1. Introduction

1) It is becoming increasingly important for new physics searches to identify single jet objects which originate from and contain the decay products of a hadronically decaying massive W boson produced with high transverse momentum.

2) It is possible to improve the signal efficiency by using the hadronic decay modes.

3) Jet substructure techniques are used for identifying single jets that have originated from highly boosted, hadronically decaying bosons and distinguishing them from the usual quark- and gluon-initiated ”QCD” jets.

4) Employed jet substructure techniques for identifying (”tagging”) merged W and Z bosons. These include searches with merged W bosons in hadronically decaying $t\bar{t}$ final states, single and pair produced merged V=W, Z bosons in the dijet final state, and searches in the VV final state, where one of the vector bosons decays leptonically.

5) In these searches, various observables have been used for identifying merged vector bosons. In this study, we investigate the performance of jet substructure observables that are used for identifying merged W bosons, so-called ”W-jets”.

Outline

In Sec. 3, we discuss the data and simulated samples used in this study.

In Sec. 4, the reconstruction methods, jet reconstruction inputs, and jet clustering algorithms are introduced and the event selection is described.

In Sec. 5, we discuss the performance of various substructure variables for different signal benchmarks and different kinematic regimes. We compare different observables and identify the most performant ones.

In Sec. 6, we compare the substructure observables in data and simulation in dijet, W+jet and $t\bar{t}$ samples.

In Sec. 7, efficiencies and mistag rates are computed in dijet and $t\bar{t}$ data samples.

In Sec. 8, we summarize our results from this study

2. CMS detector

1) A 3.8 T superconducting solenoid of 6 m internal diameter.

2) Within the field volume are the silicon tracker, the crystal electromagnetic calorimeter (ECAL), and the brass-scintillator hadron calorimeter (HCAL). The muon system is installed outside the solenoid and embedded in the steel return yoke.

3) The CMS tracker consists of 1440 silicon pixel and 15148 silicon strip detector modules.

4) The ECAL consists of nearly 76 000 lead tungstate crystals, which provide coverage in pseudorapidity $|\eta| < 1.479$ in the central barrel region and $1.479 < |\eta| < 3.0$ in the two forward endcap regions.

5) The HCAL consists of a sampling calorimeter which utilizes alternating layers(交替層) of brass or steel as absorber and plastic scintillator as active material.

6) The muon system includes barrel drift tubes covering the pseudorapidity range $|\eta| < 1.2$, endcap cathode strip chambers ($0.9 < |\eta| < 2.5$), and resistive plate chambers ($|\eta| < 1.6$).

3. Data and simulated samples

1) We aim to distinguish W-jets from quark- and gluon-initiated jets

2) The W+jets topology, where the W decays leptonically, is considered a benchmark process. Additionally, the dijet topology is considered in order to make comparisons in the high pT regime. Finally, we consider a semileptonic $t\bar{t}$-enriched sample, which provides a source of W-jets in data.

3) The data sample for this analysis was recorded by the CMS detector in 2012 at a center of-mass energy of 8 TeV and corresponds to an integrated luminosity of approximately 19.6 fb^{-1}.

4) We define benchmark signal samples which are required to decay into the WW final state and which we consider as the source of W-jets. We consider as the default signal sample a resonance decaying to a pair of longitudinally polarized W bosons.

5) To study the effect of W polarization on substructure distributions we compare the SM Higgs-like couplings case with a purely pseudoscalar Higgs case, which yields only transversely polarized W bosons.

6) The main background Monte Carlo (MC) simulation samples considered are QCD multijets, W+jets, WW/WZ, Drell-Yan, $t\bar{t}$, and single top.

4. Event reconstruction

4.1 Physics objects

1) Jets reconstructed with the anti-kT algorithm of radius R=0.5 (AK5) and jets reconstructed with the Cambridge-Aachen algorithm of radius R=0.8 (CA8)

2) CA8 jets are used as the default jets for identifying merged W bosons because of the increased acceptance for moderately boosted signal W-jets.

3) AK5 jets are used for some supporting selection criteria, e.g., the number of additional jets that originate from a “b” quark.

4) In order to mitigate the effect of multiple interactions occurring in the same bunch crossing, so called pileup (PU) interactions, the PF candidates are cleaned of charged hadrons which are not associated to the primary vertex.
   - An event-by-event jet-area-based correction is applied to remove the remaining pileup energy that is due to neutral particles originating from the other vertices.
5) In this analysis we use jets with pT > 30 GeV. We require |η| < 2.4 so that the jets fall within the tracker acceptance.

6) An accurate E_T^{miss}(missing transverse energy) measurement is essential for distinguishing the semileptonic W+jet sample from QCD backgrounds. The E_T^{miss} is computed from the negative of the vector sum of all PF candidates.

8) Electrons are reconstructed in CMS using a Gaussian-Sum-Filter algorithm (GSF); in addition it is required that each GSF electron must pass the identification and isolation criteria optimized for high pT electrons.
   - Electrons are required to lie in $|\eta| < 2.5$ and tight electrons have pT > 90 GeV, while loose electrons have pT > 35 GeV. Loose electrons are used to veto the presence of any additional isolated leptons in the event.

9) Muons are reconstructed by tracker and global algorithms: one proceeds from the inner tracker outwards, the other one starts from tracks measured in the muon chambers and matches them with the ones reconstructed in the silicon tracker. Muons are identified using the high-pT selection. The selected muon candidates also have to be isolated from charged hadron activity in the detector, requiring that the sum of tracks transverse momentum (I_{tk}) within a cone of $\Delta R = 0.3$ around the muon track, relative to the muon pT, should be I_{tk}/pT < 0.1.
   - We define a tight muon as a muon with pT > 50 GeV and $|\eta| < 2.1$, while a loose muon has pT > 20 GeV and $|\eta| < 2.4$.

4.2 Event-level requirements

1) The $t\bar{t}$ sample is used to study the efficiency of W-tagging, while the dijet sample is used to study the fake rate of W-tagging. Both the W+jet and dijet samples are used to study the discrimination power of different W-tagging algorithms.

2) The dijet and W+jet topologies are chosen to be in the kinematic regime typically considered in searches for new physics. The W+jet sample accesses the low pT regime, while the dijet sample reaches to higher pT. we study the highest pT jet in the event.

3) Then we define bins of jet pT where comparisons can be performed and the signal resonance mass is defined accordingly, as follows:

   - W+jet topology: jet pT = 250-350 GeV ⇔ mresonance = 600 TeV
   - dijet topology: jet pT = 400-600 GeV ⇔ mresonance = 1 TeV
   - dijet topology: jet pT = 1.1-1.4 TeV ⇔ mresonance = 2.5 TeV

   We choose the corresponding signal samples such that the pT distribution in these bins are similar between signal and background.

4.2.1 Dijet selection

1) The dijet sample is collected using the logical “or” of a set of triggers based on requirements on HT = ∑jets pT (pT is the transverse momentum of a jet) and the invariant mass of the two highest pT jets in an event.

2) Events are initially selected by requiring at least two jets with pT > 30 GeV and $|\eta| < 2.4$. The two highest-pT jets are required to have a pseudorapidity separation $|\Delta\eta| < 1.3$, which rejects a large fraction of QCD multijet background in view of a resonance search. Finally, the dijet invariant mass is required to be larger than 890 GeV. Two pT bins are studied in this paper: 400 < pT < 600 GeV, 1.1 < pT < 1.4 TeV.

4.2.2 W+jet selection

1) The main goal of the kinematic selection for the W+jet sample is to isolate a highly boosted topology which is consistent with the W+hard recoil(後坐力) jet system. The W+jet sample, as well as the $t\bar{t}$ sample discussed below, is collected using single lepton triggers.

2) The pT thresholds of these triggers are 24 and 27 GeV for the muon and electron channels, respectively. The pT of the leptonic W and of the jet are required to be greater than 200 GeV. The E_T^{miss} is required to be above 50 GeV (80 GeV) for the muon (electron) channel to suppress the contribution from QCD.
3) Events with additional loose electrons and muons are vetoed, in order to improve the purity of W+jet events.

4) To ensure that the W bosons are back-to-back in the transverse plane: the distance ∆R between the lepton and the jet must be greater than π/2; the azimuthal distance ∆φ between the missing energy and the jet must be greater than 2.0 radians; and the azimuthal distance ∆φ between W_{lep} and W_{had} must be greater than 2.0 radians.

5) Finally, we also apply cuts on additional jet activity in the event to reduce the amount of $t\bar{t}$ background, and the number of AK5 jets with a b-tag that are not matched with the CA8 jet in the event is required to be zero.

4.2.3 $t\bar{t}$ selection

1) To select the $t\bar{t}$ sample, we use the standard kinematic preselection cuts with single lepton triggers but invert(顛倒) the cut on the number of b-tagged AK5 jets that are not matched with the CA8 jet, requiring that there is at least one AK5 b-jet.

2) To increase the statistics, we choose the CA8 jet with the highest mass in the opposite hemisphere of the lepton. This is contrary to the W+jet selection which uses the high pT CA8 jet in the event. The cut on the muon (electron) pT is 50 (90) GeV and the cut on the E_T^{miss} is 50 (80) GeV.

5. Algorithms for W-jet identification

- In this section, we perform simulation-only studies of the various observables used in identifying W-jets.
- First we define the observables under investigation. Then we study the effect of pileup and reconstruction on these observables. We also look at the effect of the W polarization on the observables as well as the different signatures of quark and gluon initiated jets. The behavior at very high pT is also examined. Finally, performance comparisons of the observables are made.

5.1 Substructure observables

1) The mass of a jet is the main observable in distinguishing a W-jet from a QCD jet. Further, the use of a jet grooming method such as filtering, trimming, or pruning improves discrimination by pushing the jet mass for QCD jets towards lower values, while maintaining the jet mass for W-jets at the W-mass.

2) We reconstruct the jet mass using jet pruning, a technique which removes the softest components of the jets. A jet is reclustered using all the particles used to build a CA8 jet, ignoring in each recombination step the softer “protojet” if the recombination is softer than a given threshold $z_{cut} = 0.1$ and forms an angle ΔR wider than $D_{cut} = 0.5m^{orig}/p_T^{orig}$ with respect to the previous recombination step, where m^{orig} and p_T^{orig} are the mass and transverse momentum of the original CA8 jet. The hardness z of a recombination is defined as $z = \min(p_T^i, p_T^j)/p_T^p$ where p_T^i and p_T^j are the pT of the two protojets to be combined and p_T^p is the pT of the combination of the two protojets.

3) we can use additional information about the jet substructure in order to further discriminate W-jets from quark and gluon-initiated QCD jets.

- **Mass drop, μ**: Two subjets are obtained by undoing the last clustering iteration of the pruned jet clustering. The ratio of the highest mass subjet (m_1) and the total pruned jet is defined as the mass drop $\mu = \frac{m_1}{m_{jet}}$.

- **N-subjettiness, $\boldsymbol{\tau_N}$**: N-subjettiness is a generalized jet shape observable. For N candidate subjets of a given jet, we can define the N-subjettiness observables as: $\tau_N = \frac{1}{d_0} \sum_k p_{T,k} \min\left(\Delta R_{1,k}, \Delta R_{2,k}, \ldots, \Delta R_{N,k}\right)$ where k runs over all constituent particles. The normalization factor is $d_0 = \sum_k p_{T,k} R_0$and R0 is the original jet radius. The τ_N observable has a small value if the jet is consistent with having N subjets. Thus, for discrimination of W-jets with 2 subjets and QCD jets consistent with 1 subjet, the ratio τ_1/τ_2 is of particular interest and tends to smaller values for signal W-jets.

- **Qjet volatility, Γ_{Qjets}** : Qjets was introduced as a statistical interpretation of jet trees. A typical jet tree is defined by its cluster sequence; however, the jet can be reinterpreted as a distribution of trees. The process for deriving a distribution of trees is defined by: (1) at every state of clustering, assign a weight to each constituent pair, w_{ij} and (2) generating a random number to choose the $2 \rightarrow 1$ clustering from the available pairs. The default weight is defined as: $w_{ij} = exp\left\{-\alpha\frac{d_{ij}+d^{min}}{d^{min}}\right\}$ where d_{ij} is the distance between the ij pair. The Qjet volatility is defined as the RMS of the distribution over the average jet mass, $\Gamma_{\text{Qjets}} = RMS/< m >$.
- **Generalized energy correlation functions, $\mathbf{C}_{\mathbf{2}}^{\boldsymbol{\beta}}$**: The 3-point correlation function is particularly useful for W-tagging:

$$C_2^{\beta} = \frac{\sum_{i,j,k} p_{Ti}p_{Tj}p_{Tk}\left(R_{ij}R_{ik}R_{jk}\right)^{\beta} \sum_i p_{Ti}}{\left(\sum_{i,j} p_{Ti}p_{Tj}\left(R_{ij}\right)^{\beta}\right)^2}$$

- **Jet charge, $\boldsymbol{Q^k}$** : A measure of the electric charge of the particle originating the jet. It is defined as: $Q^k = \frac{\sum_i q_i\left(p_T^i\right)^K}{\left(p_T^{jet}\right)^K}$ where “i” runs over all particles in a jet. It can be used to provide additional discrimination between quark jets and gluon jets or also to distinguish between, for example, a charged W’ and a Z’ new physics signal.

5.2 Effect of pileup and detector effects