Published for SISSA by Springer

Received: *March 9, 2016*
Revised: *May 3, 2016*
Accepted: *May 5, 2016*
Published: *May 26, 2016*

Thinking outside the ROCs: Designing Decorrelated Taggers (DDT) for jet substructure

JHEP05(2016)156

James Dolen,[a] Philip Harris,[b] Simone Marzani,[a] Salvatore Rappoccio[a] and Nhan Tran[c]

[a] *Department of Physics, University at Buffalo, The State University of New York,*
Buffalo, NY, 14260-1500 U.S.A.

[b] *CERN, European Organization for Nuclear Research,*
Geneva, Switzerland

[c] *Fermi National Accelerator Laboratory (FNAL),*
Batavia, IL, 60510 U.S.A.

E-mail: james.william.dolen@cern.ch, philip.coleman.harris@cern.ch, smarzani@buffalo.edu, srrappoc@buffalo.edu, ntran@fnal.gov

Abstract: We explore the scale-dependence and correlations of jet substructure observables to improve upon existing techniques in the identification of highly Lorentz-boosted objects. Modified observables are designed to remove correlations from existing theoretically well-understood observables, providing practical advantages for experimental measurements and searches for new phenomena. We study such observables in W jet tagging and provide recommendations for observables based on considerations beyond signal and background efficiencies.

Keywords: Jets, QCD Phenomenology

ArXiv ePrint: 1603.00027

doi:10.1007/JHEP05(2016)156

Contents

1 Introduction

Techniques that aim to exploit the substructure of jets in order to identify highly Lorentz-boosted objects [1–4] have become an essential component of the LHC phenomenology toolkit. Several grooming and tagging algorithms, e.g. [5–15], have been developed, successfully tested, and are currently used in experimental analyses. Considerable theoretical progress has also been made and theoretical calculations that describe the action of groomers and taggers on both background [16, 17] and signal jets [18, 19] have been performed. More recently, calculations have been extended to interesting case in which a jet shape is measured in conjunction with a cut on the jet mass in [20–23] and [24].

Despite this enormous amount of progress, experimental collaborations have yet to fully exploit these advantages to reduce systematic uncertainties in analyses using substructure techniques. Much study has been focused on the relationship of numerous identification observables in order to construct the most optimal heavy object taggers. Dedicated phenomenological studies [4] and detailed analysis by CMS [25–28] and ATLAS [29–32] employing multivariate techniques were performed in order to understand how to best identify boosted W/Z bosons, top quarks and Higgs bosons optimizing the statistical discrimination power of background rejection and signal efficiency. Moreover, there has been recent interest in using computer vision techniques to combine individual calorimeter cells into non-linear optimal observables [33–35]. However, a quantitative study of the reduction of systematic uncertainties by taking advantage of theoretical improvements has not yet been performed.

In the following study, we aim to build a tagger based not only on statistical discrimination power, but also the robust behavior of the inherent QCD background. This tagger will be designed such that, after applying a flat cut on the tagging variable, the shape of the QCD background jet mass distribution remains stable and flat. We demonstrate our methodology, entitled "designed decorrelated taggers (DDT)", by performing an example analysis in which hadronically decaying W boson jets are distinguished from quark- and gluon-initiated jets. The DDT approach is applicable to the identification of any heavy boosted objects, such as Z, H, and top jets.

Samples. The Monte Carlo samples used in this study were originally used for studies in the BOOST13 report [4]. Samples were generated at $\sqrt{s} = 8\,\text{TeV}$ for QCD dijets, and for W^+W^- pairs produced in the decay of a scalar resonance. The QCD events were split into subsamples of gg and $q\bar{q}$ events, allowing for tests of discrimination of hadronic W bosons, quarks, and gluons. QCD samples were produced at leading order (LO) using `MADGRAPH5` [36], while WW samples were generated using the `JHU GENERATOR` [37]. The samples were then showered through `PYTHIA8` (version 8.176) [38] using the default `tune 4C` [39]. The samples were produced in exclusive p_T bins of width 100 GeV at the parton level. The p_T bins investigated in this report were 300–400 GeV, 500–600 GeV and 1.0–1.1 TeV.

The stable particles in the generator-level events are clustered into jets with the anti-k$_T$ jet algorithm [40] with three different distance parameters, $R = 0.4, 0.8, 1.2$, using `fastjet 3.1` [41, 42]. No multiple parton interactions (or pileup) is used in these samples, although previous LHC measurements [43, 44] have shown that grooming algorithms are more resilient to pileup effects than standard jet algorithms. Furthermore, it was shown in those measurements that the Monte Carlo simulation can accurately reproduce the data for regions of high jet mass, whereas there are disagreements below the Sudakov peak. The grooming algorithms, however, mitigate this disagreement very strongly as well. As such, we study jets with a grooming algorithm applied. The algorithms we have investigated are the "modified" mass-drop tagger (mMDT) [5, 16] with $z_{\text{cut}} = 0.1$, jet trimming [10] with $R_{\text{sub}} = 0.3$ and $f_{\text{cut}} = 0.1$, jet pruning [8, 9], and soft drop [12] with $z_{\text{cut}} = 0.1$ for both $\beta = 1$ and $\beta = 2$ (note that the case of $\beta = 0$ is equivalent to the mMDT). We have found that the conclusions are not strongly dependent on the groomer used, so have used soft-drop with $\beta = 0$ (mMDT) for most of our comparisons due to its smoother scaling behavior than other groomers [16].

2 Current taggers

Current heavy object jet substructure taggers employed by CMS and ATLAS often cut on some number of observables directly or through some algorithm. Take, for example, something similar to the CMS Run 1 W tagger that uses simple cuts on the N-subjettiness ratio τ_2/τ_1 [11] and the soft drop jet mass [12]. In this study, we consider the τ_2/τ_1 variable where the subjet axes are chosen using the k_T one-pass axes optimization technique.

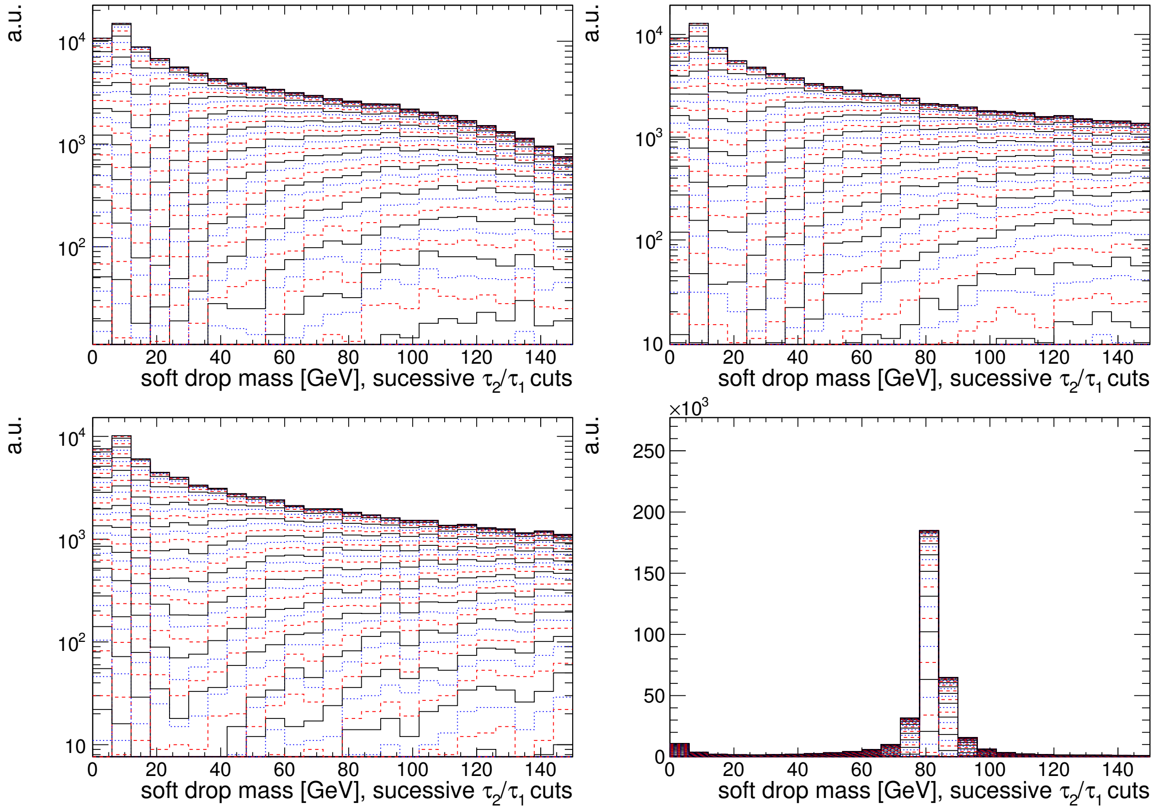

Figure 1. Soft drop mass distribution ($z_{\text{cut}} = 0.1$ and $\beta = 0$) for gluon jets after various cuts on τ_2/τ_1 ($\beta_\tau = 1$) for different jet p_T bins: $p_T = 300$–$400\,\text{GeV}$ (top left), $p_T = 500$–$600\,\text{GeV}$ (top right), $p_T = 1$–$1.1\,\text{TeV}$ (bottom left) and also for the signal (bottom right), distributions for signal are stable versus p_T. The cuts in τ_2/τ_1 vary from 1.0 to 0.0 in steps of 0.02; the changing line styles for successive cuts are meant to visually aid the reader.

In order to distinguish hadronically decaying W bosons (which give rise to jets that are intrinsically two-pronged) from QCD background, a flat cut on on τ_2/τ_1 is typically performed. As expected, this procedure greatly reduces the background, but it also leads to an unwanted sculpting of the soft drop jet mass distribution (an undesirable feature also discussed in ref. [45]), as shown in figure 1.

After cutting on τ_2/τ_1 to select jets which are two-pronged, the QCD background soft drop jet mass distribution becomes more peak-like in shape, making it harder to distinguish QCD jets from W jets which also have a peak in the jet mass distribution. The shape of the sculpted jet mass distribution, and the location of this artificial peak, varies for different jet p_T regions. This p_T dependent sculpting of the jet mass distributions makes sideband methods of background estimation more difficult. In this case and in further examples, we primarily consider gluon-initiated jets though performance with quark-initiated jets is similar. Differences will be explored in greater detail in future studies.

In ref. [16] it was argued that flat QCD mass distributions could be obtained by tuning the value of the soft drop energy fraction threshold (z_{cut}), and optimal values for quark- and gluon-initiated jets were analytically derived. However, the presence of the τ_2/τ_1 cut makes this situation more complex and it requires reconsidering the issue.

Therefore, we propose additional criterion in determining a better tagging observable beyond pure statistical discrimination power. For similarly discriminant observables, we would like to find an observable which is (1) primarily uncorrelated with the groomed jet mass observable (or rather that has complementary correlations as far as discrimination is concerned) and (2) maintains a desirable groomed mass behavior while scaling p_T. Observables satisfying this criterion would, after applying a rectangular cut, still produce a flat groomed jet mass distribution.

3 Shape observable scaling in QCD

We start our study of the correlations of substructure variables with the jet mass and p_T by introducing the appropriate scaling variable for QCD jets:

$$\rho = \log(m^2/p_T^2). \tag{3.1}$$

Here we have differed from the typical definition of jet ρ by removing the jet distance parameter R^2 from the denominator of the definition. For now we keep $R = 0.8$ fixed and leave this for future study. Note that when we apply soft-drop, we take the mass in eq. (3.1) to be computed on the constituents of the soft-drop jet, while the transverse momentum is the one of the original (ungroomed) jet.

We now compute, on both our background and signal samples, the average value of the N-subjettiness ratio τ_2/τ_1 (computed on the full jet) as a function of the soft-drop ρ. This is shown in figure 2, on the left. The signal W jets are shown in open circles while the background, here gluon jets, are shown in closed circles. The various colors are different bins in jet p_T. We note the typical behavior showing τ_2/τ_1 for the signal tending to lower values than the background and at a given value in ρ due to the mass scale of the signal jet in a given p_T bin. The signal tends to be fixed around the W mass and thus shifts for different values of p_T and is otherwise most concentrated in the dip region. Now, let us focus on the background curves (solid points). We notice a strong dependence on τ_2/τ_1 which is what causes the sculpting of the mass distributions shown in the previous section. However, we note that there exist a region in ρ for which this relationship is conspicuously linear. This is an interesting behavior, which we will exploit shortly in section 4. We also observe that, even in this linear region, there is still a residual p_T dependence, which looks like, to a very good approximation, a constant shift. The behavior observed in figure 2 for soft drop ρ is also observed for other groomers, such as trimming and pruning, within the p_T ranges consdidered. At lower values of ρ differences in the groomers become more apparent, most likely because in that region trimming and pruning acquire further sensitivity to soft physics [16]. Thus, in the current study, we concentrate on the soft-drop mass due to its stable behavior.

This approximate linear relation between τ_2/τ_1 can be (qualitatively) understood by noting that, in the case $\beta_\tau = 2$, τ_2 essentially measures the subjet mass, while τ_1 corresponds to the jet mass itself. This leads to an approximately linear relation between τ_2/τ_1 and ρ in the region of the (soft-collinear) phase-space where all-order effects can be

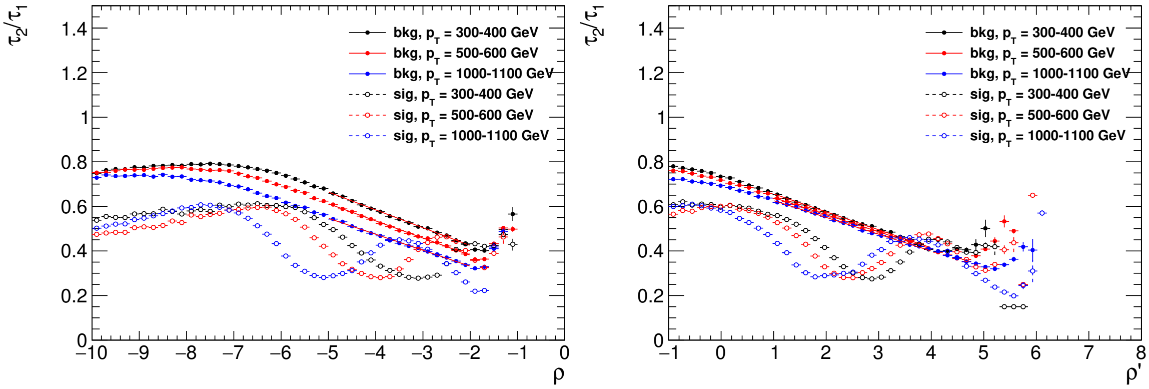

Figure 2. Profile distributions, $\langle\tau_2/\tau_1\rangle$, as a function of $\rho = \log(m^2/p_T^2)$ (left) and as a function of $\rho' = \log(m^2/p_T/\mu)$ (on the right). Solid dots correspond to background, while hollow ones to signal. The different colors correspond to different p_T bins.

neglected.[1] Furthermore, ref. [24] performed calculations for jet mass distributions in the presence of a τ_2/τ_1 cut to an accuracy which is close to next-to-leading logarithmic (NLL) accuracy. Despite the fact that the calculation corresponding to the profile plot in figure 2 were not performed, it could in principle be derived because the authors do provide the double differential distribution in τ_2/τ_1 and ρ. However, some important differences between our current set-up and the one of ref. [24] prevent us from using their results to get more quantitative insight in the behaviors we observe beyond the existence of a region with linear correlation. First ref. [24] did not consider the soft drop ρ and, second, the definition of N-subjettiness differs in the two studies both in regards of the angular exponent ($\beta_\tau = 1$ versus $\beta_\tau = 2$) and of the choice of axes. We note that, at fixed-coupling, all the transverse momentum dependence is accounted for in the definition of the shape and ρ. We have checked whether the origin of the p_T dependence that we see in figure 2 (on the left) could be traced back to the transverse momentum used in the definition of the ρ (ungroomed vs groomed) but this was found not to be the case. Running coupling contributions, as well as other subleading corrections, do introduce a p_T dependence and they are likely to responsible for the observed p_T dependence. However, a quantitative understanding of these effects would require a calculation using the techniques of ref. [24]. This goes beyond the scope of this work and for this study we limit ourselves to a phenomenological solution, while leaving a first-principle analysis for future work. Thus, in order to remove the constant p_T dependence in the τ_2/τ_1 profile, we introduce a modified version of ρ:

$$\rho' = \rho + \log\frac{p_T}{\mu} = \log\left(\frac{m^2}{p_T\mu}\right). \tag{3.2}$$

This change of variable, together with the choice $\mu \sim 1\,\text{GeV}$, appears to perform an excellent job in getting rid of the p_T dependence, as shown in figure 2, on the right, though of course we note this is purely an empirical observation.

So far, we have only considered τ_2/τ_1 versus soft drop mass. We also noted that a similar linear correlation exists between τ_2/τ_1 and other groomed masses, though not

[1] We thank Andrew Larkoski for raising this point.

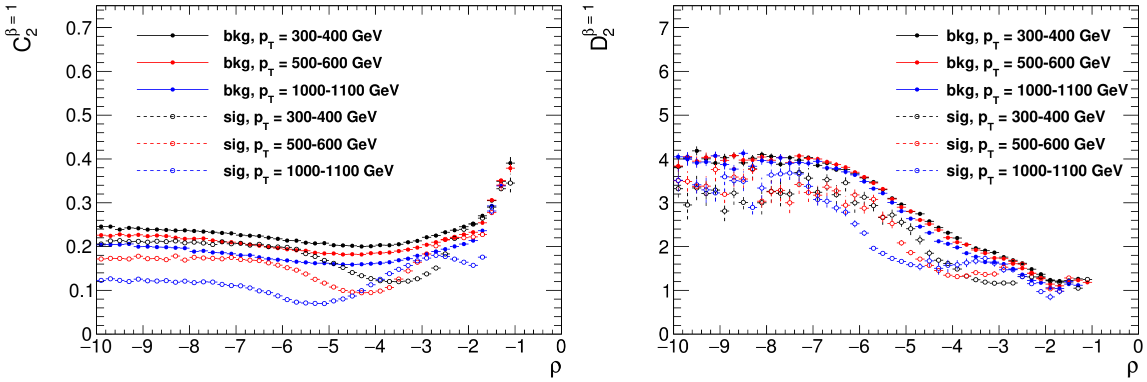

Figure 3. Profile distributions, $\langle C_2^{\beta=1}\rangle$ (left) and $\langle D_2^{\beta=1}\rangle$ (right), as a function of $\rho = \log(m^2/p_T^2)$. Solid dots correspond to background, while hollow ones to signal. The different colors correspond to different p_T bins.

shown explicitly. We can also consider other shape variables, though we leave an exhaustive exploration of all shape variables to a later study. As an example, we show also energy correlation functions $C_2^{\beta=1}$ and $D_2^{\beta=1}$ as a function of ρ in figure 3. On the left, $C_2^{\beta=1}$ shows a relatively flat distribution versus ρ which is desirable although the behavior is not quite linear. On the right, $D_2^{\beta=1}$ is highly correlated with ρ. In both cases, the correlations have some p_T-dependence that is not trivially empirically determined.

4 Designing decorrelated taggers (DDT)

4.1 Transforming τ_2/τ_1

By performing the transformation $\rho \to \rho'$, we have successfully accounted for most of the p_T dependence of the profile distribution. Next we would like to perform a further transformation with the aim of flattening the profile dependence on ρ', with the idea that this will in turn reduce the mass-sculpting discussed earlier.

In order to determine the transformation we are after, we concentrate on the region in which the relationship between τ_2/τ_1 and ρ' is essentially linear. Thus, we introduce

$$\tau'_{21} = \tau_2/\tau_1 - M \times \rho', \tag{4.1}$$

where the slope M is numerically fitted from figure 2 (red fit lines). The comparison between the τ_2/τ_1 and τ'_{21} distributions is shown in figure 4, for different jet p_T bins. The transformed variable, τ'_{21}, looks similar to the original variable τ_2/τ_1 although the behavior of the correlation with the groomed mass is now practically removed. We note that a p_T-dependence on the signal shape is introduced which is, in hindsight, expected given the transformation takes advantage of scaling properties of the background. This can cause a p_T-dependence in the signal efficiency with a cut on τ'_{21} not present in the original τ_2/τ_1; however, we note this is not necessarily an undesirable feature. For example, as backgrounds decrease at higher p_T it may be desirable to allow a larger signal efficiency and this should be studied in more detail in the experiments within the context of particular

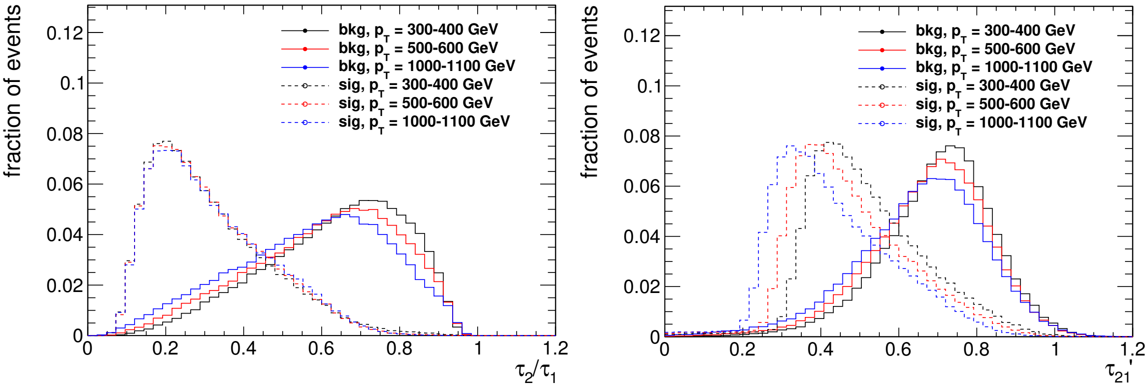

Figure 4. Raw τ_2/τ_1 distributions on the left and transformed distribution, τ_{21}', on the right.

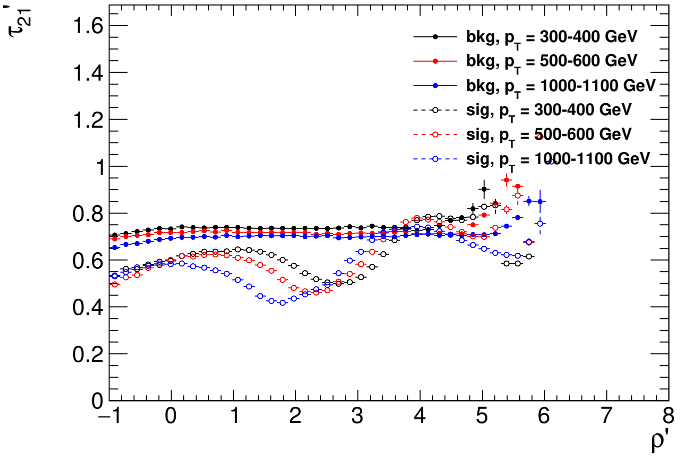

Figure 5. Profile distributions, $\langle\tau_{21}'\rangle$, as a function of $\rho' = \log(m^2/p_T/\mu)$. Solid dots correspond to background, while hollow ones to signal. The different colors correspond to different p_T bins.

analyses. This can be seen in figure 5 which shows the profile of τ_{21}' as a function of ρ' with the intended decorrelated behavior.

Now, we can explore the sculpting of the mass distributions making a flat cut in τ_{21}'. This is shown in figure 6 which should be contrasted with figure 1 which was obtained with a flat cut in τ_2/τ_1. Notice that now the sculpting of the mass distribution is considerably reduced, particularly in the region of interest where the W boson peak is. With a simple transformation, we can now preserve mass sidebands for background estimations and make robust predictions of the p_T dependence of the backgrounds. This practical consequences of a well-behaved background shape will be explored in section 5. Generally speaking, a non-linear dependence is not a technical obstacle to performing an observable transformation and we discuss this in section 6; however, studying the behavior in a simple analytic regime allows us to better understand the underlying physical behavior. The final component to evaluating the success of the observable transformation is to understand the performance of the new observable in terms of rejecting backgrounds.

4.2 Performance of DDT

To evaluate the performance of the transformed variable we use the traditional receiver operating characteristic (ROC) curve, defined as the signal efficiency as a function of the

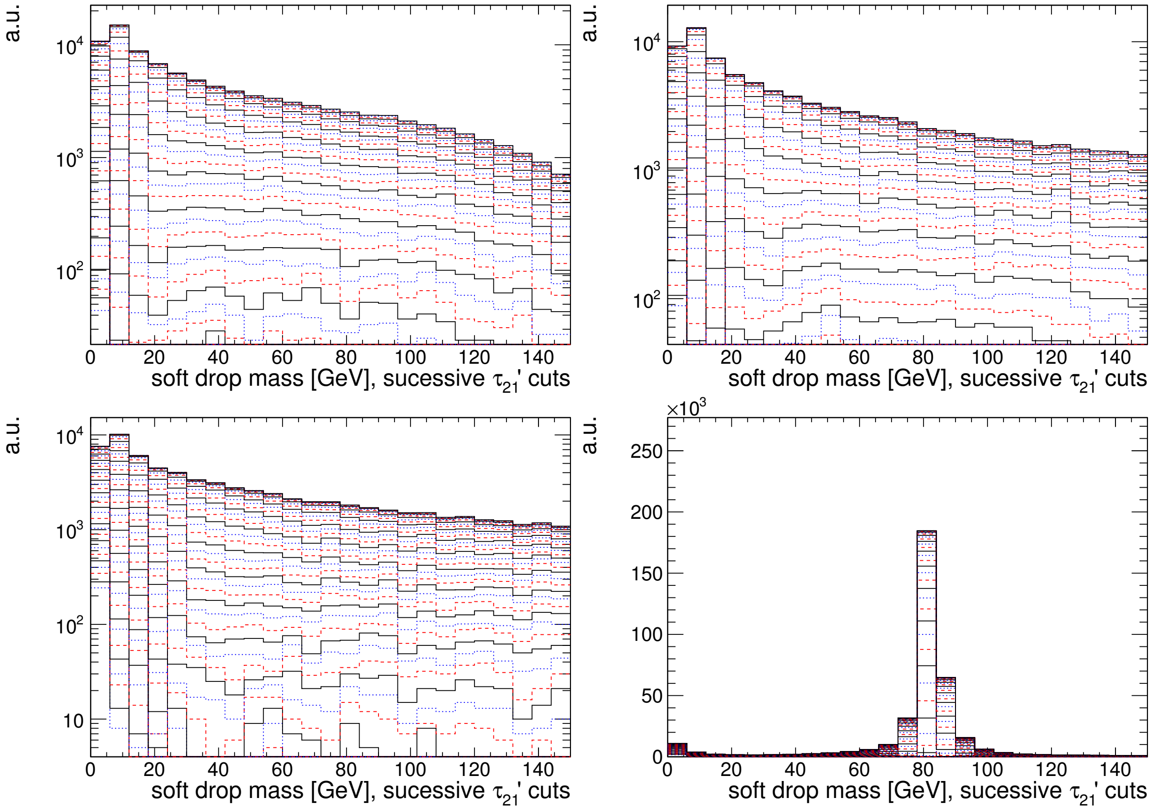

Figure 6. Soft drop mass distribution for gluon jets after various cuts on τ'_{21} for different jet p_T bins: $p_T = 300$–400 GeV (top left), $p_T = 500$–600 GeV (top right), $p_T = 1$–1.1 TeV (bottom left) and also for the signal (bottom right), distributions for signal are stable versus p_T. The cuts in τ'_{21} vary from 1.0 to 0.0 in steps of 0.02; the changing line styles for successive cuts are meant to visually aid the reader.

background efficiency. A better discriminating tagger is characterized by higher signal efficiency and lower background efficiency. The discriminating performance of τ_2/τ_1 and the transformed τ'_{21} are shown in the left of figure 7 for jets within a soft drop mass window of [60–120] GeV (corresponding to the W signal mass region). From the ROC curve, we note that after transforming the variable the discriminating power does not degrade and even shows modest improvement in this kinematic regime. We can see where this comes from in the right panel of figure 7. After cutting on raw τ_2/τ_1 the QCD soft drop jet mass distribution is sculpted such that many of the jets surviving the cut fall into the W mass region. In contrast, cutting on τ'_{21} leaves a more linearly falling distribution which preserves the low sideband. The mass distributions on the right side of figure 7 are after making a cut on the shape observable to maintain a signal efficiency of 50%.

5 Case studies

Currently, the systematic uncertainties in extracting the efficiency are large (and usually dominant) sources of uncertainty in SM and BSM analyses at the LHC [46–52]. There are

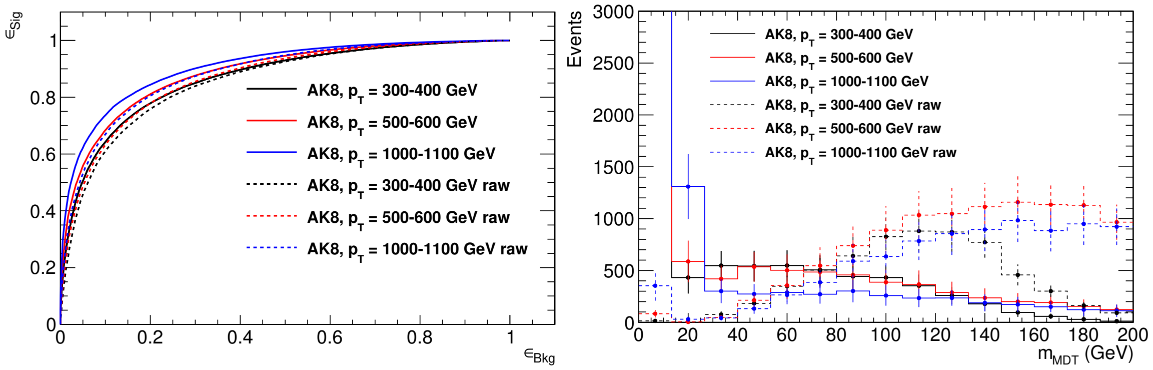

Figure 7. (Left) ROC discrimination curve: W-tagging efficiency versus QCD jet tagging efficiency for three p_T regions for the transformed τ'_{21} variable (solid) and the raw τ_2/τ_1 variable (dashed). Here efficiency is defined as the number of jets with mass satisfying $60 < \mathrm{mMDT} < 120\,\mathrm{GeV}$ which are tagged. (Right) Soft drop mass distributions after a cut on the transformed τ'_{21} variable (solid) and the raw τ_2/τ_1 variable (dashed), where the cut corresponds to 50% signal efficiency. Here the uncertainties on each bin signify the expected variation for a 10% uncertainty on the W boson tag efficiency.

several places where the improved scaling behavior can reduce these systematics, in addition to the performance improvements in the ROC curves shown in figure 7. We will present two improvements, the preservation of mass sidebands in the kinematic fit to extract the W tagging efficiency from semileptonic $t\bar{t}$ events, and the overall background estimate in diboson analyses. Both cases take advantage of the flatter background distributions to improve the uncertainties in shape-based fits.

5.1 Preservation of mass sidebands

The shape of the jet mass spectrum is used in the LHC experiments to determine the W tagging efficiency; for instance, CMS relies on a simultaneous fit to the jet mass in events that pass and fail the τ_{21} selection. However, as shown in figure 1, the τ_{21} selection significantly kinematically sculpts the background distribution in this variable. This can lead to significant fitted uncertainties when extracting the background normalization, and thus directly translates to large uncertainties in the W tagging efficiency measurement. By using the τ'_{21}, a significant improvement is observed.

To demonstrate this, we examine two cases, modified mass drop tagging with $\tau_{21} < 0.45$, and modified mass drop tagging with a scale-dependent selection $\tau_{21} < 0.6\text{–}0.08 \times \rho'$, where $\rho' = \log\left(m^2/p_T/\mu\right)$. This translates into a cut on $\tau'_{21} < 0.6$. These selections have approximately the same signal efficiency. For simplicity, the same signal and background MC samples are used as in the previous sections, but the events are weighted with an easily specifiable fraction of background jets. In this case, the background fraction for the entire sample is 40%. This gives a comparable fraction of merged to unmerged W bosons in a semileptonic $t\bar{t}$ selection at 13 TeV at the LHC, but allows us to easily tune the fraction. In addition, to mimic the approximate detector resolution, the intrinsic resolution of the $W \to qq$ system is smeared with a Gaussian of width 10 GeV. This is indicative of the resolutions obtained at the CMS and ATLAS experiments.

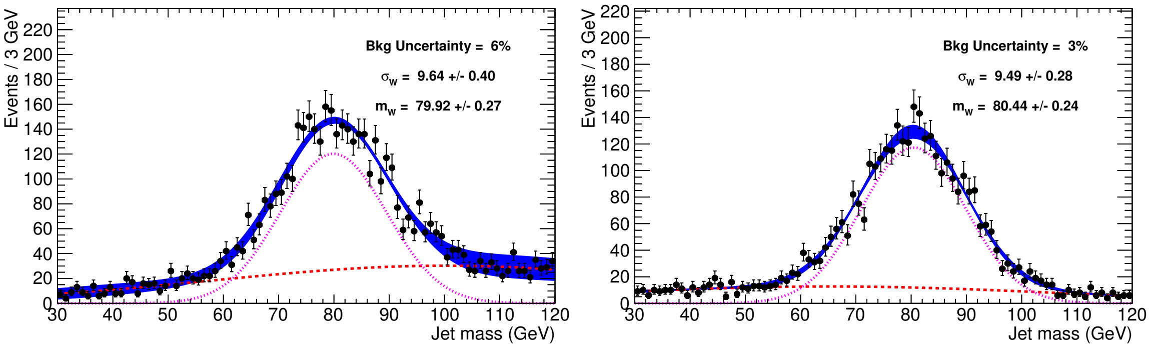

Figure 8. Jet mass for jets with $p_T = 300$–$400\,\mathrm{GeV}$ for the modified mass-drop tagger after requiring $\tau_{21} < 0.45$ (left) and $\tau'_{21} < 0.6$ (right), respectively. These two selections have approximately the same signal efficiency. The background fraction of the entire sample (for all jet masses) is set to 40%. The points are the observed MC events, after smearing the jet mass resolution to $\sim 10\%$. The purple dotted line corresponds to the smeared W signal jets. The red dashed line corresponds to the fitted background component, modeled as a Gaussian distribution. The blue band corresponds to a fit to the signal plus background, where the thickness of the line corresponds to the uncertainty in the fitted component.

Figure 8 shows simple fits to the jet mass for 5000 MC events in the range $50 < m_J < 120\,\mathrm{GeV}$, after a selection on the N-subjettiness variable. The model is a double Gaussian, one for the QCD continuum and one for the W mass peak. The jet p_T range considered is $p_T = 300$–$400\,\mathrm{GeV}$, to give a typical p_T range of the W bosons from top quark decays from SM $t\bar{t}$ production. The first fit shows the modified mass drop algorithm after $\tau_{21} < 0.45$. The second fit shows the modified mass drop algorithm after $\tau'_{21} < 0.6$. The fits successfully capture the mass of the W and the input width of 10%.

It is interesting to note that the jet mass of the QCD jets after the τ'_{21} selection are significantly pushed below 10 GeV. In addition, the remaining distribution is flat. However, for the standard τ_{21} selection, the distribution is rising, with significantly more background under the W signal peak.

The background uncertainty on the fit is is 6% when using the standard τ_{21} selection. However, it is reduced by a factor of two to 3% by using the τ'_{21} selection. This is driven by the fact that the fitter can more easily handle sidebands that are flatter, so the τ'_{21} variable outperforms the τ_{21} variable in this metric.

This would translate directly into a decreased systematic uncertainty for the LHC experiments. While newer and more clever algorithms can achieve better performance in MC simulations, this does not always translate directly to improvements in actual analyses due to the need to characterize the systematic uncertainties. We therefore propose this test as an appropriate metric to characterize the systematic performance of new substructure algorithms.

5.2 Diboson background estimate

The diboson background estimate for the LHC experiments is much the same as the extraction of the W tagging efficiency, except that the background fraction is significantly

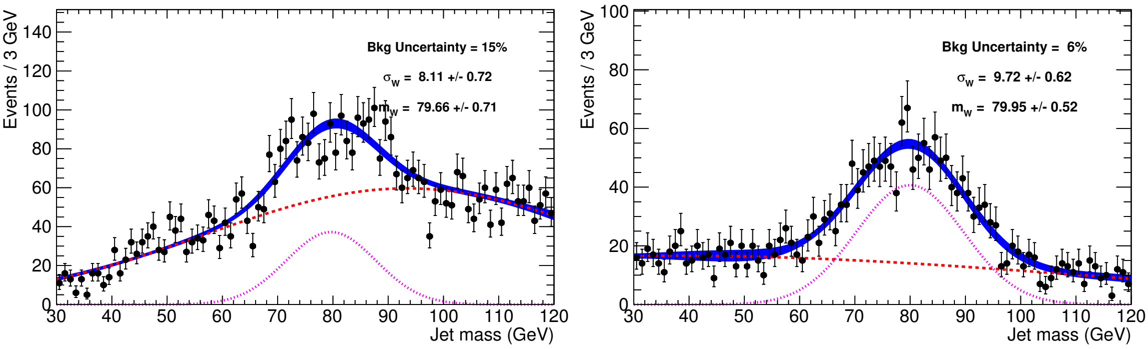

Figure 9. Jet mass for jets with $p_T = 500$–$600\,\text{GeV}$ for the modified mass-drop tagger after requiring $\tau_{21} < 0.45$ and $\tau'_{21} < 0.6$, respectively. These two selections have approximately the same signal efficiency. The background fraction of the entire sample (for all jet masses) is set to 80%. The points are the observed MC events, after smearing the jet mass resolution to $\sim 10\%$. The purple dotted line corresponds to the smeared W signal jets. The red dashed line corresponds to the fitted background component, modeled as a Gaussian distribution. The blue band corresponds to a fit to the signal plus background, where the thickness of the line corresponds to the uncertainty in the fitted component.

higher. We have chosen a value of 80% (integrated over the entire spectrum of events) as an indicative fraction, with the same number of events (5000). We have considered two different p_T ranges, $p_T = 500$–$600\,\text{GeV}$ and $p_T = 1000$–$1100\,\text{GeV}$.

One somewhat obvious but important point is that as the p_T increases, the Sudakov peak from QCD-generated jets shifts further to the right. As this occurs, the fits to discriminate boosted W bosons from QCD-generated jets are less and less able to distinguish between the categories.

Figures 9 and 10 show similar fits as shown in figure 8. However, the background fraction is raised from 40% to 80% (again integrated over the entire mass spectrum), and the p_T ranges are set to $p_T = 500$–$600\,\text{GeV}$ and $p_T = 1000$–$1100\,\text{GeV}$, respectively.

For the range $p_T = 500$–$600\,\text{GeV}$, it is plain to see that there is a significant improvement of the τ'_{21} variable, where the background uncertainty decreases from 15% to 6%. This is even more apparent for the range $p_T = 1000$–$1100\,\text{GeV}$, where the uncertainty decreases from 23% to 6%.

6 Generalized scale invariance

Decorrelation schemes can be extended beyond a pair of variables to decorrelate classes of many variables. Such a procedure can be used to allow for a class of variables to be merged into a single multi-variate analysis discriminator (MVA), while preserving decorrelation against one or a set of variables that are further used in the analysis. Consider, for example, building an MVA W tagger using both τ_2/τ_1 and $C_2^{\beta=1}$. Both of these variables have correlations with p_T and mass, so the resulting classifier that combines the variables will also be correlated with mass and p_T. Decorrelating the space of variables against mass and p_T before or during the construction of the MVA can thus preserve the mass and

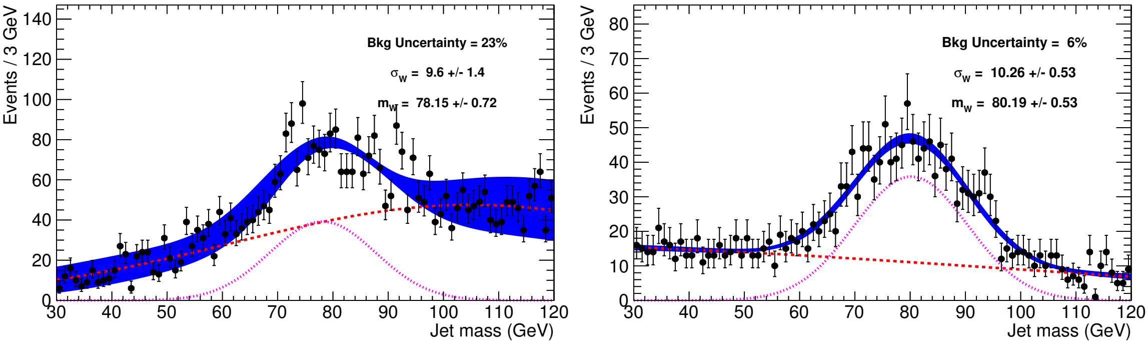

Figure 10. Jet mass for jets with $p_T = 1000$–$1100\,\mathrm{GeV}$ for the modified mass-drop tagger after requiring $\tau_{21} < 0.45$ and $\tau'_{21} < 0.6$, respectively. These two selections have approximately the same signal efficiency. The background fraction of the entire sample (for all jet masses) is set to 80%. The points are the observed MC events, after smearing the jet mass resolution to $\sim 10\%$. The purple dotted line corresponds to the smeared W signal jets. The red dashed line corresponds to the fitted background component, modeled as a Gaussian distribution. The blue band corresponds to a fit to the signal plus background, where the thickness of the line corresponds to the uncertainty in the fitted component.

p_T invariance resulting in an uncorrelated tagger. This idea has previously been pursued in b-physics utilizing an MVA that minimizes the mass dependence, while simultaneously constructing a classifier [53].

In light of building an example based on previously presented studies, we split $\rho = \log(m^2/p_T^2)$ by into it components $\log(m)$ and $\log(p_T)$. Combining this with either C_2^1 or τ_2/τ_1 gives a class of three variables for which we decorrelate into a set of three independent linear combinations of variables. The independent variables can be viewed as properties of the data which span the space of distinctive features. This space can be explored to further understand behavior of the data. Additionally, a subset of the independent components can be merged through an MVA while maintaining the decorrelation of the remaining set of variables. In this way, mass sidebands or other sideband methods can be used on the merged MVA discriminator with the decorrelated variable.

As has previously been noted, decorrelating variables which are not implicitly linearly correlated is poorly defined [54]. We thus consider two generalized approaches that attempt to decorrelate discriminators that are not necessarily linearly correlated. We consider two decorrelation approaches: Principle Component Analysis (PCA) of transformed variables and Independent Component Analysis (ICA).

Decorrelation by PCA and ICA. Given a set of variables need not be linearly correlated, we consider a transformed variable (v'_i) of the original variable v_i defined by

$$v'_i = f(v_i)\,. \tag{6.1}$$

For this transformation, we train a gradient boosted decision tree [55] with the boosted W boson as a signal and a high p_T QCD jet as a background. This transformation places the variables into a space that enables the possibility of linearized correlations of the original variables.

The resulting correlation matrix of the transformed variables can be decorrelated through principle component analysis by taking the eigenvectors of the matrix. This yields a set of n-independent vectors for a n-dimensional correlation matrix.

The decorrelated vectors for the triplet of transformed τ_2/τ_1, $\log(p_T)$, and $\log(m)$ is shown in figure 11. The correlation of the resulting vectors is compared with a gradient boosted decision tree using all variables and with the transformed mass. From this correlation, we observe two discriminating dimensions and the p_T. These we can write as

$$v_1 = \log(m/\mu_1) + K_1(\tau_2/\tau_1) \tag{6.2}$$

$$v_2 = \tau_2/\tau_1 + K_2 \log(m^{3.5}/p_T\mu_2^{2.5}), \tag{6.3}$$

where $K_{1,2}$ correspond to coefficients and $\mu_{1,2}$ are scales, typically $\mu_{1,2} \sim 1\,\text{GeV}$ to make the observables dimensionless. The first variable corresponds to the transformed mass and the second corresponds the transformed τ_2/τ_1. The second variable is not too different from ρ' decorrelated τ_2/τ_1.

An alternative decorrelation approach, known as independent component analysis (ICA), involves diagonalization of the matrix constructed by computing the pairwise mutual information of each pair of variables on the sample of QCD jets. This differs to previous approaches, which rely on the mutual information to truth. Here, we focus on identifying features in the data and not necessarily discriminating power. We perform the ICA with an algorithm that uses k-nearest neighbor to expedite the diagonalization process (MILCA) [56]. The right panel of figure 11 shows the ICA decomposed vectors. As with the transformed PCA, the ICA decorrelates the p_T, however the mass τ_2/τ_1 interdependence is stronger than in the transformed case.

Finally, the equivalent decorrelated matrix for a combined set of observables is shown in figure 12, here we show just the transformed PCA approach. From the combined set, we observe the largest orthogonal set of discrimination power comes from the $C_2^{\beta=1}$ as oppose to τ_2/τ_1. When comparing the two approaches, we have found variable transformed PCA yields a more consistent performance with our previous observations.

7 Conclusion and outlook

In this note, we explore the scale-dependence and correlations of jet substructure observables. The goal is not only to improve the statistical power of such observables, which we also demonstrate, but also to consider practical issues related to using such observables in searches for new physics. In order to design decorrelated taggers (DDT), we transform the shape observable, here $\tau_2/\tau_1 \to \tau'_{21}$, by decorrelating it from groomed mass observables also factoring in the p_T scale-dependence. In addition to improving the statistical discrimination between signal and background, we also preserve a robust, flat background shape and which has more stable behavior when scaling of the background going from lower p_T bins to higher p_T bins. We demonstrate the advantages of such an approach in various case studies such as predicting background normalizations and determining heavy object tagging scale factors related to new physics searches.

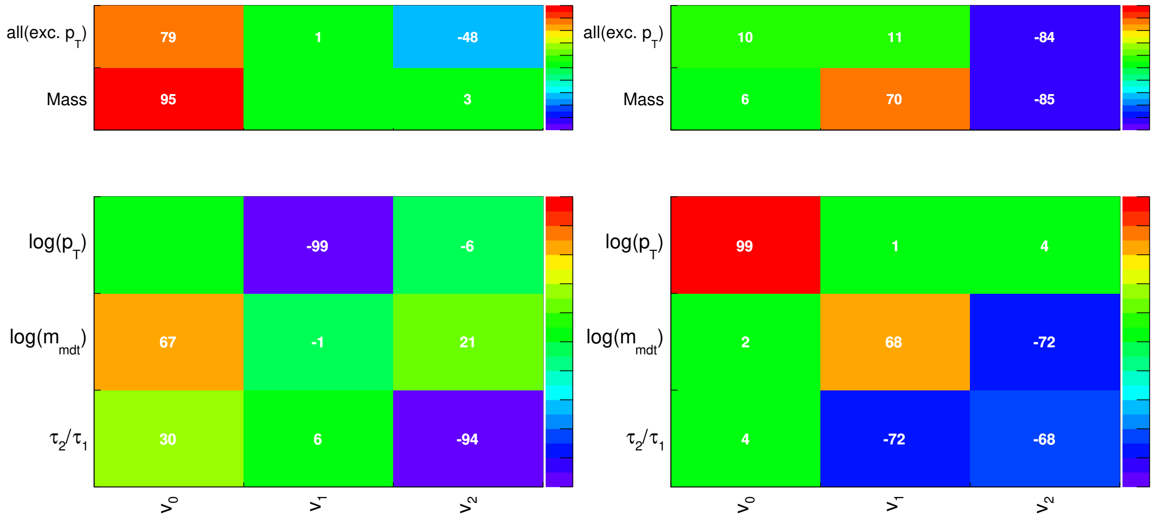

Figure 11. Decorrelated vectors from variable transformed PCA (left) and ICA (right). The bottom panel corresponds to the vectors in columns with their relative fraction labeled by row. The top panel corresponds to the correlation to the soft dropped mass and a gradient boosted decision tree trained with all variables excluding the p_T.

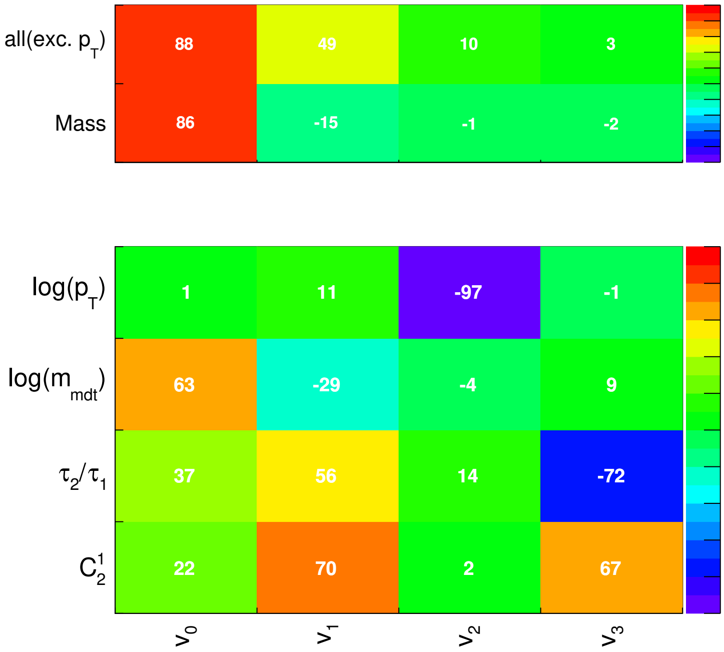

Figure 12. Decorrelated vectors from variable transformed PCA Using the full suite of observables studied in this paper. The bottom panel corresponds to the vectors in columns with their relative fraction labeled by row. The top panel corresponds to the correlation to the soft dropped mass and a gradient boosted decision tree trained with all variables excluding the p_T.

The intention of this note is not to perform a detailed study of all possible heavy object taggers, but instead, to introduce further considerations when designing taggers and propose a method by which all considerations can be addressed, namely via observable decorrelation. We leave studies related to variations on jet mass groomers and shape observables, R-scaling, quark-gluon fractions, scaling background predictions, behavior at extremely high p_T, and top tagging to future works. We have explored more generic determinations of observable decorrelation with complex taggers using multivariate techniques and numerical principle-component analysis.

Acknowledgments

We thank Andrew Larkoski and Petar Maksimovic for their critical reading of the manuscript. The authors would also like to thank Matteo Cremonesi, Matthew Low, Cristina Mantilla Suarez, Siddharth Narayanan, Gavin Salam, Gregory Soyez, and Michael Spannowsky for useful discussion and inputs. JD and SR are partially supported by the U.S. National Science Foundation, under grant PHY–1401223. PH is supported by CERN. The work of SM is partly supported by the U.S. National Science Foundation, under grant PHY–0969510, the LHC Theory Initiative. NT is supported by the Fermi Research Alliance, LLC under Contract No. De-AC02-07CH11359 with the United States Department of Energy.

References

[1] A. Abdesselam et al., *Boosted objects: A Probe of beyond the Standard Model physics*, *Eur. Phys. J.* **C 71** (2011) 1661 [arXiv:1012.5412] [INSPIRE].

[2] A. Altheimer et al., *Jet Substructure at the Tevatron and LHC: New results, new tools, new benchmarks*, *J. Phys.* **G 39** (2012) 063001 [arXiv:1201.0008] [INSPIRE].

[3] A. Altheimer et al., *Boosted objects and jet substructure at the LHC. Report of BOOST2012, held at IFIC Valencia, 23rd–27th of July 2012*, *Eur. Phys. J.* **C 74** (2014) 2792 [arXiv:1311.2708] [INSPIRE].

[4] D. Adams et al., *Towards an Understanding of the Correlations in Jet Substructure*, *Eur. Phys. J.* **C 75** (2015) 409 [arXiv:1504.00679] [INSPIRE].

[5] J.M. Butterworth, A.R. Davison, M. Rubin and G.P. Salam, *Jet substructure as a new Higgs search channel at the LHC*, *Phys. Rev. Lett.* **100** (2008) 242001 [arXiv:0802.2470] [INSPIRE].

[6] D.E. Kaplan, K. Rehermann, M.D. Schwartz and B. Tweedie, *Top Tagging: A Method for Identifying Boosted Hadronically Decaying Top Quarks*, *Phys. Rev. Lett.* **101** (2008) 142001 [arXiv:0806.0848] [INSPIRE].

[7] CMS collaboration, *A cambridge-aachen (C-A) based jet algorithm for boosted top-jet tagging*, CMS-PAS-JME-09-001 (2009) [INSPIRE].

[8] S.D. Ellis, C.K. Vermilion and J.R. Walsh, *Techniques for improved heavy particle searches with jet substructure*, *Phys. Rev.* **D 80** (2009) 051501 [arXiv:0903.5081] [INSPIRE].

[9] S.D. Ellis, C.K. Vermilion and J.R. Walsh, *Recombination Algorithms and Jet Substructure: Pruning as a Tool for Heavy Particle Searches*, *Phys. Rev.* **D 81** (2010) 094023 [arXiv:0912.0033] [INSPIRE].

[10] D. Krohn, J. Thaler and L.-T. Wang, *Jet Trimming*, *JHEP* **02** (2010) 084 [arXiv:0912.1342] [INSPIRE].

[11] J. Thaler and K. Van Tilburg, *Maximizing Boosted Top Identification by Minimizing N-subjettiness*, *JHEP* **02** (2012) 093 [arXiv:1108.2701] [INSPIRE].

[12] A.J. Larkoski, S. Marzani, G. Soyez and J. Thaler, *Soft Drop*, *JHEP* **05** (2014) 146 [arXiv:1402.2657] [INSPIRE].

[13] D.E. Soper and M. Spannowsky, *Finding physics signals with shower deconstruction*, *Phys. Rev.* **D 84** (2011) 074002 [arXiv:1102.3480] [INSPIRE].

[14] D.E. Soper and M. Spannowsky, *Finding top quarks with shower deconstruction*, *Phys. Rev.* **D 87** (2013) 054012 [arXiv:1211.3140] [INSPIRE].

[15] D.E. Soper and M. Spannowsky, *Finding physics signals with event deconstruction*, *Phys. Rev.* **D 89** (2014) 094005 [arXiv:1402.1189] [INSPIRE].

[16] M. Dasgupta, A. Fregoso, S. Marzani and G.P. Salam, *Towards an understanding of jet substructure*, *JHEP* **09** (2013) 029 [arXiv:1307.0007] [INSPIRE].

[17] M. Dasgupta, A. Fregoso, S. Marzani and A. Powling, *Jet substructure with analytical methods*, *Eur. Phys. J.* **C 73** (2013) 2623 [arXiv:1307.0013] [INSPIRE].

[18] I. Feige, M.D. Schwartz, I.W. Stewart and J. Thaler, *Precision Jet Substructure from Boosted Event Shapes*, *Phys. Rev. Lett.* **109** (2012) 092001 [arXiv:1204.3898] [INSPIRE].

[19] M. Dasgupta, A. Powling and A. Siodmok, *On jet substructure methods for signal jets*, *JHEP* **08** (2015) 079 [arXiv:1503.01088] [INSPIRE].

[20] A.J. Larkoski, I. Moult and D. Neill, *Toward Multi-Differential Cross Sections: Measuring Two Angularities on a Single Jet*, *JHEP* **09** (2014) 046 [arXiv:1401.4458] [INSPIRE].

[21] A.J. Larkoski, I. Moult and D. Neill, *Power Counting to Better Jet Observables*, *JHEP* **12** (2014) 009 [arXiv:1409.6298] [INSPIRE].

[22] A.J. Larkoski, I. Moult and D. Neill, *Building a Better Boosted Top Tagger*, *Phys. Rev.* **D 91** (2015) 034035 [arXiv:1411.0665] [INSPIRE].

[23] A.J. Larkoski, I. Moult and D. Neill, *Analytic Boosted Boson Discrimination*, arXiv:1507.03018 [INSPIRE].

[24] M. Dasgupta, L. Schunk and G. Soyez, *Jet shapes for boosted jet two-prong decays from first-principles*, *JHEP* **04** (2016) 166 [arXiv:1512.00516] [INSPIRE].

[25] CMS collaboration, *Identification techniques for highly boosted W bosons that decay into hadrons*, *JHEP* **12** (2014) 017 [arXiv:1410.4227] [INSPIRE].

[26] CMS collaboration, *Boosted Top Jet Tagging at CMS*, CMS-PAS-JME-13-007 (2014) [INSPIRE].

[27] CMS collaboration, *V Tagging Observables and Correlations*, CMS-PAS-JME-14-002 (2014) [INSPIRE].

[28] CMS collaboration, *Top Tagging with New Approaches*, CMS-PAS-JME-15-002 (2016) [INSPIRE].

[29] ATLAS collaboration, *Performance of jet substructure techniques for large-R jets in proton-proton collisions at $\sqrt{s} = 7$ TeV using the ATLAS detector*, JHEP **09** (2013) 076 [arXiv:1306.4945] [INSPIRE].

[30] ATLAS collaboration, *Identification of boosted, hadronically decaying W bosons and comparisons with ATLAS data taken at $\sqrt{s} = 8$ TeV*, Eur. Phys. J. **C 76** (2016) 154 [arXiv:1510.05821] [INSPIRE].

[31] ATLAS collaboration, *Boosted hadronic top identification at ATLAS for early* 13 *TeV data*, ATL-PHYS-PUB-2015-053 (2015).

[32] ATLAS collaboration, *Identification of boosted, hadronically-decaying W and Z bosons in $\sqrt{s} = 13$ TeV Monte Carlo Simulations for ATLAS*, ATL-PHYS-PUB-2015-033 (2015).

[33] J. Cogan, M. Kagan, E. Strauss and A. Schwarztman, *Jet-Images: Computer Vision Inspired Techniques for Jet Tagging*, JHEP **02** (2015) 118 [arXiv:1407.5675] [INSPIRE].

[34] L.G. Almeida, M. Backović, M. Cliche, S.J. Lee and M. Perelstein, *Playing Tag with ANN: Boosted Top Identification with Pattern Recognition*, JHEP **07** (2015) 086 [arXiv:1501.05968] [INSPIRE].

[35] L. de Oliveira, M. Kagan, L. Mackey, B. Nachman and A. Schwartzman, *Jet-Images — Deep Learning Edition*, arXiv:1511.05190 [INSPIRE].

[36] J. Alwall et al., *The automated computation of tree-level and next-to-leading order differential cross sections and their matching to parton shower simulations*, JHEP **07** (2014) 079 [arXiv:1405.0301] [INSPIRE].

[37] Y. Gao, A.V. Gritsan, Z. Guo, K. Melnikov, M. Schulze and N.V. Tran, *Spin determination of single-produced resonances at hadron colliders*, Phys. Rev. **D 81** (2010) 075022 [arXiv:1001.3396] [INSPIRE].

[38] T. Sjöstrand, S. Mrenna and P.Z. Skands, *A Brief Introduction to PYTHIA 8.1*, Comput. Phys. Commun. **178** (2008) 852 [arXiv:0710.3820] [INSPIRE].

[39] A. Buckley et al., *General-purpose event generators for LHC physics*, Phys. Rept. **504** (2011) 145 [arXiv:1101.2599] [INSPIRE].

[40] M. Cacciari, G.P. Salam and G. Soyez, *The anti-k_t jet clustering algorithm*, JHEP **04** (2008) 063 [arXiv:0802.1189] [INSPIRE].

[41] M. Cacciari and G.P. Salam, *Dispelling the N^3 myth for the k_t jet-finder*, Phys. Lett. **B 641** (2006) 57 [hep-ph/0512210] [INSPIRE].

[42] M. Cacciari, G.P. Salam and G. Soyez, *FastJet User Manual*, Eur. Phys. J. **C 72** (2012) 1896 [arXiv:1111.6097] [INSPIRE].

[43] ATLAS collaboration, *Jet mass and substructure of inclusive jets in $\sqrt{s} = 7$ TeV pp collisions with the ATLAS experiment*, JHEP **05** (2012) 128 [arXiv:1203.4606] [INSPIRE].

[44] CMS collaboration, *Studies of jet mass in dijet and W/Z+jet events*, JHEP **05** (2013) 090 [arXiv:1303.4811] [INSPIRE].

[45] G. Kasieczka, T. Plehn, T. Schell, T. Strebler and G.P. Salam, *Resonance Searches with an Updated Top Tagger*, JHEP **06** (2015) 203 [arXiv:1503.05921] [INSPIRE].

[46] CMS collaboration, *Search for a massive resonance decaying into a Higgs boson and a W or Z boson in hadronic final states in proton-proton collisions at $\sqrt{s} = 8$ TeV*, JHEP **02** (2016) 145 [arXiv:1506.01443] [INSPIRE].

[47] CMS collaboration, *Search for Narrow High-Mass Resonances in Proton-Proton Collisions at $\sqrt{s} = 8$ TeV Decaying to a Z and a Higgs Boson*, Phys. Lett. **B 748** (2015) 255 [arXiv:1502.04994] [INSPIRE].

[48] CMS collaboration, *Search for massive resonances decaying into pairs of boosted bosons in semi-leptonic final states at $\sqrt{s} = 8$ TeV*, JHEP **08** (2014) 174 [arXiv:1405.3447] [INSPIRE].

[49] CMS collaboration, *Search for massive resonances in dijet systems containing jets tagged as W or Z boson decays in pp collisions at $\sqrt{s} = 8$ TeV*, JHEP **08** (2014) 173 [arXiv:1405.1994] [INSPIRE].

[50] CMS collaboration, *Search for heavy resonances in the W/Z-tagged dijet mass spectrum in pp collisions at 7 TeV*, Phys. Lett. **B 723** (2013) 280 [arXiv:1212.1910] [INSPIRE].

[51] ATLAS collaboration, *Search for high-mass diboson resonances with boson-tagged jets in proton-proton collisions at $\sqrt{s} = 8$ TeV with the ATLAS detector*, JHEP **12** (2015) 055 [arXiv:1506.00962] [INSPIRE].

[52] ATLAS collaboration, *Search for Higgs boson pair production in the $b\bar{b}b\bar{b}$ final state from pp collisions at $\sqrt{s} = 8$ TeV with the ATLAS detector*, Eur. Phys. J. **C 75** (2015) 412 [arXiv:1506.00285] [INSPIRE].

[53] A. Rogozhnikov, A. Bukva, V. Gligorov, A. Ustyuzhanin and M. Williams, *New approaches for boosting to uniformity*, 2015 JINST **10** T03002 [arXiv:1410.4140] [INSPIRE].

[54] A.J. Larkoski, J. Thaler and W.J. Waalewijn, *Gaining (Mutual) Information about Quark/Gluon Discrimination*, JHEP **11** (2014) 129 [arXiv:1408.3122] [INSPIRE].

[55] H. Voss, A. Hoecker, J. Stelzer and F. Tegenfeldt, *TMVA, Toolkit for Multivariate Data Analysis with ROOT*, PoS(ACAT)040 [physics/0703039] [INSPIRE].

[56] A. Kraskov, H. Stögbauer and P. Grassberger, *Estimating mutual information*, Phys. Rev. **E 69** (2004) 066138 [cond-mat/0305641].