.P. Salam, and G. Soyez, J. High Energy Phys. **04**, 063 (2008).

[5] Y.L. Dokshitzer, G.D. Leder, S. Moretti, and B.R. Webber, J. High Energy Phys. **08**, 001 (1997).

[6] CMS Collaboration, CMS-PAS-JME-13-006.

[7] CMS Collaboration, CMS-PAS-JME-10-013.

[8] CMS Collaboration, CMS-PAS-BTV-13-001.

[9] CMS Collaboration, CMS-PAS-JME-13-005.

[10] CMS Collaboration, CMS-PAS-EXO-12-024.

[11] CMS Collaboration, CMS-PAS-HIG-13-008.

[12] S.D. Ellis, C.K. Vermilion, and J.R. Walsh, Phys. Rev. D **80**, 051501 (2009).

[13] S.D. Ellis, C.K. Vermilion, and J.R. Walsh, Phys. Rev. D **81**, 094023 (2010).

[14] J.M. Butterworth, A.R. Davison, M. Rubin, and G.P. Salam, Phys. Rev. Lett. **100**, 242001 (2008).

[15] J. Thaler and K. Van Tilburg, J. High Energy Phys. **03**, 015 (2011).

[16] D.E. Kaplan, K. Rehermann, M.D. Schwartz, and B. Tweedie, Phys. Rev. Lett. **101**, 142001 (2008).

[17] CMS Collaboration, CMS-PAS-B2G-12-005.

[18] T. Plehn and M. Spannowsky, J. Phys. G **39**, 083001 (2012).

[19] S. Chatrchyan *et al.* (CMS Collaboration), arXiv:1309.2030 [hep-ex].

[20] CMS Collaboration, CMS-PAS-B2G-12-019.

[21] CMS Collaboration, CMS-PAS-B2G-12-012.

[22] CMS Collaboration, CMS-PAS-B2G-12-015.

[23] CMS Collaboration, CMS-PAS-EXO-12-021.

[24] CMS Collaboration, CMS-PAS-EXO-12-024.