CMS Physics Analysis Summary

Contact: cms-pog-conveners-jetmet@cern.ch 2013/08/19

Pileup Jet Identification

The CMS Collaboration

Abstract

High pileup in LHC collisions can increase incidence of jets by several large factors. To reduce the incidence of jets from pileup and to preserve the rate of good jets, a jet identification based on both vertex information and jet shape information has been developed. The construction of this jet identifier is described and the performances are evaluated using both Z+jets MC simulated samples and Z+jets data collected in the 2012 $\sqrt{s} = 8$ TeV run. The effectiveness of this jet identifier is discussed in the context of jet vetoes and vector boson fusion production.

1 Introduction

The current running of the large hadron collider (LHC) is at such high intensities that multiple proton-proton collisions per bunch interaction occur with high likelihood. In this instance, one is typically concerned about identifying and reconstructing a single primary collision where a physics event of interest occurs amongst the background of the additional proton-proton collisions. Such backgrounds are due to processes that occur with high likelihood like low-p_{T} jet production. These additional collisions are known as pileup (PU). The rate of pileup is quoted in units of the number of additional collisions. The 2012 LHC run at $\sqrt{s} = 8$ TeV had an average pileup rate of 23 additional collisions, with some events exhibiting well over 40 pileup collisions.

In the current CMS detector, some of the sub-detectors also read data in an extended window about the time of the current collision. This allows for pileup from both previous and following proton bunches to affect the reconstructed event. This effect is known as out-of-time pileup (as opposed to in-time-pileup). The influence of out-of-time pileup on the event is much smaller. In this paper both effects are combined and referred to generically as pileup.

To reconstruct pileup in events with the CMS detector a vertex reconstruction is performed on all charged tracks. The resulting number of vertices indicates the level of pileup. The vertex reconstruction efficiency is 0.7 for a pileup vertex; thus, a pileup of 25 corresponds to 17 reconstructed vertices.

In the current running of the CMS detector, pileup exists in ubiquity. The typical p_{T} density of pileup (PU) is roughly 0.7 GeV per unit area (in the η, ϕ plane) per reconstructed primary vertex. For the 2012 running of CMS, this gives a total pileup p_{T} of 10 GeV for a typical anti-k_T jet with radius parameter $R = 0.5$.

The origin of pileup deposits are varied, however most pileup jet are built from low p_{T} QCD jet production resulting from pileup collisions. This implies that the pileup itself is clustered. Additionally, it is known from extrapolations of the inclusive jet cross sections [1] down to low p_{T} that a single jet with a $p_{\mathrm{T}} > 5$ GeV occurs with nearly every collision. Such a large incidence of low p_{T} jets induces a phenomenon whereby the low p_{T} jets combine to form one single high p_{T} jet. The resulting jet formed from overlapping jets is known as a pileup jet.

1.1 Incidence of Pileup Jets

Consider a numerical model for the rate of two overlapping jets. The probability of two overlapping jets with added total p_{T} give by p_{T}, while integrating over both η and ϕ, can be written by

$$p(overlap|p_T) \quad = \quad N_{pu}\left(N_{pu}-1\right)a_{jet}^2\int_0^{p_T} dp_T^{'}\frac{d\sigma}{dp_T}\left(p_T^{'}\right)\frac{d\sigma}{dp_T}\left(p_T-p_T^{'}\right) \tag{1}$$

where a_{jet} represents the area of a jet, N_{pu} is the number of pileup events, and the rightmost integral represents the convolution of the inclusive differential cross section as a function of p_T for two sub-jets having p_{T} values of $p_{\mathrm{T}}^{'}$ and $p_{\mathrm{T}} - p_{\mathrm{T}}^{'}$. The measured cross section [1] can be expressed in the form of a falling exponential as

$$\frac{d\sigma}{dp_T}\left(p_T^{'}\right) \quad = \quad \frac{A}{p_T^5} \tag{2}$$

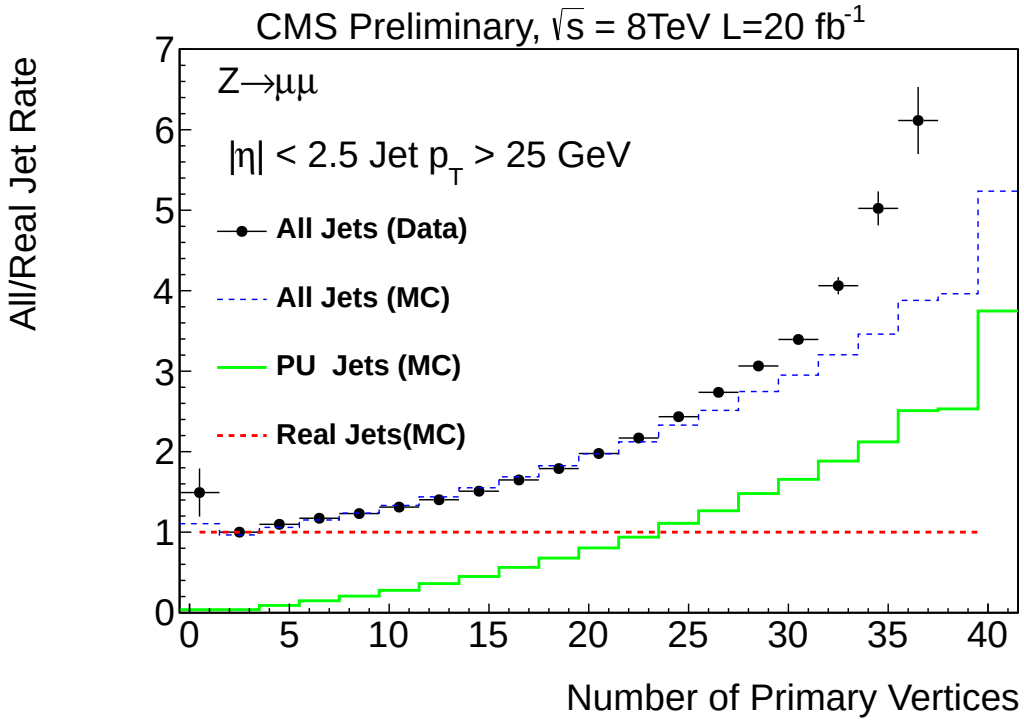

Figure 1: Rate of data and MC PU jets with p_{T} >25 GeV relative to the expected rate of real jets as a function of the number of reconstructed primary vertices.

where the term A is a constant roughly equal to 300 mb. Expanding out the full form of the convolution integral numerically gives an expression of the form

$$p(overlap|p_T) \approx N_{pu}^2 a_{jet}^2 \frac{A^2}{p_T^{6.2}} \quad (3)$$

The key result from this calculation is that the rate of overlapping jets grows quadratically with pileup. If one considers the rate of three overlapping jets or more, this rate grows even more rapidly with pileup. Taking the full form of the convolution, the p_{T} distribution falls more rapidly than the inclusive p_{T} spectrum, making it such that for higher p_{T} objects the rate of overlapping pileup is small. However, the fact that overlapping jets combine to make a larger jet with the equivalent sum p_{T} of all the internal jets allows for a mechanism of pileup jets which can lead to large p_{T} pileup jets. One last observation is that the rate of overlapping jets grows quadratically with the area of the jet cone size. Reducing the area would thus allow for a smaller incidence of pileup jets.

Figure 1 shows the expected inclusive jet spectrum based on the analytic model discussed above. As one extends from ten pileup to forty pileup a clear excess in the growth rate of overlapping pileup jets is present. Of particular importance is the contribution of three and four jet rates, which becomes rapidly larger and extends out to higher p_{T}. The inclusive growth rate for data and pileup is also shown in Fig. 1. A rapid growth is present giving a roughly quadratic increase in the rate of pileup jets.

1.2 Identification and Use of Pileup

Due to the fact that pileup jets primarily come from overlapping jets incurred during pileup interactions, pileup jets exhibit two characteristic features: they are both diffuse and, where charged particle identification is possible, some fraction of the charged particles will not point to the primary vertex. These characteristics allow for the identification of pileup jets in both regions where charged particle tracking is available and regions where jet shape identification is possible. Both vertex and shape information are combined through a multivariate analysis technique, to give a single discriminator targeting the identification of pileup jets. This technique is known as the pileup jet id.

Another technique commonly used in CMS and orthogonal to the pileup jet id is known as charged hadron subtraction. In this technique charged particle flow candidates pointing at another vertex are removed and the jets are allowed to recluster. This technique will not be discussed further.

1.3 Usage examples

For jets with $p_T < 25$ GeV, pileup jets are the largest single source of jets at running conditions of 2012. Their contribution to the total source of jets remains substantial (beyond the few percent level) for jets with $p_T < 40$ GeV. Thus pileup jet identification and removal is critical for jet identification at low p_T. With this in mind, a large number of papers, including all the Higgs papers, have utilized the pileup jet id to mitigate the effect of pileup on jet category migration [2–12], background reduction for searches of vector boson fusion processes [13–21], and construction of a pileup free missing transverse energy [2, 11, 22]. Currently, most analyses that use the pileup jet id select jets with a $p_T > 30$ GeV as a prerequisite for all jets in the analysis. For a few instances, the pileup jet id has been applied on jets with lower p_T [3, 6, 8, 12, 14, 18].

1.3.1 Jet Veto Performance

The initial motivation for the development of the pileup jet id resulted from large event migrations observed between different jet bin categories. This migration is particularly large in the presence of out-of-time pileup [23]. Application of the pileup jet id in conjunction with a well calibrated jet energy scale has reduced the rate of migration for all jets with $p_T > 20$ GeV to below 1%.

This feature was used successfully in Higgs searches where jet categories are used to isolate Higgs signal from additional backgrounds. In one such search, the H→WW search, a b-tag veto on all jets with $p_T > 10$ GeV and an explicit category requiring no jets with $p_T > 30$ GeV are used in order to reduce $t\bar{t}$ background. Migration of signal events out of this category leads to a loss in the sensitivity of $H \rightarrow$WW directly proportional to the rate of migration [3, 6, 8, 12]. Application of the pileup jet id in this analysis allowed for a stabilized jet yield restoring the sensitivity in the high pileup region.

1.3.2 Vector Boson Fusion Background Reduction

Vector boson fusion(VBF) identification poses a particular challenge due to the very low rates and the requirement to tag events with low p_T jets at high η, typically around p_T of 30 GeV and $|\eta|$ of 2.75. These jets suffer from the highest rate of background from pileup jets, making pileup rejection in this region most critical. With the application of the pileup jet id, a clear reduction in the pileup jet rate by more than a factor of ten is present for jets inside the tracker volume, and by more than a factor of two outside the tracker volume. The pileup jet id is currently being used by all analyses where a vector boson fusion is present [24], [25]. For most

analyses a p_{T} cut of $p_{\mathrm{T}} > 30\,\mathrm{GeV}$ is applied, however for the $h \to \gamma\gamma$ a cut of $p_{\mathrm{T}} > 20$ GeV is applied [14, 18].

1.3.3 Missing Transverse Energy

A key use of the pileup jet id is the construction of a pileup insensitive missing E_T. The pileup jet id is the most effective approach at isolating jets which are from pileup. To demonstrate the effect of this on the missing E_T, the performance of the hadronic recoil, $\vec{u}$, in Z$\to \mu\mu$ events is considered. The hadronic recoil is the vector sum in the transverse plane of pileup insensitive objects: tracks from the primary vertex and neutrals in jets with a $p_{\mathrm{T}} > 5\,\mathrm{GeV}$ that pass the pileup jet id. For such a calculation the recoil response with respect to the true recoil is found to plateau at 0.95. If one is to apply the ρ area subtraction from the jets [26], one obtains a plateau response of 0.85, with a response corrected resolution that is the same.

A measure of the sensitivity to pileup is the dependence of the resolution of $u_{\perp}$, the component of the recoil perpendicular to the Z($\to ll$) direction, on the pileup. This resolution is found nearly insensitive to pileup and at high pileup it yields a reduction of 80% in the resolution when compared to that of the conventional missing E_T. It is for this reason that the dominant input to the multivariate particle flow missing E_T is defined by the summing over the tracks from the primary vertex and jets passing the pileup jet id [22].

2 The CMS detector

A detailed description of the CMS detector can be found in [27]. The central feature of the CMS detector is a superconducting solenoid of 6 m internal diameter, providing a magnetic field of 3.8 T. The superconducting solenoid volume is instrumented with the tracker and calorimetry. Gas-ionization detectors embedded in the steel return yoke outside the solenoid are used to reconstruct and identify muons. CMS uses a right-handed coordinate system, with the origin at the nominal interaction point, the x axis pointing to the centre of the LHC, the y axis pointing up (perpendicular to the LHC plane), and the z axis along the anticlockwise-beam direction. The polar angle θ is measured from the positive z axis and the azimuthal angle ϕ is measured in the x-y plane. Charged particle trajectories are measured by the silicon pixel and strip tracker, with full azimuthal coverage within $|\eta|$ <2.5, where the pseudorapidity η is defined as $\eta = -ln[tan(\theta/2)]$. A lead-tungstate crystal electromagnetic calorimeter (ECAL) and a brass/scintillator hadron calorimeter (HCAL) surround the tracking volume and cover the region $|\eta|$ <3. A steel/quartz-fibers forward calorimeter (HF) extends the calorimetric coverage to $|\eta|$ <5.0.

3 Data Samples and Object Definition

The analysis is performed using samples of Z+jets events, with the Z boson decaying to muons. This allows for a clean definition of the recoiling p_{T}, for which jets can be balanced against.

The data events are selected from the full 2012 run at $\sqrt{s}$ = 8 TeV and amount to a total integrated luminosity of 19.8 fb^{-1}. In this running period, the LHC bunch spacing was 50 ns.

Events are required to pass the di-muon trigger, with thresholds on the muon transverse momenta of 17 GeV and 8 GeV respectively. $Z \to \mu\mu$ events are selected by requiring two isolated muons with $p_{\mathrm{T}} > 20$ GeV and $|\eta| < 2.4$, with an invariant mass in a window of 30 GeV around the nominal Z mass. The muon isolation is computed as the sum of the transverse energy of the particles inside a cone of radius ΔR =0.3 around the muon direction divided by the muon

transverse momentum. A correction for the pileup contribution to the energy inside the cone is applied. The resulting isolation is required to be lower than 0.1.

Jets are reconstructed using the CMS Particle Flow (PF) algorithm [28][29], which reconstructs and identifies single particles produced in a collision with an optimized combination of all sub-detector information. The particles are classified into mutually exclusive categories: charged hadrons, photons, neutral hadrons, muons, and electrons. These objects are then clustered into jets with the anti-kT algorithm [30] with a distance parameter R = 0.5. Jet energy corrections are applied to account for the non-linear response of the calorimeters to the particle energies and other instrumental effects. In this analysis jets with p_{T} >25 GeV and $|\eta|$ <5 are considered.

The primary interaction vertex (PV) is defined as the vertex with the highest $\sum p_T^2$ of charged tracks associated to it. Vertices are required to satisfy the good vertex selection: they must have at least 4 tracks and a maximum distance from the nominal interaction point $<$ 24 cm along the z axis.

Data are compared to a Drell-Yan MC sample simulated with Madgraph [31] and Pythia 6.426 [32] for showering. An additional cross check is performed with Drell-Yan MC simulated with Herwig++[33]. This MC sample is corrected to match the true pileup distribution from the 2012 run. For both the Pythia and Herwig++ samples, the pileup is simulated from a minimum bias sample generated with Pythia 6.426. For each event, the number of pileup events is chosen randomly from a Poisson distribution whose mean is distributed over the allowed range of expected pileup. The pileup distribution is matched to the measured CMS instantaneous luminosity through a re-weighting scheme based on the initial sampled distribution. The pileup events are selected randomly from a large minimum bias sample and overlayed at the simulation level allowing for reconstruction of the merged real and pileup event. The out-of-time pileup is simulated for bunch crossings in a time window of $\pm$ 50 ns around the nominal one, which, for 50 ns bunch spacing, corresponds to one bunch crossing before and one after the nominal one.

3.1 Definition of Pileup Jet

The definition of a pileup jet is subject to a number of different interpretations. The definition used here is based on an attempt to isolate good jets with low pileup contamination from jets which have either a large or total contribution from pileup. To perform this, a jet in the MC simulation is identified to be a good jet (not from pileup) if it is matched to a generator level jet found from clustered simulated particles from the hard scatter with $p_{\mathrm{T}} > 8$ GeV within a cone of radius $\Delta R < 0.25$. This matching was determined by taking the minimum ΔR distribution when comparing all generator level jets with p_{T} above 8 GeV. Changing the definition to either a lower p_{T} or larger ΔR yields a small variation in the final performance of the pileup jet identification.

The matching to jets is further divided into jet flavor, separately isolating gluons and quarks. The jet flavor assignment is defined by matching to the closest generated jet where a single parton initiated jet production.

Figure 2 shows the p_{T} and η distributions for pileup jets and non-pileup jets. The contribution of pileup dominates at a p_{T} of 25 GeV. However, this cross over rate grows rapidly higher as the pileup is increased.

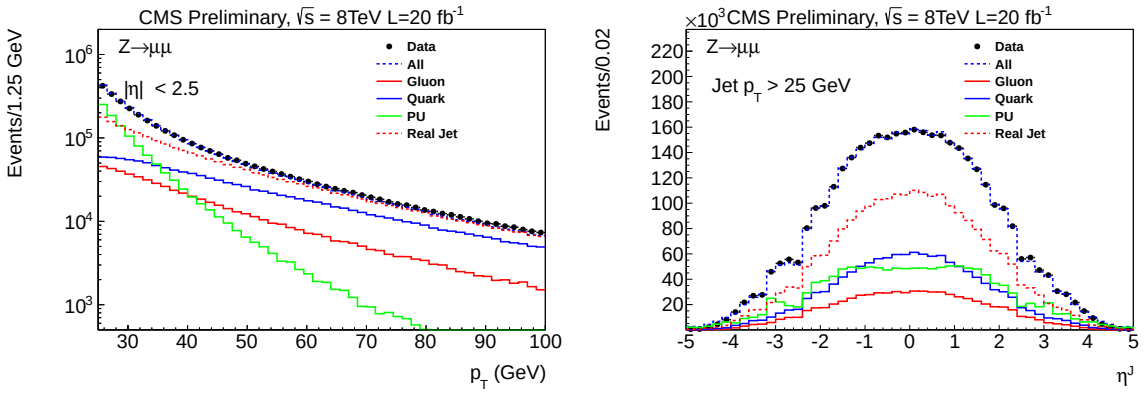

Figure 2: Jet p_T distribution (left) and jet η distribution (right) for all jets having a $p_T > 25$ GeV for the full 2012 dataset.

4 Pileup Jet Id Algorithm

Pileup jet identification (id) relies on two distinct classes of variables:

- vertexing related variables
- shape related variables

Charged PF candidates with tracks contribute to roughly half of the total pileup. Two thirds of the pileup in the tracker volume is charged, the other half of the pileup originates from either neutral candidates from charged particles which are outside of the tracker volume or true neutral candidates where no track is linked. Inside or near the tracker volume a distinct enhancement in the ability to discriminate against pileup is possible by exploiting the compatibility of the jet tracks to come from the PV. Outside the tracker volume, this use of vertexing is not possible, thus jet shower shapes are the only handle to distinguish pileup jets. Since characteristically overlapping pileup jets tend to result in wider jets, shape related variables are precisely designed to target the diffuseness of a jet.

To perform the identification of pileup jets twelve distinct variables, four of which relate to the charged tracking information, are combined in a boosted decision tree (BDT) yielding a single discriminator which can be cut on to give jets of varying pileup contamination. This is known as the Pileup Jet multivariate analysis (MVA).

The training of the BDT and optimization of the jet id working points are done separately in four regions corresponding to the four different regions of the calorimeters: the tracker volume ($|\eta| < 2.5$), the tracker-endcap transition region ($2.5 < |\eta| < 2.75$), the endcap region ($2.75 < |\eta| < 3.0$) and the HF region ($3.0 < |\eta|$). The tracker volume corresponds to the region where tracks are reconstructed. The transition region corresponds to the region where part of the jet is typically within the tracker volume and thus tracking variables can still be used, however their behavior is different to those within the tracker volume. The endcap region corresponds to the region where the HCAL and ECAL endcap are still present. The HF region corresponds to the region where the central jet axis lies in HF.

The training is done on the Z+jets MC sample with target good jets and pileup jets given by the definitions in Sec. 3.

The BDT based pileup jet id represents a baseline for usage by the CMS collaboration.

A cut-based pileup jet id, consisting in a simple jet selection based on the two most discriminating variables, has also been studied. It is used, for example, in [34].

An additional *different* pileup jet id MVA discriminator has been developed for the construction of a pileup insensitive missing transverse energy (missing E_T), known as the particle flow MVA missing E_T [22]. This second MVA discriminator differs from the default Pileup Jet mva in that the jet kinematic variables p_T, η and ϕ are added to the BDT and one inclusive training (as opposed to four η bins) is performed. Plots concerning this specific training are not shown in the rest of this paper.

4.1 Input Variables

To determine the most discriminating variables against pileup jets a systematic scan of the Receiver Operator Characteristic (ROC) of the MVA classifier over a set of approximately eighty variables was performed, first separating them into blocks of similar discrimination and then systematically removing variables until a minimal set retaining most of the discrimination power was determined.

4.1.1 Track related variables

The track related variables in the pileup jet id are constructed to explicitly target the PV the jet is coming from. Four track related variables are used in the computation of the pileup jet id:

- β
- β^*
- d_Z
- $\mathrm{n}_{vertices}$

Each variable explicitly targets a different set of vertexing parameters. All of them are closely related, however each one gives a small gain in performance when added on top.

The variable β is defined as the sum of the p_T of all PF charged candidates originating from the PV divided by the sum of the p_T of all charged candidates in the jet:

$$\beta = \frac{\sum_{i \in PV} p_{Ti}}{\sum_i p_{Ti}} \tag{4}$$

To be identified as coming from the PV, the charged PF candidate must have a $|\Delta_Z| < 0.2$ cm where Δ_Z is the distance with respect to the PV along the z axis.

The variable β^* is defined as the sum of the p_T of all PF charged candidates associated to another PV divided by the sum of the p_T of all charged candidates in the jet:

$$\beta^* = \frac{\sum_{i \in otherPV} p_{Ti}}{\sum_i p_{Ti}} \tag{5}$$

β^* is found to be the most discriminating tracking based variable in the pileup jet id algorithm. β^* and β are decorrelated due to the tracks that are not matched to any vertex.

The variable d_Z is defined as the distance along the z axis between the primary vertex the highest p_T charged candidate in the jet.

Finally, the number of vertices is used in the training of the BDT. Addition of this variable in the BDT allows for varied choice of optimal discriminating variables as the pileup is increased. At

high pileup, vertexing variables have less discriminating power and shape variables become more powerful in discrimination against pileup.

Figure 3 shows the distribution of the four tracking variables for jets in the tracker region. A clear separation is present in both the β and β^* variables. Some disagreement is present in the variables β and β^* resulting from incorrect simulation of the ratio of pileup to real jets. Additionally disagreement is also a result of a smaller resolution term for pileup jets in data when compared with the Monte-Carlo β^*. This disagreement for the signal shape is almost equivalent.

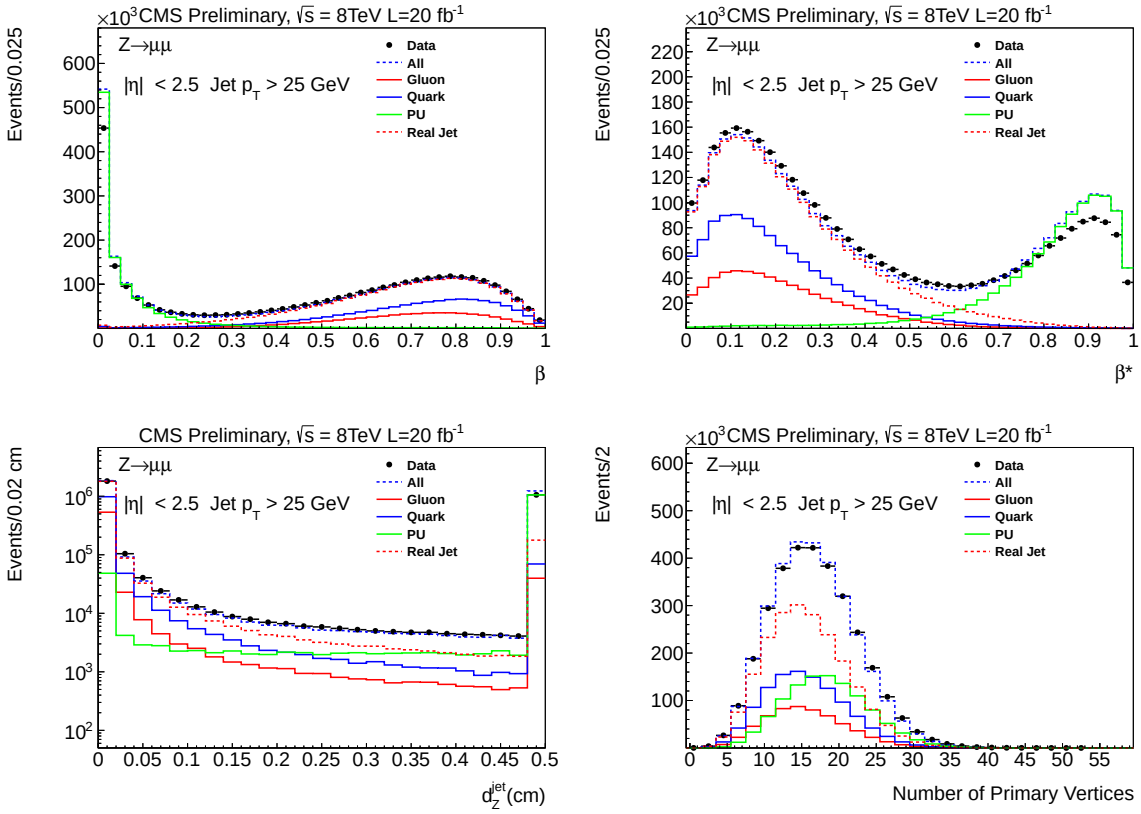

Figure 3: Comparison between jet flavors and pileup for jets with $p_T > 25$ GeV for the four track related variables: β (top-left), β^* (top-right), d_Z (bottom-left), and number of vertices (bottom-right). For the d_Z plot on the bottom-left, the last bin includes all events outside of the plotted axis.

4.1.2 Shape based variables

Shape based variables are related to how the p_T is shared among jet constituents and as a function of their distance from the jet axis. In addition to shape based variables, variables sensitive to the quark-gluon separation are added to allow for an optimized discrimination between pileup and either quark or gluon jets separately.

The shape related variables used in the pileup jet id are

- $\langle \Delta R^2 \rangle$
- $A < (\Delta R) < A + 0.1$
- $N_{charged}$

- $N_{neutrals}$
- p_T^D

The first variable, which is found to be the most discriminating single radial variable, is defined as

$$\langle \Delta R^2 \rangle \quad = \quad \frac{\sum_i \Delta R_i^2 p_{Ti}^2}{\sum_i p_{Ti}^2} \tag{6}$$

where the sum runs over all PF candidates inside the jet and $\Delta R = \sqrt{\Delta\eta^2 + \Delta\phi^2}$ is the distance of the PF candidate with respect to the jet axis. This variable is shown for two different η bins in Fig. 4. The variable for real jets peaks relatively close to zero, whereas for pileup jets it tends to correspond to a value of 0.05, which is slightly smaller than the expected value originating for a uniformly dense jet. The degradation in separation is clear as one extends out to higher η as a result of the coarse granularity in the forward calorimeters. In addition, as the p_{T} of the jet becomes higher, the ΔR^2 tends to get smaller for both pileup jets and non pileup jets. This trend in the current pileup jet id MVA yields an increase in the rate of both pileup jets and real jets at higher p_{T}.

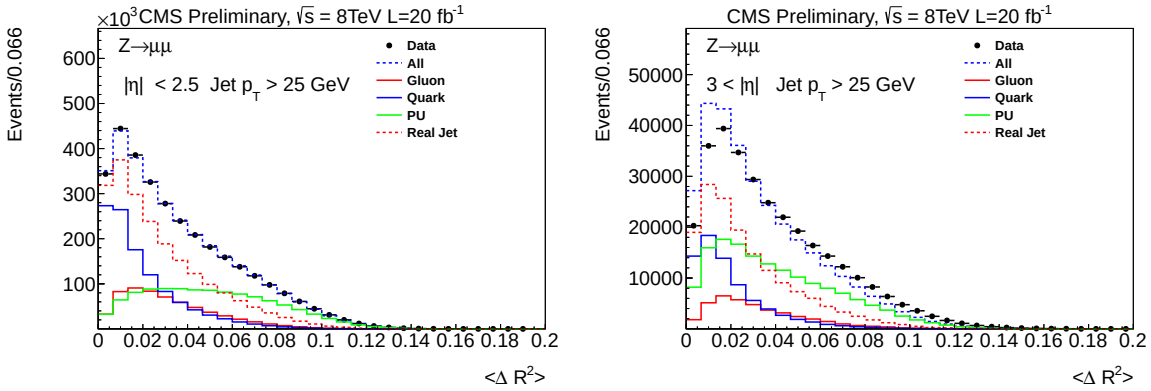

Figure 4: $\langle \Delta R^2 \rangle$ for PF jets with $p_T > 25$ GeV and $|\eta| < 2.5$ (left), and $3.0 < |\eta| < 5.0$ (right).

Enhanced discrimination of pileup comes from adding the full jet shower shape information to the BDT. This is done through the five variables $A < (\Delta R) < A + 0.1$ which consist in the fractional energy deposits in five annuli about the jet axis. They are defined as:

$$A < (\Delta R) < A + 0.1 \quad = \quad \frac{1}{p_T^{jet}} \sum_{i \in A < \Delta R < A + 0.1} p_{Ti} \tag{7}$$

where A is in the 0.1 intervals from 0 to 0.5 about the jet cone axis. These five variables are shown in Fig. 5 for jets in the tracker volume. Comparing them a clear feature is observed: pileup jets contain a large fraction of their energy in the regions $\Delta R = 0.2 - 0.4$ and not in the nearby regions about $\Delta R = 0$. Gluon jets also have a similar characteristic trend, however they tend to be less diffuse than pileup jets.

In addition to these variables, the class of radial variables was studied. They can generically be expressed as

$$W_{ij} \quad = \quad \frac{1}{\sum_i p_T^2} \sum_i \begin{pmatrix} (\Delta\phi_i)^2 \, p_T^2 & (\Delta\eta_i \Delta\phi_i) \, p_T^2 \\ (\Delta\phi_i \Delta\eta) \, p_T^2 & (\Delta\eta_i)^2 \, p_T^2 \end{pmatrix} \tag{8}$$

where the sum is over all PF candidates i in the jet and the $\Delta\eta$ and $\Delta\phi$ terms are with respect to the jet axis. The variables scanned consist in the the jet major and minor axes of W_{ij}, the eigenvalues of W_{ij}, the jet width (quadratic mean of the major and minor) and the η and ϕ moments. They present similar or slightly worse performances or in separating pileup from good jets with respect to other radial variables. Being highly correlated with the radial annuli, their addition to the BDT on top of the $A < (\Delta R) < A + 0.1$ variables provides only a small improvement in the final discrimination. Thus, only the annuli and the most discriminating radial variable $\langle \Delta R^2 \rangle$ are used.

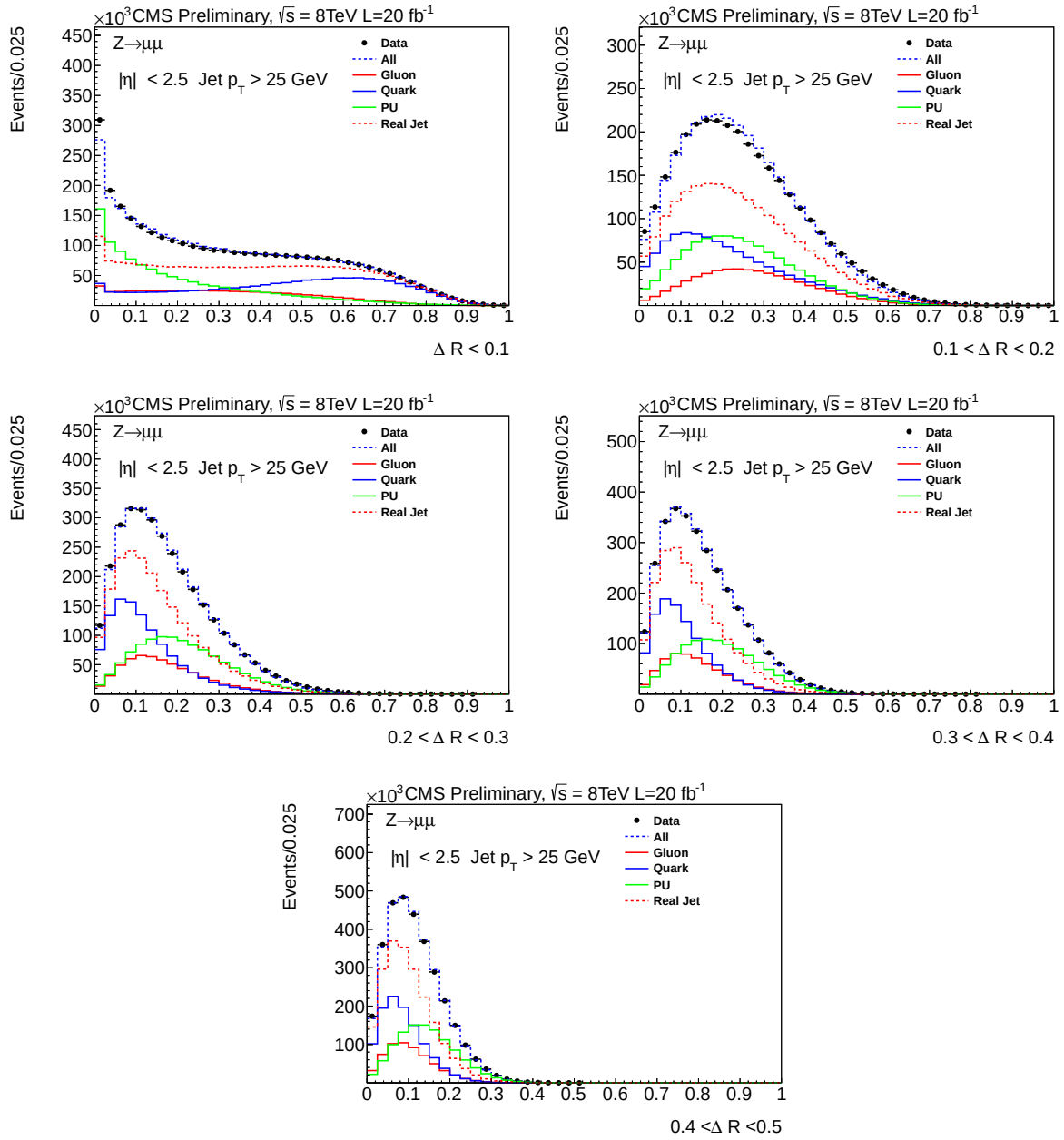

Figure 5: $A < (\Delta R) < A + 0.1$ for PF jets with $p_T > 25$ GeV and $|\eta| < 2.5$ for the concentric rings going from $A = 0.1$ (top-left), $A = 0.2$ (top-right), $A = 0.3$ (middle-left), $A = 0.4$ (middle-right), $A = 0.5$ (bottom).

The charged and neutral multiplicities, $N_{charged}$ and $N_{neutrals}$, are also added to the pileup jet id so as to play the dual role of separately enhancing the quark versus pileup and gluon versus pileup separation by allowing for splitting of quarks and gluons into categories and also by further enhancing the pileup separation. A comparison of the number of charged and neutral particles is shown in Fig. 6. As with the radial variables, the pileup characteristically has more associated candidates than both the quark and gluon jets, with the gluon jets having slightly larger multiplicities.

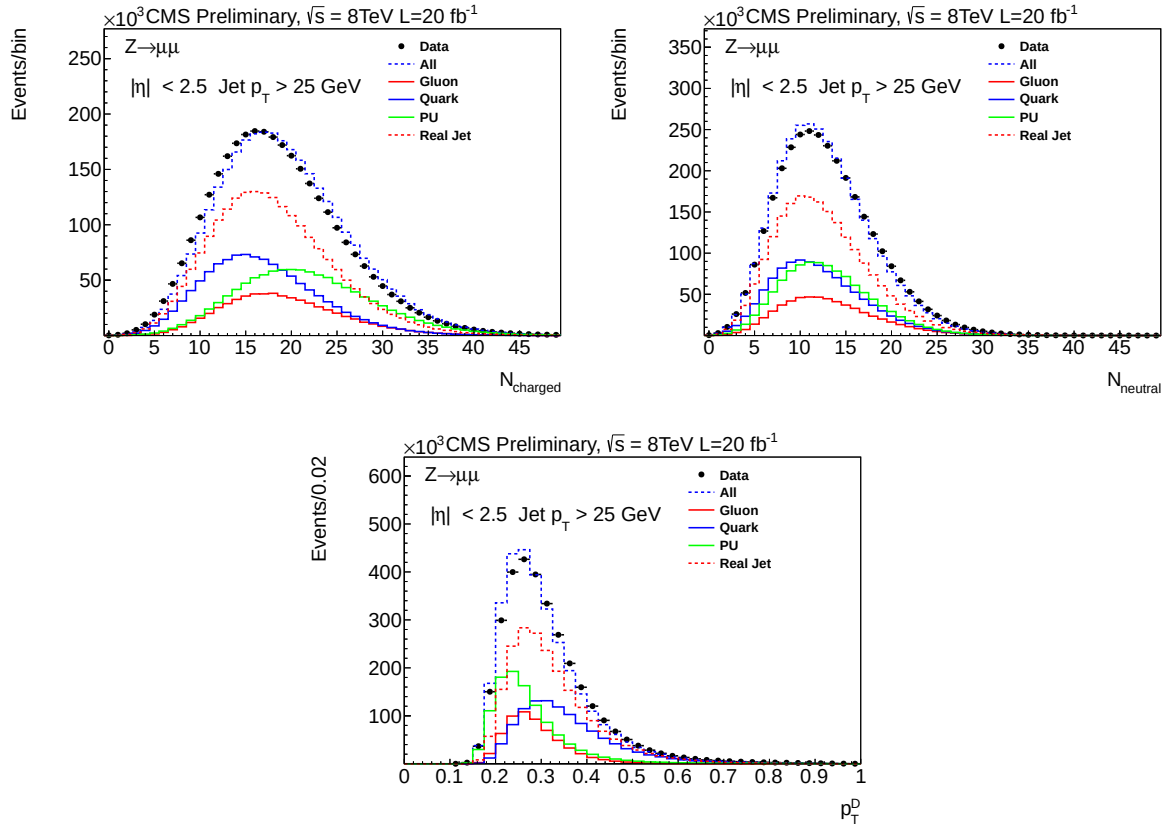

Figure 6: Number of charged particles (left) and neutral particles (right) and p_T^D (bottom) for PF jets with $p_T > 25$ GeV and $|\eta| < 2.5$.

Finally, the variable p_T^D, used in the current CMS quark-gluon discriminator [35], is also considered in the construction of the pileup jet id to enhance the ability to separate quark and gluon jets from pileup jets. In this case, all the neutral candidates of a jet are used, whereas for the CMS quark-gluon discriminator neutral candidates having a $p_{\mathrm{T}} > 1\,\mathrm{GeV}$ are used.

The variable p_T^D is defined as

$$p_T^D = \frac{\sqrt{\sum_i p_{Ti}^2}}{\sum_i p_{Ti}} \tag{9}$$

where the sums run over all the PF constituents inside the jet. Its distribution for PF jets in the tracker acceptance is shown in Fig. 6. As pileup jets tend to have lower p_T^D than gluon jets, the addition of this variable enhances the gluon-pileup separation, particularly at high η.

4.2 Training

To perform the training all the aforementioned shape and tracking variables are added to a boosted decision tree. The training is performed separately for the four different bins in η for all jets with $p_{\mathrm{T}} > 20$ GeV. A further p_T binning was considered separately training the BDT in 10 GeV p_T bins from zero up to 40 GeV, however it was determined that no additional gain in discrimination resulted from such a training and therefore no p_{T} binning was adopted.

Figure 7 shows the pileup jet id BDT output distribution for jets with $p_T > 25$ GeV. Some disagreement between data and MC is observed in the higher η bins.

In the region $2.5 < |\eta| < 3$, the effect is the result of an imperfect modeling of the out-of-time pileup in the simulation and its interplay with the ECAL energy reconstruction. The ECAL data read out consists of 10 consecutive digitizations, corresponding to a sequence of samplings of the signal at 40 MHz. The ECAL amplitude reconstruction uses 5 signal samples and 3 pre-samples for dynamic pedestal subtraction [36]. Amplitude weights are defined so that the pedestal averages to zero only for uniform out-of-time pileup at all bunch crossings. This is not the case in the current MC simulation, where out-of-time pileup is simulated in a window of $\pm$50 ns around the nominal bunch crossing, resulting in an increase of the effective noise.

The data/MC disagreement in HF is mainly related to the Geant4 [37] simulation based on the GFlash parametrization which is currently not satisfactory from the energy flow point of view. An additional contribution comes from the accuracy of the calibration to compensate for response losses due to radiation damage. If the pileup contribution is removed by cutting on the azimuthal angle between the Z boson and the jet, $\Delta\phi(Z, j) > 3.0$, the agreement between data and MC simulation is restored.

The feature about 0.5 in the BDT output in the region $2.75 < |\eta| < 3$ is due to jets with $\beta^* = 0$, $\beta = 0$ and number of vertices in the event < 15.

5 Performance

The performance of the pileup jet identification algorithms is evaluated with simulated $Z \to \mu\mu$ events. As discussed above, certain MVA output values are used to classify the events as either good jets or pileup jets. For each such MVA output value, the probability for a good jet to have a higher value defines the signal efficiency ε(signal), whereas the probability for a pileup jet to have a higher value gives the background efficiency ε(background), which is related to the background rejection $1 - \varepsilon$(background).

The performance is characterized by the ROC curves for the MVA classifier. The results are derived yielding working points for a number of different jet η and p_{T} categories to account for the expected differences in performance. The categorization is analogous in η to the training, with an additional p_{T} bin between 20 and 30 GeV. Furthermore, the efficiencies are determined separately for quark and gluon jets to get a hold of potential efficiency differences due to differences in the jet shapes.

5.1 Efficiency for simulated events

Quark and gluon jets have different properties that affect the discrimination from pileup jets. Most importantly, gluon jets are less collimated than quark jets, and they have a higher charged multiplicity as well as a softer fragmentation function. For the shape-based variables, this implies that gluon jets exhibit more pileup-like properties than quark jets. However, the larger charged multiplicity in conjunction with the softer fragmentation function leads to narrower

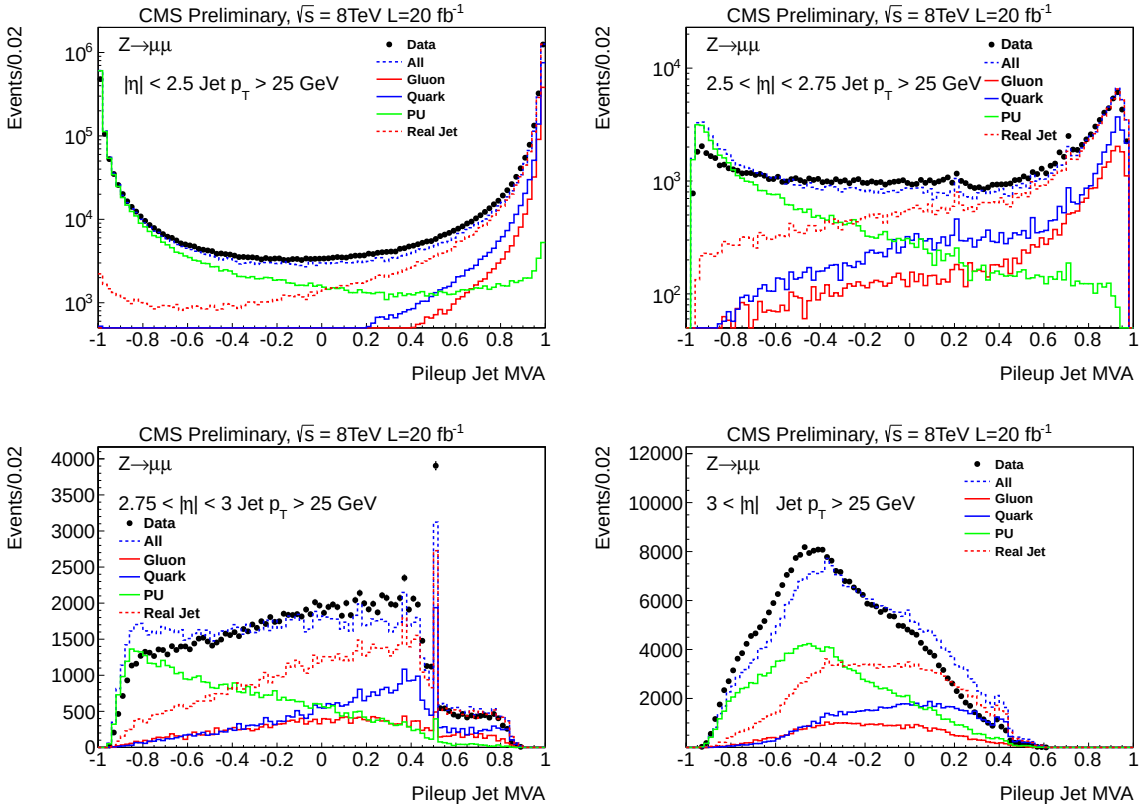

Figure 7: MVA discriminator for particle flow jets with $p_T > 25$ GeV and $|\eta| < 2.5$ (top-left), $2.5 < |\eta| < 2.75$ (top-right), $2.75 < |\eta| < 3.0$ (bottom-left) and $3.0 < |\eta| < 5.0$ (bottom-right). Disagreement in the pileup region of the MVA is present in the region where $2.5 < |\eta|$. This is a known effect, which results from improper simulation of out-of-time pileup.

distributions of the β and β^* variables for gluon jets, resulting in a higher discrimination between gluon and pileup jets at low values of β/high values of β^*.

p_T bin	η bin	Pile-up	Quark	Gluon
$20\,\text{GeV} < p_T < 30\,\text{GeV}$	$\|\eta\| < 2.5$	14.0%	98.6%	99.3%
	$2.5 < \|\eta\| < 2.75$	32.4%	94.0%	93.1%
	$2.75 < \|\eta\| < 3.0$	40.4%	89.5%	84.0%
	$3.0 < \|\eta\| < 5.0$	37.2%	85.1%	73.7%
$30\,\text{GeV} < p_T < 50\,\text{GeV}$	$\|\eta\| < 2.5$	13.1%	99.3%	99.7%
	$2.5 < \|\eta\| < 2.75$	41.3%	95.5%	94.9%
	$2.75 < \|\eta\| < 3.0$	57.8%	93.0%	88.2%
	$3.0 < \|\eta\| < 5.0$	60.3%	87.7%	78.6%

Table 1: Comparison of identification efficiency for quark and gluon jets split in p_T and η bins.

The performance for the different detector regions is given by the ROC curves in Fig. 8. The corresponding identification efficiencies for the given working point can be found in Table 1.

For central jets, signal efficiencies of $\sim 99\%$ are reached for background rejection of 90–95% for $30 < p_T < 50$ GeV and around 85% for $20 < p_T < 30$ GeV.

The fraction of pileup jets can still be significantly reduced in the tracker-endcap transition region. For the given working point, a signal efficiency of ∼95% corresponds to a background rejection of ∼70% (60%) for $20 < p_T < 30$ GeV ($30 < p_T < 50$ GeV). For jets in the endcap and

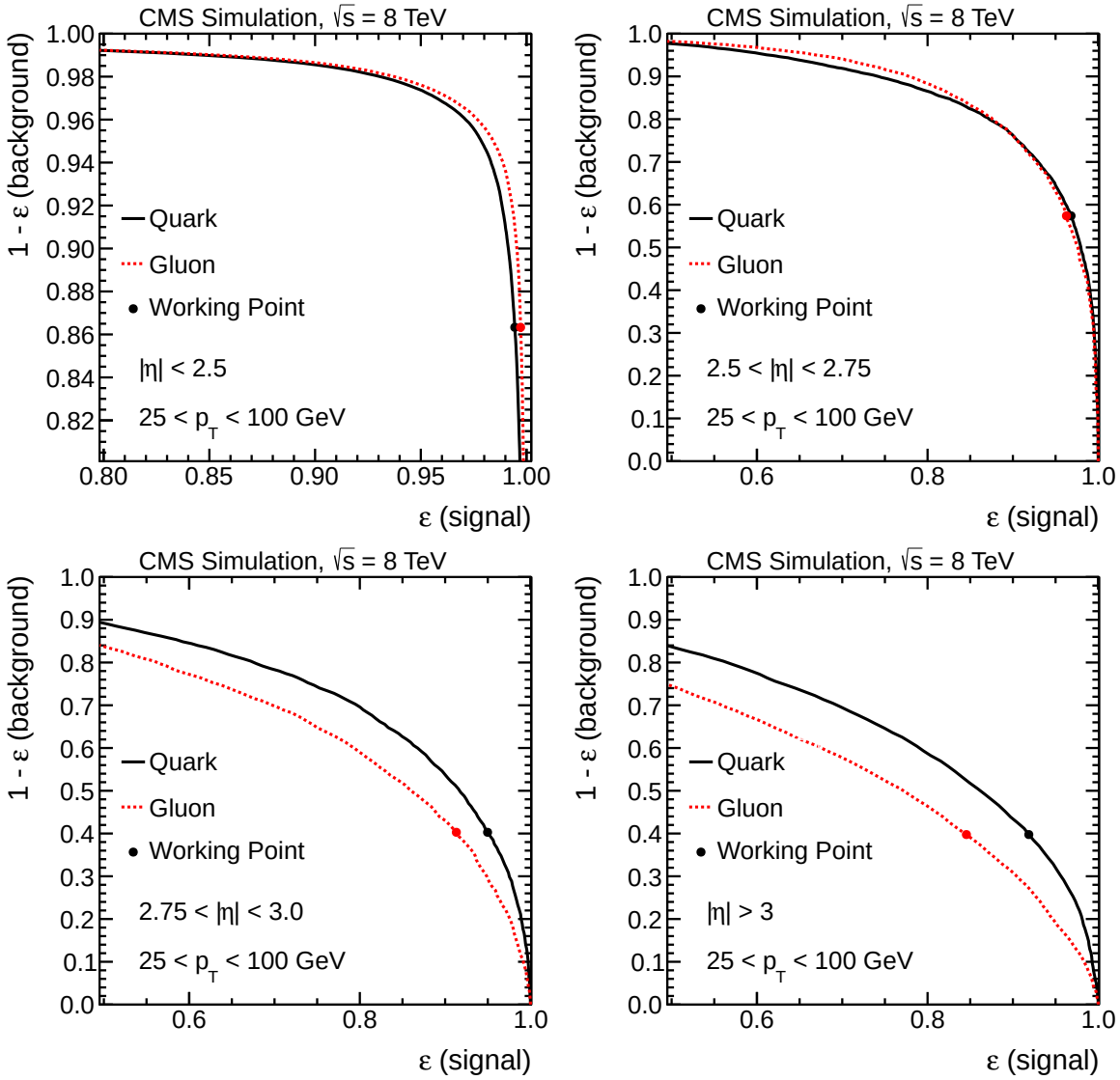

Figure 8: ROC curves for quark and gluon jets with $25 < p_{\mathrm{T}} < 100$ GeV in the four different η regions.

forward regions, the background rejection is ∼60% (40%) for $20 < p_T < 30$ GeV ($30 < p_T < 50$ GeV) at signal efficiencies of ∼90% and ∼80%.

The identification efficiency is higher for gluon jets than for quark jets in the central and the tracker-endcap transition regions, where the β and β^* variables provide the highest discrimination power, and vice versa in the endcap and forward regions. The differences in efficiency for gluon and quark jets are at or below the 1% level for the central and the tracker-endcap transition region; for jets in the endcap and forward regions, the absolute differences are in the range of 5–12%.

To check the effect of using a different showering and hadronisation model, the signal efficiencies are compared for simulated Z+jets events produced with either PYTHIA or HERWIG. For the given working point, the resulting efficiencies are compatible within statistical uncertainties of $\sim$ 1% for central jets. In the tracker-endcap transition region, the efficiencies agree within 2%, and within 5–10% beyond.

5.2 Data/MC scale factors for efficiencies

The efficiency of the pileup jet identification criteria on real jets is checked using a tag-and-probe method on a control sample of $Z(\rightarrow \mu\mu)$+jets events, where the jet recoiling against the Z is used as a probe. In order to reduce the pileup contamination on the probe side, requirements on the balancing between the Z and the hardest jet momenta are applied: the absolute azimuthal separation $|\Delta\phi(Z,j)|$ between the Z and the jet must be larger than 2.5 and the ratio between the jet p_{T} and the Z p_{T} must be between 0.5 and 1.5. With these selections the purity of the control sample is between 80% and 98%, depending on the considered jet momentum and pseudorapidity. Under the assumption that the $\Delta\phi(Z,j)$ distribution is flat for pileup jets, the residual background due to pileup jets in the control sample (both before and after applying the pileup jet id) is estimated from the pileup enriched region with $|\Delta\phi(Z,j)| < 1.5$. The efficiency on real jets is therefore computed as:

$$\epsilon = \frac{N_{passId,sig} - k \cdot N_{passId,bkg}}{N_{all,sig} - k \cdot N_{all,bkg}} \tag{10}$$

where $N_{all,sig}$ is the total number of jets in the control region ($|\Delta\phi(Z,j)| > 2.5$) , $N_{all,bkg}$ is the total number of jets in the pileup enriched region ($|\Delta\phi(Z,j)| < 1.5$), $N_{passId,sig}$ is the number of jets in the control region passing the jet identification, $N_{passId,bkg}$ is the number of jets passing the jet identification in the pileup enriched region and, finally, $k = (\pi - 2.5)/1.5$ is the scaling factor to extrapolate the number of pileup jets from the pileup enriched region to the control sample.

The results of the efficiency measured in data and MC simulation and of their ratio are reported in Fig. 9. As shown, the agreement between data and MC is within 2-10% depending on the jet pseudorapidity and transverse momentum range. The largest data/MC scale factors are observed for the forward region as a consequence of the data/MC differences on the pileup discriminator discussed in Sec. 4. The efficiency of the pileup jet id on pileup jets (estimated in the pileup enriched region defined by $|\Delta\phi(Z,j)| < 1.5$) measured on data is found to be in agreement with MC within $\pm$20% for jets with $p_{\mathrm{T}} > 25$ GeV.

6 Conclusions

Pileup jets are a ubiquitous background under the current 8 TeV running conditions of the Large Hadron Collider. Their presence typically arises from overlapping low p_T jets and grows roughly quadratically with the number of pileup collisions. Due to their unusual formation, pileup jets exhibit distinct features that allow them to be separated from real jets that have originated from either quarks or gluons.

Identification and removal of pileup jets is performed in two ways in the CMS detector, either through the use vertex information or through the use of shape information. Vertex information allows for a highly efficient removal of pileup, however it can only be exploited in the central region of the CMS detector, where tracking is available. Shape information, although less effective than vertexing, extends throughout the whole detector volume and in conjunction with vertex information enhances the ability to identify pileup jets. Shape and vertex information can be combined through a multivariate BDT to give the pileup jet id available for all jets used in CMS.

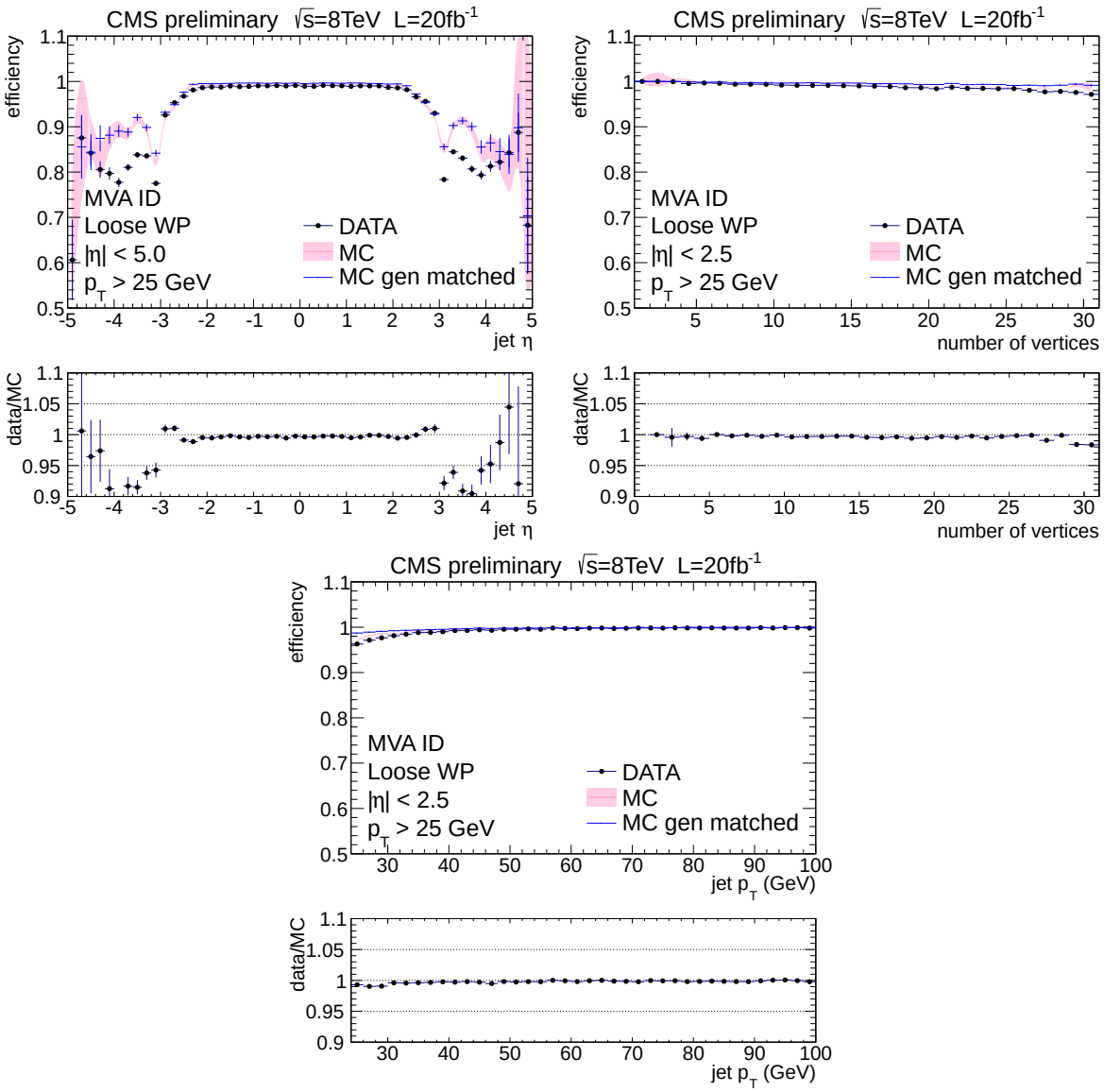

Figure 9: Data-MC comparison of the MVA (loose working point) pileup jet identification efficiency on the $Z(\rightarrow \mu\mu)$+jets sample for PF jets with $p_T > 25$ GeV: the efficiency is shown as a function of the jet pseudorapidity (top-left), as a function of the number of reconstructed vertices for jets with $|\eta| < 2.5$ (top-right) and as a function of p_T for jets with $|\eta| < 2.5$ (bottom).

References

[1] CMS Collaboration, "Measurement of the Inclusive Jet Cross Section in pp Collisions at $\sqrt{s}$=7 TeV", *Physics Review Letters* **107** (2011) 132001, doi:10.1088/1748-0221/3/08/S08004.

[2] CMS Collaboration, "Higgs to tau tau (SM) (HCP)", CMS Physics Analysis Summary CMS-PAS-HIG-12-043, (2012).

[3] CMS Collaboration, "Evidence for a particle decaying to W+W- in the fully leptonic final state in a standard model Higgs boson search in pp collisions at the LHC", CMS Physics Analysis Summary CMS-PAS-HIG-12-042, (2012).

[4] CMS Collaboration, "Higgs to tau tau (MSSM) (HCP)", CMS Physics Analysis Summary CMS-PAS-HIG-12-050, (2012).

[5] CMS Collaboration, "Observation of a new boson with mass near 125 GeV in pp collisions at $\sqrt{s}$ = 7 and 8 TeV", technical report, (2012).

[6] CMS Collaboration, "Search for the standard model Higgs boson decaying to a W pair in the fully leptonic final state in pp collisions at sqrt(s) = 8 TeV", CMS Physics Analysis Summary CMS-PAS-HIG-12-038, (2012).

[7] CMS Collaboration, "Search for a standard-model-like Higgs boson with a mass in the range 145 to 1000 GeV at the LHC", technical report, (2012).

[8] CMS Collaboration, "Search for the standard model Higgs boson decaying to a W pair in the fully leptonic final state in pp collisions at sqrt(s) = 8 TeV", CMS Physics Analysis Summary CMS-PAS-HIG-12-017, (2012).

[9] CMS Collaboration, "Search for the Standard Model Higgs boson in the H to WW to lnujj decay channel in pp collisions at the LHC", CMS Physics Analysis Summary CMS-PAS-HIG-12-046, (2012).

[10] CMS Collaboration, "Combination of standard model Higgs boson searches and measurements of the properties of the new boson with a mass near 125 GeV", CMS Physics Analysis Summary CMS-PAS-HIG-13-005, (2013).

[11] CMS Collaboration, "Search for the Standard-Model Higgs boson decaying to tau pairs in proton-proton collisions at sqrt(s) = 7 and 8 TeV", CMS Physics Analysis Summary CMS-PAS-HIG-13-004, (2013).

[12] CMS Collaboration, "Evidence for a particle decaying to W+W- in the fully leptonic final state in a standard model Higgs boson search in pp collisions at the LHC", CMS Physics Analysis Summary CMS-PAS-HIG-13-003, (2013).

[13] CMS Collaboration, "Combination of standard model Higgs boson searches and measurements of the properties of the new boson with a mass near 125 GeV", CMS Physics Analysis Summary CMS-PAS-HIG-12-045, (2012).

[14] CMS Collaboration, "Evidence for a new state decaying into two photons in the search for the standard model Higgs boson in pp collisions", CMS Physics Analysis Summary CMS-PAS-HIG-12-015, (2012).

[15] CMS Collaboration, "Search for a standard model Higgs bosons decaying to tau pairs in pp collisions", CMS Physics Analysis Summary CMS-PAS-HIG-12-018, (2012).

[16] CMS Collaboration, "Search for a Higgs boson decaying into a Z and a photon in pp collisions at $\sqrt{s}$ = 7 and 8 TeV", technical report, (2012).

[17] CMS Collaboration, "Properties of the Higgs-like boson in the decay H to ZZ to 4l in pp collisions at sqrt s =7 and 8 TeV", CMS Physics Analysis Summary CMS-PAS-HIG-13-002, (2013).

[18] CMS Collaboration, "Updated measurements of the Higgs boson at 125 GeV in the two photon decay channel", CMS Physics Analysis Summary CMS-PAS-HIG-13-001, (2013).

[19] CMS Collaboration, "Measurement of the hadronic activity in events with a Z and two jets and extraction of the cross section for the electroweak production of a Z with two jets in pp collisions at $\sqrt{s}$ = 7 TeV", technical report, (2013).

[20] CMS Collaboration, "Search for a heavy Higgs boson in the H to ZZ to 2l2nu channel in pp collisions at sqrt(s)= 7 and 8 TeV", CMS Physics Analysis Summary CMS-PAS-HIG-13-014, (2013).

[21] CMS Collaboration, "Higgs to bb in the VBF channel", CMS Physics Analysis Summary CMS-PAS-HIG-13-011, (2013).

[22] CMS Collaboration, "Performance of Missing Transverse Momentum Reconstruction Algorithms in Proton-Proton Collisions at p sqrt s = 8 TeV with the CMS Detector", CMS Physics Analysis Summary CMS-PAS-JME-12-002, (2012).

[23] CMS Collaboration, "Higgs Physics Studies for the HCAL Upgrade TDR", CMS Physics Analysis Summary CMS-PAS-HIG-12-030, (2013).

[24] CMS Collaboration, "Observation of a new boson at a mass of 125 GeV with the CMS experiment at the LHC", paper Phys. Lett. B 716 (2012) 30, (2013).

[25] CMS Collaboration, "EWK Z+2j analysis", *CMS Physics Analysis Summary* **PAS-FSQ-12-019** (2013).

[26] G. P. S. Matteo Cacciari, "Pileup subtraction using jet areas", *Accepted by Phys. Lett. B* (2012) arXiv:0707.1378.

[27] CMS Collaboration, "The CMS experiment at the CERN LHC", technical report, (2008).

[28] CMS Collaboration, "Particle–Flow Event Reconstruction in CMS and Performance for Jets, Taus, and $E_{\mathrm{T}}^{\mathrm{miss}}$", CMS Physics Analysis Summary CMS-PAS-PFT-09-001, (2009).

[29] CMS Collaboration, "Commissioning of the Particle-flow Event Reconstruction with the first LHC collisions recorded in the CMS detector", *CMS Physics Analysis Summary* **CMS PAS PFT-10-001** (2010).

[30] M. Cacciari, G. P. Salam, and G. Soyez, "The anti-kt jet clustering algorithm", *JHEP* **04** (2008) 063, doi:10.1088/1126-6708/2008/04/063, arXiv:0802.1189.

[31] J. Alwall et al., "MadGraph 5: Going Beyond", *JHEP* (2011) arXiv:1106.0522.

[32] T. Sjöstrand, S. Mrenna, and P. Skands, "PYTHIA 6.4 physics and manual", *JHEP* **05** (2006) 026, doi:10.1088/1126-6708/2006/05/026, arXiv:hep-ph/0603175.

[33] S. Gieseke et al., "Herwig++ 2.5 Release Note", arXiv:1102.1672.

[34] CMS Collaboration, "Updated measurements of the Higgs boson at 125 GeV in the two photon decay channel", CMS Physics Analysis Summary CMS-PAS-HIG-13-001, (2013).

[35] CMS Collaboration, "Performance of quark/gluon discrimination in 8 TeV pp data at CMS", CMS Physics Analysis Summary CMS-PAS-JME-13-002, (2013).

[36] P. Adzic et al., "Reconstruction of the signal amplitude of the CMS electromagnetic calorimeter", *Eur. Phys. J. C* **46** (2006) 23–35, doi:10.1140/epjcd/s2006-02-002-x.

[37] GEANT4 Collaboration, "GEANT4: A Simulation toolkit", *Nucl. Instrum. Meth. A* **506** (2003) 250, doi:10.1016/S0168-9002(03)01368-8.