Constructing mass-decorrelated hadronic decay taggers in ATLAS

Andreas Søgaard on behalf of the ATLAS Collaboration
School of Physics and Astronomy, University of Edinburgh, James Clerk Maxwell Building, Peter Guthrie Tait Road, Edinburgh EH9 3FD, UK

E-mail: andreas.sogaard@ed.ac.uk

Abstract. A large number of physics processes as seen by the ATLAS experiment manifest as collimated, hadronic sprays of particles known as 'jets.' Jets originating from the hadronic decay of massive particles are commonly used in searches for new physics. ATLAS has employed multivariate discriminants for the challenging task of identifying the origin of a given jet. However, such classifiers exhibit strong non-linear correlations with the invariant mass of the jet, complicating analyses which make use of the mass spectrum. A comprehensive study of different mass-decorrelation techniques is performed with ATLAS simulated datasets, comparing designed decorrelated taggers (DDT), fixed-efficiency k-NN regression, convolved substructure (CSS), adversarial neural networks (ANNs), and adaptive boosting for uniform efficiency (uBoost). Performance is evaluated using suitable metrics for classification and mass-decorrelation.

ATL-SOFT-PROC-2019-002
16 May 2019

1. Introduction

Jets are used to reconstruct hadronic decays of massive particles, e.g. W and Z bosons, in several searches in the ATLAS experiment [1] at the Large Hadron Collider (LHC). In the high-p_{T}, or "boosted," regime the products of these resonance decay are collimated enough to be reconstructed as individual large-radius jets. Several searches for physics beyond the Standard Model (BSM) are performed by using jet mass spectra and rely on the identification of the hadronic decays of hypothesised massive Z' resonances [2–6]. Distinguishing these two-body resonance decays from the dominant non-resonant jet backgrounds is possible by studying the angular energy distributions inside the jet, i.e. the so-called jet substructure [7, 8]. In the case of two-body decays, the jet substructure can be quantified using analytically calculated observables such as the N-subjettiness ratio τ_{21} [9] or the energy correlation function ratio $D_2^{(\beta=1)}$ [10]. In ATLAS, reconstruction techniques combine multiple jet substructure observables using multivariate analysis (MVA) techniques such as deep neural networks (DNNs) or boosted decision trees (BDTs), thereby improving e.g. W boson jet identification [11]. However, MVA-based jet taggers learn to exploit the fact that the jet mass, peaking for resonant production, is a powerful feature for identifying hadronic resonance decays against non-resonant backgrounds. Therefore, MVA taggers using only jet substructure observables as inputs yield tagger observables which exhibit non-linear correlations with the jet mass. This means that simple selections on such MVA tagger observables tend to distort the non-resonant background jet mass distribution, sculpting it to resemble the resonance jet mass peak, and thereby complicating resonance searches in the jet mass spectrum. This contribution presents the recent work in ATLAS on constructing mass-decorrelated jet substructure taggers, detailed in Ref. [12].

2. Samples and metrics

Binary classification of W jets ('signal') vs. non-resonant QCD jets ('background') is used as the benchmark task, although results should generalise to Z' resonances with other masses. Monte Carlo (MC) samples are generated with PYTHIA8 [13] using a $W' \to WZ$ process and QCD dijet production for the signal and background jets, respectively. Jets are reconstructed from calorimeter clusters using the anti-k_t algorithm [14, 15] with radius parameter $R = 1.0$ and are groomed using the trimming algorithm [16] with $R_{\text{sub}} = 0.5$ and $f_{\text{cut}} = 5\%$. The combined jet mass m defined in Ref. [17] is used throughout. Jets with 200 GeV $< p_{\text{T}} <$ 2 TeV and 50 GeV $< m <$ 300 GeV are selected to emulate relevant analysis selections. The MC signal sample is re-weighted to the background jet p_{T} spectrum to avoid potential biases from modelling differences. A sample of 10^6 signal and 10^6 background samples are used for training the MVA-based jet taggers, and a separate sample of 10^6 signal and 10^7 background samples are retained for final performance evaluation.

To evaluate the performance of each mass-decorrelated jet taggers, suitable metrics for both classification performance and degree of mass-decorrelation are necessary. Classification performance is quantified using the background rejection rate

$$\frac{1}{\varepsilon_{\text{bkg}}^{\text{rel}}} = \frac{N_{\text{bkg}}^{\text{total}}}{N_{\text{bkg}}^{\text{pass}}} \tag{1}$$

for a threshold selection corresponding to a fixed signal selection efficiency of $\varepsilon_{\text{sig}}^{\text{rel}} = 50\%$. The degree of mass-decorrelation is quantified using the Jensen-Shannon divergence (JSD) of the jet mass distributions for background jets passing and failing, respectively, a selection with signal selection $\varepsilon_{\text{sig}}^{\text{rel}} = 50\%$

$$\text{JSD} \equiv \text{JSD}\left(\frac{1}{N_{\text{bkg}}^{\text{pass}}} \frac{\text{d}N_{\text{bkg}}^{\text{pass}}}{\text{d}m} \,\middle\|\, \frac{1}{N_{\text{bkg}}^{\text{fail}}} \frac{\text{d}N_{\text{bkg}}^{\text{fail}}}{\text{d}m} \right). \tag{2}$$

The JSD is a symmetrisation of the Kullback-Leibler (KL) divergence, with a base-2 logarithm used in this contribution, resulting in $0 \leq \text{JSD} \leq 1$. Smaller values of JSD indicate a smaller correlation between the tagger observable and the jet mass, and therefore better mass-decorrelation performance; and *vice versa*.

3. Mass-decorrelation techniques

This section briefly introduces each of the five mass-decorrelation techniques studied in this contribution. Additional details are given in Ref. [12].

3.1. Designed decorrelated taggers (DDT)

For the background process, the average value of τ_{21} is linear as a function of the kinematic variable $\rho^{\text{DDT}} = \log\left(m^2/(p_{\text{T}} \times \mu)\right)$, with $\mu = 1$ GeV, in the range $\rho^{\text{DDT}} \in [1.5, 4.0]$. The DDT method [18] proposes to correct for this dependence using a linear fit

$$\tau_{21}^{\text{DDT}} = \tau_{21} - a \times (\rho^{\text{DDT}} - 1.5) \tag{3}$$

resulting in a new mass- and p_{T}-decorrelated jet tagger τ_{21}^{DDT}. This method only corrects for the mean bias and is limited by the validity of the linear approximation, which breaks down in the low- and high-mass limits. In addition, the DDT method is only applicable to τ_{21}, which is the only jet substructure observable known to exhibit this characteristic linear dependence.

3.2. Fixed-efficiency regression (k-NN)

Fixed-efficiency k-nearest neighbours (k-NN) regression can be considered a non-parametric generalisation of DDT. In this contribution, the method is used to construct a new substructure observable based on D_2, which, for the background process, is decorrelated from the jet mass and p_{T}. The decorrelation is performed for a selection with a fixed target background efficiency $\varepsilon_{\mathrm{bkg}}^{\mathrm{rel}} = 16\%$, corresponding to a signal efficiency of $\varepsilon_{\mathrm{sig}}^{\mathrm{rel}} = 50\%$. The 16% background efficiency contour of D_2 is measured in bins of (ρ, p_{T}), with $\rho = \log\left(m^2/p_{\mathrm{T}}^2\right)$, and the k-NN-fitted profile $D_2^{(16\%)}(\rho, p_{\mathrm{T}})$ is subtracted to obtain a mass- and p_{T}-decorrelated observable

$$D_2^{k\text{-NN}} = D_2 - D_2^{(16\%)}(\rho, p_{\mathrm{T}}). \tag{4}$$

This observable is fully decorrelated, within the statistical uncertainties, for the chosen dataset and selection efficiency. The k-NN implementation in SCIKIT-LEARN (v0.19.1) [19] is used.

3.3. Convolved substructure (CSS)

In contrast to the DDT and k-NN methods, the CSS method [20] attempts to also remove the dependence of higher-order moments, e.g. the width of a particular substructure observable distribution. In this contribution, D_2 is used as the base observable, and the decorrelation is performed by morphing the D_2 distribution in bins of the jet mass to match the distribution at a reference mass m_{ref}. This "morphing" is done by convolving the D_2 distribution with a Gamma distribution, which results in the mass-decorrelated D_2^{CSS} distribution. The parameters of the Gamma distribution are optimised in each mass-bin through a χ^2-minimisation of the morphed D_2^{CSS} distribution to the D_2 distribution at m_{ref}. This method requires sufficient statistics to have smooth distributions suitable for morphing, and the discrete mass-binning may introduce artificial discontinuities. Finally, this method only decorrelates D_2 with respect to the jet mass, and not the p_{T}, in contrast to the two previous methods.

3.4. Adversarial neural networks (ANN)

Adversarial training has been proposed for making neural network (NN) classifiers independent of certain variables, e.g. the jet mass [21, 22]. A standard classifier NN can be constructed and trained according to a binary cross-entropy loss L_{clf} to predict the jet label $y \in \{0, 1\}$ based on a number of jet substructure observables $x = \{\tau_{21}, C_2, D_2, a_3, A, \mathcal{P}, R_2^{\mathrm{FW}}, KtDR, \sqrt{d_{12}}, z_{12}\}$, resulting in a tagger variable $z_{\mathrm{NN}} \in [0, 1]$. See Ref. [11] for the definition of the substructure observables. In addition, a second adversary network can be introduced and trained according to a negative log-likelihood loss L_{adv} to infer the jet mass m for background samples based on the output of the jet classifier z. If the adversary is able to perform this task beyond random guessing, some generally non-linear correlation must exist between the two. The two networks are trained simultaneously according to the effective loss

$$\min_{\theta_{\mathrm{clf}}} \max_{\theta_{\mathrm{adv}}} L_{\mathrm{clf}}(\theta_{\mathrm{clf}}) - \lambda L_{\mathrm{adv}}(\theta_{\mathrm{clf}}, \theta_{\mathrm{adv}}) \tag{5}$$

where $\theta_{\{\mathrm{clf,adv}\}}$ denote the weights of the respective networks and λ is a regularisation parameter controlling the trade-off between classification and mass-decorrelation.

The classifier and adversary NNs are connected in a single, optimised architecture as shown in Figure 1. The models are connected via a gradient reversal layer, which multiplies the gradient from the adversary by $-\lambda$ during back-propagation, leading to the desired adversarial effect in Eq. (5). Both models are constructed in KERAS (v2.1.5) [23] using the TENSORFLOW (v1.4.1) [24] backend. The project library [25] is open-source and available on GitHub.

The (A)NN training proceeds in three stages: 1. the classifier is pre-trained without the adversary, by minimising L_{clf}, resulting in the z_{NN} observable; 2. the adversary is conditioned on

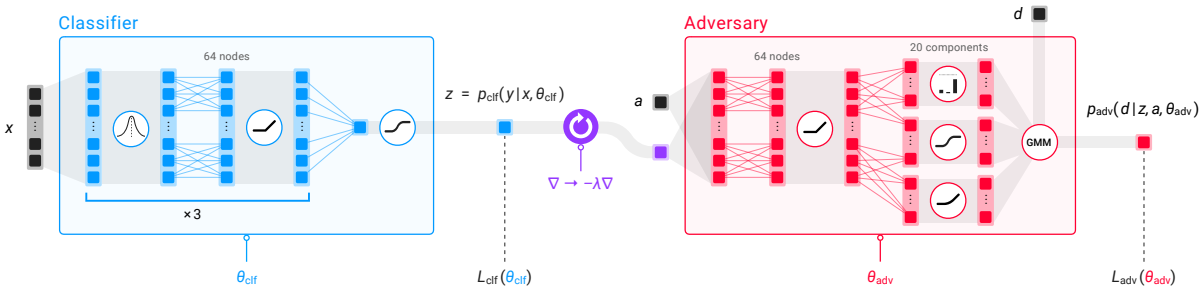

Figure 1: Adversarial neural network (ANN) architecture: The classifier network is tasked with predicting jet labels (y) based on jet substructure variable inputs (x), outputting a tagger variable (z). The adversary network is tasked with inferring the value(s) of the variables from which the classifier is to be decorrelated (d; here, the jet mass m), optionally aided by auxiliary features (a; here, $\log p_\mathrm{T}/\mu$ with $\mu = 1$ GeV), by parametrising a posterior probability density function (p.d.f.) as a Gaussian mixture model (GMM) [21]. The adversarial training is implemented using a gradient reversal layer, the trade-off between L_clf and L_adv controlled by the parameter λ. Figure from [12].

the pre-trained classifier, which is held fixed, by minimising L_adv; 3. the two networks are trained simultaneously with gradient reversal, according to Eq. (5), with a small learning rate for the classifier relative to the adversary, resulting in the z_ANN observable. During training, both signal and background samples are re-weighted to have flat p_T distributions.

3.5. Adaptive boosting for uniform efficiency (uBoost)

The second MVA-based mass-decorrelation method builds on the AdaBoost method for boosting decision trees, where at every boosting step t the relative weight w_i^t of each training example i is updated as $w_i^{t+t} = w_i^t \times c_i^t$, where the classification weight c_i^t is based on whether the i^th sample was misclassified by the decision tree at step t. This results in a jet tagger $z_\mathrm{AdaBoost} \in [0, 1]$. The uBoost method [26] introduces a uniformity weight u_i^t which is less than one if the local background selection efficiency around the i^th training example, as a function of the jet mass, is greater than a target background selection efficiency $\bar{\varepsilon} = 8\%$, corresponding roughly to a signal efficiency of $\varepsilon_\mathrm{sig}^\mathrm{rel} = 50\%$, at boosting step t; and *vice versa*. By changing the weight update to

$$w_i^{t+t} = w_i^t \times c_i^t \times u_i^t \tag{6}$$

the resulting BDT observable z_uBoost is trained to provide uniform background selection efficiency. The scale of the uniformity weight u_i^t is controlled by the so-called uniforming rate α, which therefore allows for trading off classification and mass-decorrelation. The standard AdaBoost algorithm is recovered for $\alpha = 0$. The `uBoostBDT` class from the HEP_ML (v0.5.0) [27] library is used for the implementation of both the AdaBoost and uBoost taggers. Both BDT taggers use the same substructure variables and sample re-weighting during training as the (A)NN taggers.

4. Results

The degree of mass sculpting can be assessed directly by comparing the normalised mass distribution for background jets in the inclusive sample to those passing $\varepsilon_\mathrm{sig}^\mathrm{rel} = 50\%$ selections on each tagger, cf. Figure 2. The standard MVA taggers (NN and AdaBoost) result in background distributions resembling the signal distributions, as previously found by ATLAS. The standard analytical taggers (τ_{21} and D_2) also sculpt the background distribution, but to a lesser and less

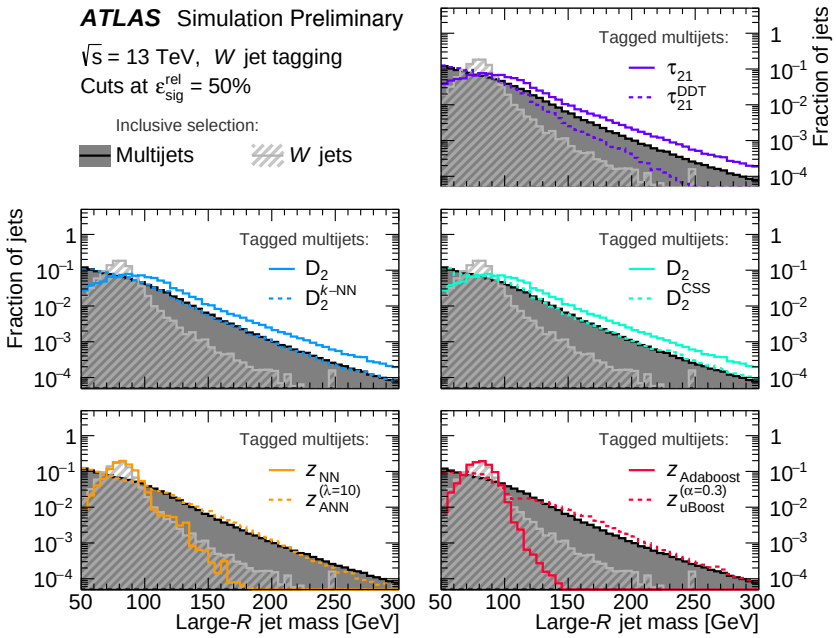

Figure 2: Normalised mass distributions for inclusive W jet (signal) and QCD multijet (background) jets, compared to the background distributions after selections the jets taggers. Selections are chosen to correspond to a signal selection efficiency of $\varepsilon_{\text{sig}}^{\text{rel}} = 50\%$. Figure from [12].

localised extent. The k-NN method results in almost perfect mass-decorrelation, while the DDT and CSS methods retain some sculpting effect. This is particularly true for DDT at high masses, which is an effect of the limitations of the linear approximation discussed in Section 3.1. Among the MVA-based methods, the ANN removes almost all sculpting effects, whereas uBoost exhibits some residual, localised sculpting at $m \approx 100$ GeV.

Figure 3 shows the classification performance, with the standard MVA taggers consistently outperforming the analytical taggers in terms of classification. Similarly, the mass-decorrelated MVA taggers perform slightly better than the standard analytical taggers. The same qualitative behaviour is also observed with the addition of a jet mass selection of $m \in [60, 100]$ GeV.

Finally, the two metrics can be studied simultaneously, cf. Figure 4. All methods presented here improve the degree of mass-decorrelation, with k-NN getting closest to the statistical limit on JSD from finite MC statistics. The MVA methods are each parametrised by a parameter (λ and α, respectively), controlling the trade-off between the classification and mass-decorrelation objectives. The classification power of both MVA-based taggers degrades with improvements in mass-decorrelation, in contrast to the analytical methods. This is because the standard MVA taggers achieve large background rejection rates exactly by exploiting mass-information, which is precisely what the mass-decorrelation methods penalise.

5. Conclusion

Five methods for constructing mass-decorrelated jet taggers are presented and evaluated according to appropriate metrics for classification and jet mass sculpting. The fixed-efficiency k-nearest neighbours (k-NN) regression method leads to the best top-level mass-decorrelation. The adversarial neural network (ANN) generally yields the largest background rejection rate for the same degree of mass sculpting. In addition, the ANN method allows for a task-specific trade-off between the two objectives.

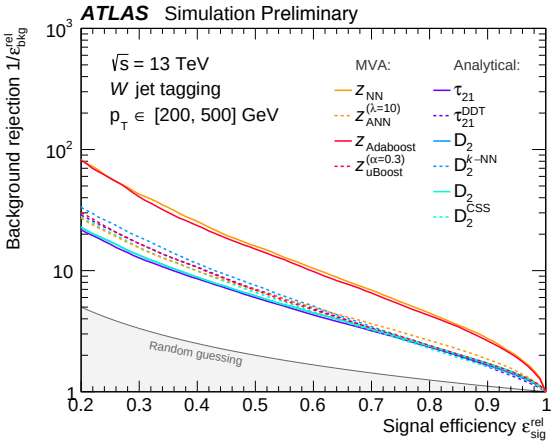

Figure 3: Rejection of QCD multijets (background) as a function of W jet (signal) selection efficiency, for standard and mass-decorrelated versions of analytical and multivariate analysis (MVA) jet taggers, without the addition of a jet mass-window selection. Figure from [12].

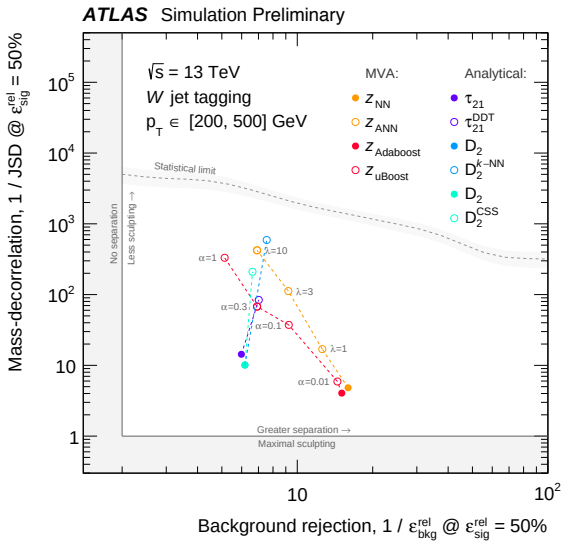

Figure 4: Background rejection, $1/\varepsilon^{\text{rel}}_{\text{bkg}}$ vs. inverse Jensen-Shannon divergence, 1/JSD, for tagger selections at $\varepsilon^{\text{rel}}_{\text{sig}} = 50\%$. Standard (mass-decorrelated) taggers shown with filled (open) markers. Shaded grey band indicates the statistical limit on 1/JSD from the finite number of simulated jets. Figure from [12].

References

[1] ATLAS Collaboration 2008 *JINST* **3** S08003
[2] ATLAS Collaboration 2018 ATLAS-CONF-2018-052 URL `https://cds.cern.ch/record/2649081`
[3] ATLAS Collaboration 2019 *Phys. Lett.* B **788** 316 (*Preprint* `1801.08769`)
[4] ATLAS Collaboration 2019 *Phys. Lett.* (*Preprint* `1901.10917`)
[5] CMS Collaboration 2018 *JHEP* **01** 097 (*Preprint* `1710.00159`)
[6] CMS Collaboration 2019 *Phys. Rev.* D **99** 012005 (*Preprint* `1810.11822`)
[7] Larkoski A J, Moult I and Nachman B 2017 (*Preprint* `1709.04464`)
[8] Asquith L *et al.* 2018 (*Preprint* `1803.06991`)
[9] Thaler J and Van Tilburg K 2011 *JHEP* **03** 015 (*Preprint* `1011.2268`)
[10] Larkoski A J, Moult I and Neill D 2014 *JHEP* **12** 009 (*Preprint* `1409.6298`)
[11] ATLAS Collaboration 2018 (*Preprint* `1808.07858`)
[12] ATLAS Collaboration 2018 ATL-PHYS-PUB-2018-014 URL `https://cds.cern.ch/record/2630973`
[13] Sjöstrand T, Mrenna S and Skands P Z 2008 *Comput. Phys. Commun.* **178** 852–867 (*Preprint* `0710.3820`)
[14] Cacciari M, Salam G P and Soyez G 2008 *JHEP* **04** 063 (*Preprint* `0802.1189`)
[15] Cacciari M, Salam G P and Soyez G 2012 *Eur. Phys. J.* **C72** 1896 (*Preprint* `1111.6097`)
[16] Krohn D, Thaler J and Wang L T 2010 *JHEP* **02** 084 (*Preprint* `0912.1342`)
[17] ATLAS Collaboration 2016 ATLAS-CONF-2016-035 URL `https://cds.cern.ch/record/2200211`
[18] Dolen J, Harris P, Marzani S, Rappoccio S and Tran N 2016 *JHEP* **05** 156 (*Preprint* `1603.00027`)
[19] Pedregosa F *et al.* 2011 *J. Mach. Learn. Res.* **12** 2825–2830 (*Preprint* `1201.0490`)
[20] Moult I, Nachman B and Neill D 2018 *JHEP* **05** 002 (*Preprint* `1710.06859`)
[21] Louppe G, Kagan M and Cranmer K 2017 *Adv. Neural. Inf. Process. Syst.* **30** 981–990 (*Preprint* `1611.01046`)
[22] Shimmin C *et al.* 2017 *Phys. Rev.* D **96** 074034 (*Preprint* `1703.03507`)
[23] Chollet F *et al.* 2018 Keras `https://github.com/fchollet/keras`
[24] Abadi M *et al.* 2016 *Proc. OSDI* 265–283 (*Preprint* `1605.08695`)
[25] Søgaard A 2018 adversarial `https://github.com/asogaard/adversarial/tree/PUBNOTE`
[26] Stevens J and Williams M 2013 *JINST* **8** P12013 (*Preprint* `1305.7248`)
[27] Rogozhnikov A hep_ml URL `https://arogozhnikov.github.io/hep_ml`