Quark vs Gluon jets in Heavy Ion Collisions

S.Draukšas[1], supervisors: M.Verweij[2], Y.Chen[2]

[1]Vilnius University, Vilnius, Lithuania
[2]CERN, Geneva, Switzerland

August 18, 2017

Contents

Abstract

The project concerned quark and gluon jets which are often used as probes of Quantum Chromodynamics(QCD) matter created in nuclear collisions at collider energies. The goal is to look for differences between quark and gluon jets, study their substructure, look for distinguishing features in unquenched (pp collisions) and quenched (heavy ion collisions) jets by using multi-variate analysis which was carried out with the help of ROOT's TMVA tool. Mapping out the modification of jets due to medium interactions could give valuable input to constraining the time evolution of the Quark Gluon Plasma created in heavy ion collisions.

1 Proton-proton collisions for building the analysis framwework

1.1 Data Sorting and Selection of Variables

The project was started by looking into jets in pp collisions and getting familiar with the differences in quark and gluon jets and also with the variables (moments) used for previous analyses [1]. The variables were sums of transverse momentum and angular distance from the jet axis, with various combinations powers, of particles belonging to the jet:

$$\sum_{i\epsilon jet} \left(\frac{p_T^i}{p_T^{jet}}\right)^{\alpha} \left(\frac{dR^i}{dR^{jet}}\right)^{\beta} \tag{1}$$

where $\alpha = 0; 1; 1.5; 2$ and $\beta = 0; 1; 2$ in common cases, p_T^{jet} and dR^{jet} are normalisation factors for each jet. dR^{jet} was set to be equal to 0.4 for each jet, which is the value used in the CMS experiment. At first, the jets acquired from pp Monte Carlo were sorted into quark and gluon jets and all of the variables were plotted out for each class of jets. To be precise, only light quarks were selected, up, down and strange because the effects of heavier quarks are negligible. Jets with transverse momentum higher than 100 GeV/c and pseudo-rapidity lower than 1 were selected (central jets). After comparison of distributions the most promising variables were selected as inputs to ROOT's Toolkit for Multi-Variate Data Analysis (TMVA). As a matter of interest a completely new variable was introduced, which was called "My moment" and had $\alpha = 2$ and a negative $\beta = -1.5$. The new variable was supposed to emphasize the differences in angular distance while also preserving the difference in transverse momentum distributions though the variable turned out to be useless as TMVA was unable to provide differentiation power from this variable and gave a flat ROC (Receiver Operation Curve) (See sec.3.2). PP Monte Carlo also showed signs of different angular hadron distributions within jets so another set of variables was introduced: multiplicity of hadrons in dR and pseudo-rapidity rings. The difference in distributions disappeared as more events were taken into account, new variables did not provide anything new and were turned off. The remaining selected variables are as follows:

- sum of p_T, where $\alpha = 1.5$ and $\beta = 0$
- p_T^D, where $\alpha = 2$ and $\beta = 0$

- p_T in a small radius (Jet shape), where $\alpha = 1$, $\beta = 0$ and $dR < 0.1$
- Mass, where $\alpha = 1$ and $\beta = 2$
- Width, where $\alpha = 1$ and $\beta = 1$
- Multiplicity, where $\alpha = 0$ and $\beta = 0$
- squared radial distance, where $\alpha = 0$ and $\beta = 2$
- MyMoment, where $\alpha = 2$ and $\beta = -1.5$

1.2 Using TMVA

In TMVA Boosted Decision Tree (BDT) methods with different variations were booked: BDT with Decorrelation and AdaBoost algorithm (BDT_DAB) and a BDT with Fisher's linear discriminant (BDT_Fisher). Quark jets were given as signal and gluon jets as background. As output TMVA provided ROCs (background rejection vs. signal efficiency), correlation plots and weight files for application on pp data, cut values of different methods, signal and background efficiency plots vs. cut values. Also, BDTs were trained with each variable separately and then with all of them combined to see the individual and combined discrimination power difference (See Sec. 4, Fig.5).

After applying TMVA weight files on pp data a lot of plots were made. Comparison of ROCs between different methods and different variables, overlay of normalised data and MC samples as well as distributions after classification, MC fitting to data was also done and the acquired coefficients were used in overlaying MC and data again to see if the difference is just in fraction of quarks and gluons or a fundamental difference in MC and data distributions. Using BDT cut value plots significance plots were made for selecting the best ROC working points (See Sec. 4).

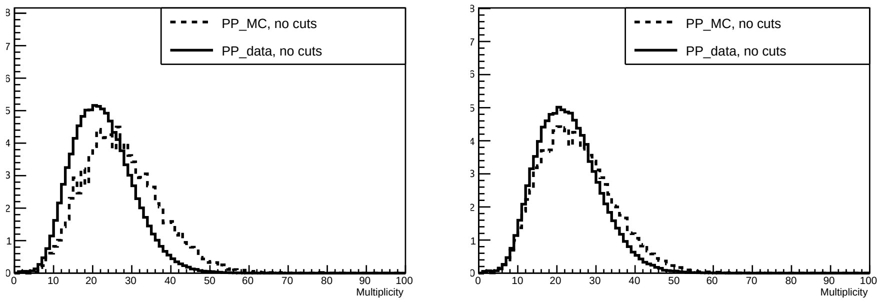
2 Heavy Ion collisions

The same approach was used for heavy ions. The only differences being that the collisions were sorted in centrality from 0%-10% (central), 10%-30%, 30%-50% to 50%-80% (pp-like).

3 Unexpected Behaviour

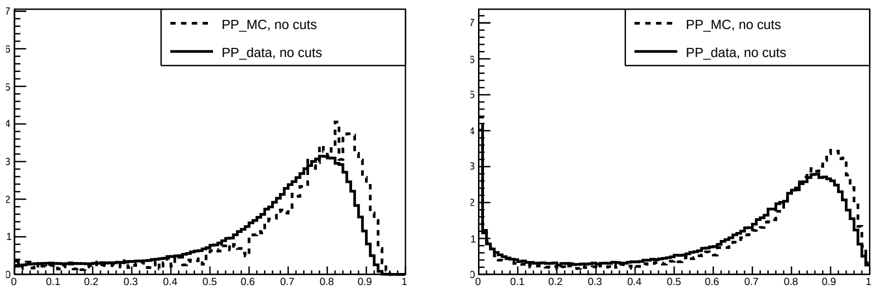
3.1 MC and Data miss-alignment

MC and data samples are not in good agreement concerning variables that include angular distance from the jet axis. For example, in squared radial distance moment there is a shift of the peak towards higher values in MC. Particle multiplicity is also exhibiting a peak shift as well as broader distribution overall. Pile up proved to have no influence on the shift. As a counter measure to miss-alignments the recalculation of jets' axes was done with included jet p_T recalculation due to the jets cone shift and change of the amount of particles. Some of the moments have improved alignment, but still did not match perfectly. In pp the distribution of squared angular distance improved a lot as well as multiplicity (See Fig. 1). Distributions of jets coming from heavy ion collisions are very dependant on centrality. Jet axis recalculation had some minor pp-like effect on the most peripheral cases, but no significant improvement was noticed in central cases. Recalculation had also changed the distribution of the Jet shape by adding a rise at low values which seems to be the correct distribution according to [1](See Fig.2). All jet selection cuts are introduced after recalculation. Further data and MC disagreements require more investigation and as of this moment it is not entirely clear what is the cause.



(a) (b)

Figure 1: In (a) particle multiplicity distribution in pp collisions before classifying (applying BDT cuts) and before jet axis recalculation, in (b) the same as in (a) but with the jet axis recalculation. Both of the MC distributions are fitted to data.



(a) (b)

Figure 2: In (a) Jet shape in pp collisions before classifying (applying BDT cuts) and before jet axis recalculation, in (b) the same as in (a) but with the jet axis recalculation. Both of the MC distributions are fitted to data. Horizontal axis is the percentage of p_T contained in the cone of $dR < 0.1$

3.2 Flat ROC of My Moment

As mentioned above the "My moment" variable did not work since TMVA was giving a completely flat ROC with 0% background rejection at any signal efficiency, which may seem to be the perfect result since it is 100% background at any chosen signal efficiency, but that is not the case since TMVA failed to compute background efficiency and it was equal to 100% at all BDT cut values (See Fig. 3). It may have happened because quarks and gluons are overlapping throughout the phase space and it is impossible to differentiate between background and signal. My moment will be included in the subsequent plots.

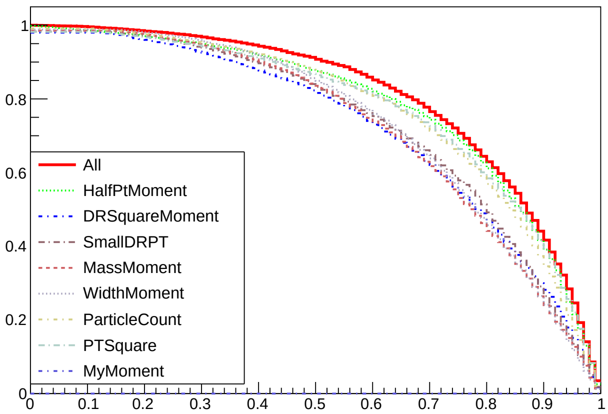


Figure 3: Plot of ROC by received from the BDT_DAB method. Vertical axis is the background rejection efficiency, vertical axis is the signal efficiency. All the used variables are seen, as well as the combined performance and a flat "My Moment" curve that matches the signal efficiency axis.

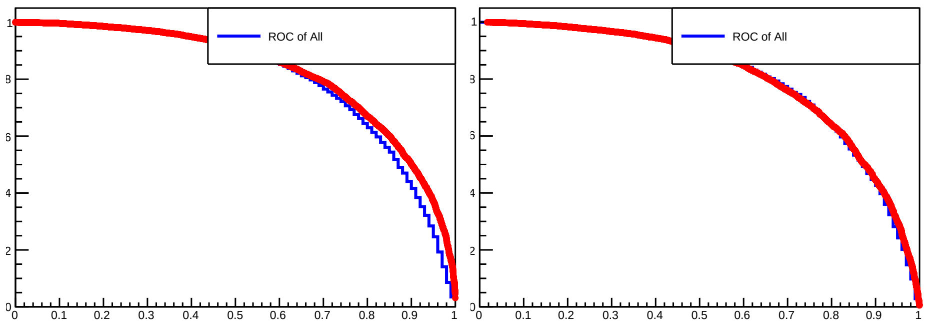
3.3 ROC inconsistencies

Another unexpected behaviour of TMVA is the inconsistency of ROCs. ROCs given by TMVA are not always the same as ROCs that are reproduced by taking signal and background efficiencies, which are also given by TMVA. The reproduced ROCs are either better (higher efficiencies and rejections) than TMVA curves or exactly the same(See Fig.4).

4 Results

4.1 Correlations of variables

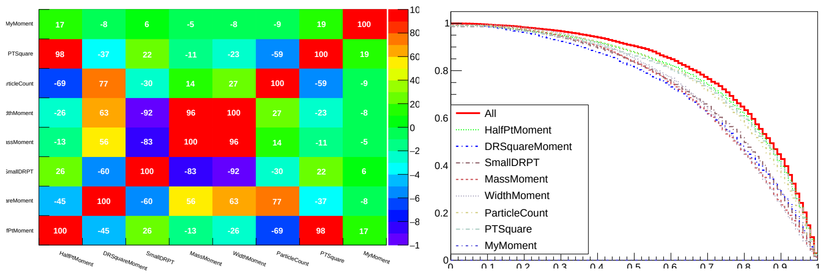
Combination of variables barely increases the overall performance of the BDT, most of the discriminating power looks to be coming from the moment with $\alpha = 1.5$ and $\beta = 0$ (HalfPtMoment as named in the legend)(See Fig.5 (b)). That very small increase of performance is due to correlations of variables. Correlated variables often share the same "truth" and do not provide additional information and therefore their combinations are not very useful. The correlation matrix of variables is provided in Figure 5 (a). The matrix shows that My Moment (top and most right rows) is very uncorrelated (green), but that is due to its uselessness as a variable overall and failing to provide any truth. The same behaviour as in "My Moment" was observed with the hadron η and dR rings and they were turned off. Variables that share p_T or dR fractions are bound to



(a) (b)

Figure 4: Plots showing the inconsistencies of TMVA ROCs, vertical axis is the background rejection efficiency and the horizontal axis is the signal efficiency. Blue line is the curve produced by TMVA, the red curve is the reproduced curve. In (a) are the curves from BDT_DAB method and in (b) from BDT_Fisher method.

be more or less correlated and if the variables are purely p_T or dR they become much less correlated.



(a) (b)

Figure 5: pp MC sample. (a) Correlation matrix of all used variables, (b) comparison of individual and combined correlation power.

4.2 General comparisons

Despite unexpected behaviour and the missalignment of data and MC BDTs are still able to pick up on variables and classify data. The same behaviour, but with Neural Networks, is seen in the article [2] where PYTHIA and HERWIG trained NNs did not have major deviations dependent on the provided sample (PYTHIA or HERWIG) (See Fig. 6). Both methods, BDT_DAB and BDT_Fisher performed very similarly or identically and provided well-agreeing ROCs (See Fig.7).

MC fitting was done after data classification using the BDT cut value histograms acquired from TMVA. In BDT cut value plots, values lower than 0 are most likely classified as gluons (background), values higher than 0 are classified as quarks(signal). The further away from zero is the value the tighter is the

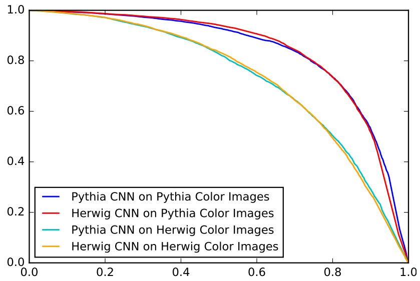


Figure 6: Plot from [2] showing stable performance of NN's regardless of the sample

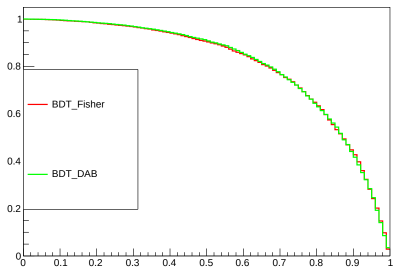
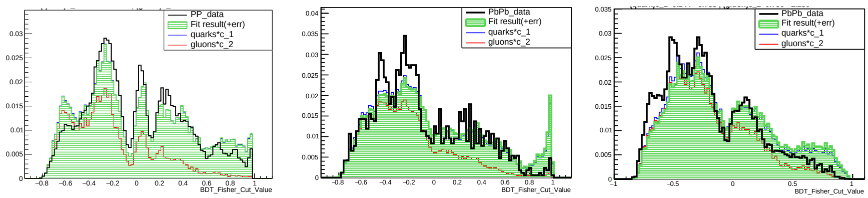


Figure 7: Plot showing almost identical performance of two different BDT's

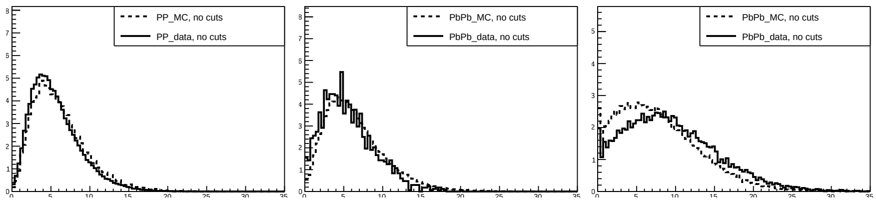
cut and more likely the classified event is signal/background. By comparing the number of events on both sides of 0 it is possible to guess whether data contains more signal or background. In very central collisions data seems to be more gluon-like than MC and it is not really possible to perfectly fit MC to data in central collisions. (See Fig.8).



(a) (b) (c)

Figure 8: Figures showing the classified data (black), fitting result(green) given by ROOT's TFractionFitter and the fit using only the coefficients from TFractionFitter (red for gluons and blue for quarks). (a) is from pp collisions, (b) the most peripheral heavy ion case (50%-80%), (c) is the most central case (0%-10%). Axis recalculation is done.

As expected, heavy ion samples (MC and data) agreed with pp more and more as collisions became more peripheral, as an example, distributions of radial distance squared are shown (See Fig.9). The peripheral case is not very different from distributions in pp, whereas the central case has a much broader distribution. Part of this effect comes from the broadening of the jet and to some extent to the broadening of multiplicity distribution in the most central case. Multiplicity is influencing all of the dR related distributions since these ratios of particle's dR to jet's dR^{jet} are going into the equations (See Eq.1) and ,unlike p_T^{jet}, dR^{jet} is fixed for every jet and is sensitive to the amount of particles.



(a) (b) (c)

Figure 9: Plots of radial distance squared from the jet axis, horizontal axis is the gained value of the moment per jet. In (a) is the pp distribution, in (b) the most peripheral heavy ion case (50%-80%), (c) is the most central case (0%-10%). MC is fitted to data and axis recalculation is done.

4.3 Working points

To choose the best ROC working points the significance function was used:

$$\frac{S}{\sqrt{S+B}} \tag{2}$$

where S is the signal(quark) efficiency and B is the background(gluon) efficiency. Background only lowers the significance and if background is not present the function behaves as the $\sqrt{S}$. According to this function the best BDT cut values were selected and best ROC working points were retrieved. Along with the ROC working points other quantities (classification efficiency, classified sample purity, etc.) together with their statistical uncertainties. All of the quantities are calculated before fitting. According to significance, the best achievable ROC working point in pp collisions is 87% signal eff. at 55% background rejection eff. at the cut of -0.062 with the BDT_DAB method and 83% signal eff. at 59% background rejection eff. at the cut of -0.235 with the BDT_Fisher method. In peripheral heavy ion collisions significance provides very similar ROC working points, which are $\approx 86\%$ signal eff. at $\approx 54\%$ background rej. eff. In central collisions significance suggests almost the same working points as in peripheral collisions

Sample purities are calculated when selecting the cut to be > 0 (mostly quarks) or < 0 (mostly gluons). In jets from pp collisions sample purities after cuts are very similar for both methods. Sample purity after quark classification is $\approx 55\%$ and after gluon classification it is $\approx 85\%$. In peripheral case purities are also very similar between methods and are $\approx 76\%$ purity after gluon classification and $\approx 70\%$ purity after quark classification. In central heavy ion collisions purity of $\approx 77\%$ is achieved after gluon classification and $\approx 67\%$ after quark classification. It is clearly seen that the sample purity is better in quark classification in heavy ion collision jets than in pp jets, though gluon sample purity is worse in heavy ion samples. Gluon classification is superior when regarding the sample purity after classification.

5 Conclusions

Non-perfect MC and data alignment requires further studies, nevertheless it is possible to discriminate between light quark (up, down, charm) and gluon initiated jets using Multi-Variate analysis, because Boosted Decision Trees are insensitive to generator systematics. The best working points are very similar between pp and heavy ion jets, but according to MC sample purities are quite different for quark and gluon classifications when taking larger intervals of BDT cuts.

In heavy ion collisions a lot depends on centrality of the collision. Most peripheral ($50\% - 80\%$) collisions are pp-like since the amount of produced Quark Gluon Plasma is low and partons do not loose a lot of energy. In most central ($0\% - 10\%$) collisions partons loose energy and as a result the jets become broader.

As differences between MC and data suggest from BDT cut value plots (See Fig.8), in central collisions amount of data classified as gluons is increasing. This increase could be caused by the increase of gluons or by quark jets that start to look more like gluon jets.

References

[1] A. J. Larkoski, J. Thaler, and W. J. Waalewijn. "Gaining (mutual) information about quark/gluon discrimination". In: *Journal of High Energy Physics* 11, 129 (Nov. 2014), p. 129. DOI: 10.1007/JHEP11(2014)129. arXiv: 1408.3122 [hep-ph].

[2] The D0 Collaboration: V. Abazov. "Combined measurements of anomalous charged trilinear gauge-boson couplings from diboson production in p-pbar collisions at sqrt(s)=1.96 TeV". In: *ArXiv e-prints* (July 2009). arXiv: 0907.4952 [hep-ex].