Classifying Heavy Ion Jets

João A. Gonçalves
joao.pedro.de.arruda.goncalves@cern.ch

Instituto Superior Técnico, Lisboa, Portugal

October 21, 2019

Abstract

This article presents the application of a probabilistic topic modelling algorithm to the separation of jets modified by the Quark Gluon Plasma (QGP), formed in heavy-ion collisions, from those that are essentially unmodified. Demix, was implemented and applied to Monte Carlo (MC) samples generated by JEWEL [1], to extract the fractions of jets modified by the QGP. We also attempt to separate quark and gluon initiated jets in the presence of the medium. Furthermore, we show that introducing a third sample improves the performance of the algorithm in the case of quark and gluon discrimination in proton-proton collisions and quenched and unquenched discrimination in lead-lead collisions.

Keywords: QCD, QGP, Jets, Probabilistic Topic Modeling, JEWEL

1. Introduction

In the heavy-ion collisions produced in particle accelerators like the Large Hadron Collider (LHC) in Europe, and the Relativistic Heavy Ion Collider in the US, a very dense and hot state of matter is produced, in which the quarks and gluons, trapped in the nucleons up to the collision, are now in a deconfined state, free to roam around this fireball. Understanding the properties of this state of matter is of fundamental importance to understand the fundamental laws that govern the sub-femtoscopic world.

Jets are a useful probe of this medium, allowing the study of the energy loss suffered by hard parton transversing. Jets are produced in the same way as in proton-proton collisions, but their fragmentation pattern, i.e. the angular and energy distribution of its constituents, will be modified. This modification is due to elastic and radiative interactions between the shower of the hard parton and the QGP. This interaction, among other things, will induce extra gluon emissions, but also an effect of anti angular ordering, meaning that the radiated partons, off the hard parton, will be radiated at higher and higher angles, as the jet evolves. This last effect competes with the vacuum angular ordering effect.

To study these effects we need to obtain a sample of jets in, say PbPb collisions. However, not all jets in this sample will have been modified by the QGP. Many will be unmodified by the QGP and hence be vacuum-like. This population may be due to their formation in the outer regions of the overlap region, or to their resolving power not being thin enough to probe the QGP's constituents, or simply because the jet escaped the region before the was formed.

These unquenched jets represent a source of systematic uncertainty in the measurement of observables that can eventually infer properties of the QGP. Learning to separate these two types of jets - quenched and unquenched - is, in this way, of the utmost importance to obtain more precise measurements of the properties of this state of matter.

To tackle this problem, we use a probabilistic topic modelling called demix, which is explained in more detail in the implementation section, 3. Results are then presented in section 4 and conclusions in 5. However, before some more background and motivation, will be given.

2. Background

The main goal is separating the jets that are strongly modified by the QGP from those that were mostly unmodified. To do this probabilistic topic modelling algorithms were applied, in particular, demix [2]. We wish to extract the modified fractions and leave a path to constrain this value for different processes, transverse momentum bins and centre of mass (CM) energies, further.

We know that in the p_T spectrum of jets in heavy-ion collisions, like the one doodled in figure 1, each bin has two contributions: 1, from jets that did not lose any transverse momentum and, 2, from jets of higher initial p_T that due to their interaction with the medium ended up in that bin. These contributions can be summarized in the following equation:

$$N_{jets}^{PbPb}(p_T \in [p_T^i, p_T^f]; N_{jets}^{pp}(p_T \geq p_T^i)) = \\ N_{jets}^{quenched}(p_T \in [p_T^i, p_T^f]; N_{jets}^{pp}(p_T > p_T^f)) + \\ N_{jets}^{unquenched}(p_T \in [p_T^i, p_T^f]; N_{jets}^{pp}(p_T \in [p_T^i, p_T^f])) \quad (1)$$

which says that the number of jets in PbPb, in a p_T bin given by $[p_T^i, p_T^f]$, depends on the number of pp jets for equivalent processes in the window $[p_T^i, \infty]$. This is a simplified model, where we assume a jet at any p_T can lose any amount of this quantity with equal probability. This dependence is translated on the two components on the right-hand side, where the number of jets with quenched and unquenched superscripts are implicitly PbPb jets. The quenched component comes only from jets that in pp would be in the interval $[p_T^f, \infty]$ - we implicitly set our criteria for a modified jet as having an energy loss more significant than $p_T^f - p_T^i$. This fact is the same as saying that the modified jets must have a larger p_T than the bin interval, before being modified, and necessarily loses transverse momentum after it is formed, which is reasonable. Furthermore, the unquenched population only depends on the number of jets in pp in that bin, since these correspond to the jets that did not lose energy.

In figure 1, we make an illustrative drawing of this effect. We draw the population p_T spectra in a *doodled* log scale, and say that they are approximately straight lines, which is reasonable solely for making the argument. As this plot is made in log scale, the bin to the right of the one represented by the dotted lines will have exponentially fewer jets, and the same happens for the one after. Hence, the contribution of quenched jets to that bin cannot be the only main population there. Many of the jets in that bin will be jets that were unmodified, maintaining its transverse momentum.

The extraction of pure quenched jet samples, from PbPb collisions, would increase the precision in measurements of the phenomena these objects can probe, quite dramatically.

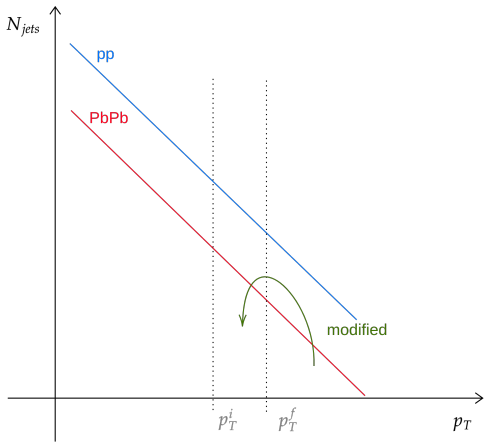

Figure 1: Doodle of the N_{jet} spectrums for pp and PbPb (log scale). Ilustrative drawing, not drawn at scale.

Probabilistic topic modelling discriminates topics at a distribution level and not at a per jet level, extracting the fractions of these topics in the input samples. With these fractions, however, one can obtain the distributions of the obtained topics for any observable, provided this observable is available from the input distributions. We aim at obtaining this fractions reliably.

2.1. Centrality in heavy-ion collisions

In heavy-ion collisions, there is an extra layer of complexity when comparing to proton collisions. Taking the lightest isotope of lead (which as 202 nucleons), per lead collision, we may have from 1 to 202 binary nucleon collisions (albeit that these limits border the unreachable). We then need to consider how many nucleon collisions occurred. So how many nucleons participated in the collision and how many were only spectators. In this way, we need to have some theoretical control of the overlap of the two nucleons, at least at a distribution level. We can control this at a distribution level through geometrical calculations.

This information can be condensed in the concept of centrality. At an experimental level, centrality is a categorical concept applied, as percentiles, to the deposited energy distribution in forward calorimeters. In the most central collisions, an enormous amount of energy is released from the interaction point, which lights up most of the detectors. On the contrary, in the most peripheral collisions, perhaps only two to four nucleons collide, and the detector has much less deposited energy. A plot produced by ATLAS taken from reference [3] is presented in figure 2 shows the centrality bins as selected from the energy deposits of the FCal (the forward calorime-

ter).

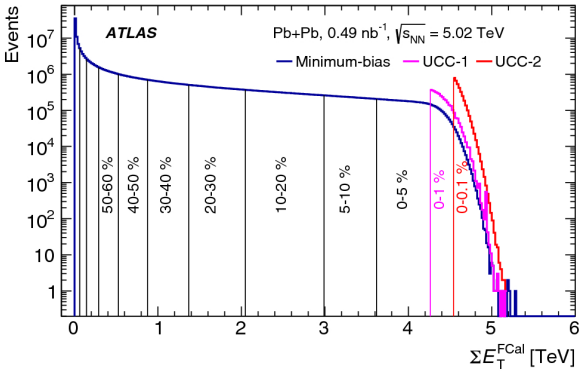

Figure 2: The sum of the transverse energy deposited in the ATLAS FCal for events selected from different experimental triggers. Plot taken from reference [3].

Centrality is an unavoidable concept in heavy-ion collisions. When most of the nucleons of the incoming ions participate in the collision, the extent of the medium produced in them is much larger than when only half of the nucleons do. This behaviour translates to the fact that jets are more likely to interact with it for its higher temperature and length. Therefore we expect to have more modified jets in the highest centrality bin, 0-10% for example, than in a median one like 40-50%.

2.2. Quark gluon plasma smoking gun

A smoking gun of the existence of the QGP, is the nuclear modification factor for charged hadrons or jets, which shows a clear modification from the pp (vacuum) environment. The jet nuclear modification factor, R_{AA}^{jet} is defined as:

$$R_{AA} = \frac{(1/N_{evt}) d^2 N_{jet}^{AA}/dp_T dy|_{cent}}{\langle T_{AA} \rangle d^2\sigma_{jet}^{pp}/dp_T dy} \qquad (2)$$

where N_{evt} is the number of events recorded at a given centrality in PbPb, $d^2 N_{jet}^{PbPb}/dp_T dy|_{cent}$ is the double differential jet yield in transverse momentum and rapidity at a given centrality, $\langle T_{AA} \rangle$ is the averaged nuclear overlap function and encodes the geometric enhancement of per-collision nucleon luminosity, $d^2\sigma_{jet}^{pp}/dp_T dY$ is the double differential cross-section of jets in transverse momentum and rapidity for a pp environment. In figure 3 we present ATLAS collaboration results taken from reference [4].

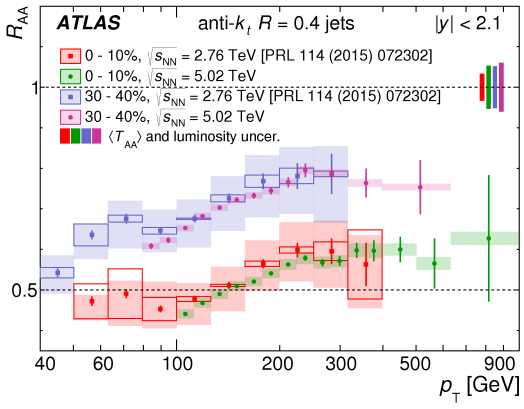

Figure 3: Jet nuclear modification factor for two centrality bins and two CM energies, as measured by ATLAS. Taken from reference [4].

In this plot, we can see the evolution of the R_{AA} with the transverse momentum of the jet, for two centrality bins and two CM energies. We can see for the most central collisions (the red and green data points) this ratio is around 50%, meaning that at this centrality half of the produced jets (as expected from the pp reference) of a given p_T were shifted to a lower p_T, and the medium absorbed many. As we go to more peripheral collisions (blue and pink data points) this absorption becomes less pronounced. This behaviour was expected: for more peripheral collisions, less nuclear matter will collide, and in turn, the extent of the QGP will be smaller and finally less jets are modified.

3. Implementation

Probabilistic topic modelling algorithms emerged in the context of natural language processing, intending to identify the underlying themes of documents, within a given collection (or corpus), [5]. The themes are extracted by considering the word occurrence distributions as a weighted sum of underlying topics (distributions), also seen in the same way.

A simple way to think about these algorithms and how they work is to consider, for example, the Instituto Superior Técnico's degree MEFT. At the Master level, students pick a branch, engineering or physics, but they still have many courses in common, with Professors from both branches. In this example, we intend to extract the distributions of words that Professors from each branch say in class, our topics. To make this extraction, we take as input the distributions of words that students from each branch ears in all classes. All students will ear all words, weather physics words, like symmetry and integral, or engineering words, like control and chips, albeit in different proportions. However,

the Professors from the physics branch say words in class that the Professors from engineering do not, and vice-versa. These are called anchor words (or bins when thinking of a histogram), words that are exclusive to each topic - high and low multiplicity for quark and gluon initiated jets for example. To extract the desired topics, the ratio of the frequencies, in which the students, from each branch, ear certain anchor words in, plays a crucial role, as it will be made clear in the following paragraphs.

These types of models can be beneficial in physics if we consider words as values of observables and documents as histograms of those observables for a given set of samples or processes. These histograms are the "word" occurrence distributions referred earlier, where words are observable values, binned according to the statistics one has. In the same way, they can be seen as a sum of underlying (topic) distributions:

$$p_i(x) = \sum_{k=1}^{K} f_k^i p_k(x) \tag{3}$$

where $p_i(x)$ is the value of our input observable distributions, i, at bin x, $p_k(x)$ is the value of the same observable for the underlying distribution, or topic, k, and f_k^i is the weight or fraction that sample i has of topic k. The sum of the fractions has to be equal to one.

3.1. Demix: the two sample case

Demix [2] is one of these models, with the advantage that the two documents, two topics case is analytical and in any appropriate modern coding language. We describe the following case in this section: given two samples, composed of different fractions of two other classes (topics), consider the distributions of a given observable for both, $p_1(x)$ and $p_2(x)$. Run through all the bins, x, to calculate the reducibility factors, defined as:

$$\kappa(1|2) = \min_x \frac{p_1(x)}{p_2(x)} \tag{4a}$$

$$\kappa(2|1) = \min_x \frac{p_2(x)}{p_1(x)} \tag{4b}$$

where the bins that satisfy equations (4) are denoted as anchor bins, a_1 and a_2, usually located in both the left and right side tails of the feature space.

From these calculate the residue distributions, which correspond to the topics' distributions:

$$p_{k=1} = \frac{p_{i=1}(x) - \kappa(1|2)p_{i=2}(x)}{1 - \kappa(1|2)} \tag{5a}$$

$$p_{k=2} = \frac{p_{i=2}(x) - \kappa(2|1)p_{i=1}(x)}{1 - \kappa(2|1)} \tag{5b}$$

For more details on the reducibility factors and residue distributions see reference [2] and references therein.

Statistical errors of the topic distributions are obtained from uncertainty propagation from the bin errors of the input samples. The bin errors of the input samples are considered to be the squared root of the number of entries within the bin.

Furthermore if we want to extract the quenched and unquenched jets' fractions - and with them, the distributions for any observable - the condition of mutual irreducibility (as defined in reference [6]) of these topics, in the chosen feature space, is a fundamental requirement. In simple terms, this condition states that each underlying topic distribution, must not be a mixture of the remaining topics plus another distribution, which corresponds to saying that the reducibility factors of the topics to one another are zero for the feature representation defined by the chosen observable(s). More simply, the topics will contain some values (words) exclusive to that topic. Moreover, sample independence and different topic fractions are required, as it is directly understood in this context. Sample independence means that the quenched and unquenched jets are of the same type for both samples, i.e. each underlying topic as the same distribution in all input samples. This sample independence can be ensured by selecting a narrow transverse momentum window for the selection. Sample independence can be further ensured if we look at quark and gluon jets independently. If these conditions are met, then the fractions are directly determined by the reducibility factors:

$$f_q^{pp} = \frac{1 - \kappa(pp|AA)}{1 - \kappa(pp|AA)\kappa(AA|pp)} \tag{6a}$$

$$f_q^{AA} = \kappa(AA|pp) f_q^{pp} \tag{6b}$$

$$f_u^{pp} = 1 - f_q^{pp} \tag{6c}$$

$$f_u^{AA} = 1 - f_q^{AA} \tag{6d}$$

where the subscript denotes the topic (q=quenched and u=unquenched) and the superscript denotes the sample to which the fraction refers to (pp = proton-proton collisions and AA = nucleus-nucleus collisions), as in equation (3). We assume these conditions are met throughout this paper.

3.2. Demix: the three (or more) sample case

In this subsection, we explain the generalization of the Demix algorithm to any number of input distributions while still requiring the reconstruction of two topics. This generalization is straightforward, corresponding to some operation to get two distributions out of the input ones and then applying the

same algorithm described in the last subsection. To test this, we use a sample of γ + Jet generated in similar conditions to the Z + Jet ones.

The procedure used to extract a lower number of distributions out of the input ones is to resample the required number of distributions as independent and uniformly distributed elements of the convex hull of the input, which is defined below (further details can be found in reference [2] and references therein):

$$conv(S) = \left\{ \sum_{i}^{|S|} \alpha_i x_i \middle| \sum_{i}^{|S|} \alpha_i = 1 \wedge \forall i : \alpha_i > 0 \right\} \tag{7}$$

where S is a set of distributions. In simple words, a convex hull of a specific set of input samples is a set containing all possible weighted sums of the inputs, where the weights sum up to one. An element of this space is then a weighted average of the inputs. For two elements (or more) to be independent and uniform, the random sampling of this space, used to obtain them, must be itself independent and uniform.

The required randomness to ensure independence of the resampled elements will introduce, by construction, randomness in the results, and not all elements under these conditions will yield the desired topics, nor acceptable error bars on them. To study this, we need to test several thousands of possibilities of random convex hull elements and present them according to how well they reconstruct the topics. This need leads us to introduce four quantities to evaluate how well the topics are reconstructed. On the one hand, we want topics that are compatible with the references we have. This compatibility can be seen as the average σ distance between each point, in reference σ or topic σ differences. In this way, we define:

$$\sigma_x = \frac{1}{N_{points}} \sum_{i=0}^{N_{points}} \frac{|y^i_{ref} - y^i_{top}|}{2e^i_x} \tag{8}$$

where N_{points} is the number of points (or bins) to be considered, y^i_{ref} and y^i_{top} are the value of point i for the reference and the topic respectively, and the e^i_x is the error of point i, with x being ref or top. For both topics, σ is calculated for both references, and the lowest value is selected - since we can only expect that a topic is compatible with only one reference. With the lowest σs of the topics one calculates the average for σ_{ref} and σ_{top}, in topics. The reference sigma is the most constraining since topic errors are arbitrarily large depending on the anchor bins. A good topic reconstruction should entail $\langle \sigma_{top} \rangle < 1$ meaning the error bars of the topic in average contain the points of the reference and in the same way, the lower the $\langle \sigma_{ref} \rangle$, the better.

On the other hand even if the topics have compatible error bars with the references, and their points themselves are close to the reference error bars, nothing tells us if the errors of the topics are not about the size of the whole distribution. With this in mind, we need two other quantities - one per topic - to characterize the amount of error that topics entails. With this in mind, we define:

$$E_j = \frac{1}{N_{points}\langle y \rangle} \sum_{i}^{N_{points}} e^i_j \tag{9}$$

where the j index refers to topic 1 and 2 and $\langle y \rangle$ is the average of the distribution in the y axis.

With these quantities, we can test several random and uniformly distributed convex hull elements of the three input distributions, but we still need a way to see them, and quickly spot trends in the "goodness" of topics. For this, we introduce the concept of ternary plots, which in the case of three input samples is directly linked to the convex hull. An example of these types of plots is given in figure 4. In these types of plots, we have three axis: A, B and C. Each axis is read at a different angle. For example, the A axis is read horizontally as in the figure, while the B axis is read at a 240-degree angle and the C axis at 120 degrees. The red lines correspond to the points where the value of A is constant, same for blue and black lines, pertaining

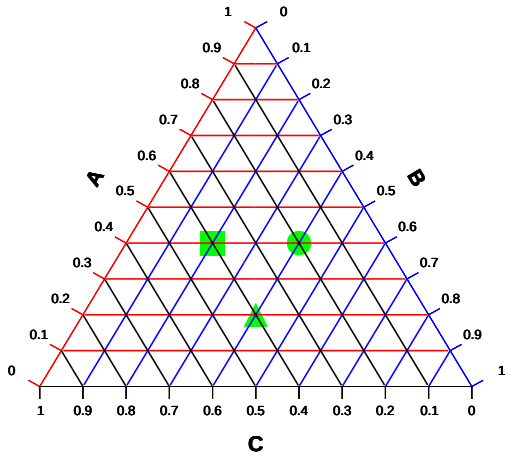

Figure 4: Example of ternary plot.

To better understand the plot, one can consider the three green points drawn therein. For example, the green circle has an A value of 0.4, the same for the B axis, and 0.2 for the C axis. The green square has $(A, B, C) = (0.4, 0.2, 0.4)$ and the green triangle

$(A, B, C) = (0.2, 0.4, 0.4)$. Note that the sum of the components is always one, meaning these values on the axis can be interpreted as the fractions, α_i for equation (7). This plot is ideal for representing points of this convex hull space. However, we need two of these points to apply the demix algorithm to extract two topics. Hence for each try of the algorithm, we will draw a line, and a different colour will be given, according to the quantities defined in (8) and (9)

4. Results

In this section, the results obtained for this article are presented. In subsection 4.1, results indicating that quark and gluon discrimination can be improved with the introduction of a third sample are shown. In subsection 4.2, results for the separation of quarks and gluons, according to JEWEL's model (from two samples), are presented. In subsection 4.3, main results for this paper, pertaining to the separation of quenched and unquenched jets, are discussed.

4.1. Quark vs gluon separation in vacuum

Quark and gluon initiated jets are expected to have distinct fragmentation patterns at, due to their different colour charges. Although the fragmentation pattern is, theoretically, expected to be different, until recently there was no operational hadron level definition of jet flavour in this sense. In reference [7], an operational definition based on jet topics, [6] and classification without labels, [8], was proposed, in a fully data-driven way.

The problem of discriminating quark and gluon jets has a long history. Already in the nineties, neural networks were being deployed to attack this problem [9]. Nowadays the field evolved further, and the systematics of quark-gluon tagging and a theory of quark-gluon discrimination have been addressed in references [10, 11].

Having pure samples of quark and gluon initiated jets would, therefore, be very useful to study not only differences in their fragmentation patterns but also to study certain observables, separately for each parton flavour (quark or gluon), allowing to test QCD and SM predictions. These pure samples are not available, however. Experimentally we need to rely on quark, and gluon enriched samples, always containing some contamination. Discriminating these classes of jets even on a distribution level is, in this way a useful contribution to the field.

In reference [6], the jet multiplicity was used to identify the underlying quark and gluon initiated jet topic distributions. This problem is addressed to validate the algorithm's implementation. This work was reproduced and it's presented in figure 5 to the left, while to the right we present a plot of both reducibility factors when these are smaller than one, with $\kappa(1|2)$ in red and $\kappa(2|1)$ in blue, where 1 and 2 correspond to the first and second input distribution in the legend from top to bottom. One, two and three sigma bands (with increasing transparency) along with green circles marking the anchor bins are also drawn.

In the presented plot, the input distributions to the algorithm are the Z + jet and dijet jet multiplicities (in green and aquamarine blue). The obtained topic distributions (in purple and blue) seem to correspond to the expected quark and gluon initiated jets' distributions (in orange and yellow). Furthermore, the Pythia quark (gluon) fractions for our processes, reconstruction and p_T selection are 37.91% (62.09%) for dijets and 86.87% (13.13%), which are contained in our fractions' uncertainty to one or two sigma.

The ternary plots for the jet multiplicity is presented in figure 6, where we made use of an additional $\gamma + jet$ sample. The conditions used to define the colour of the lines are presented in the plot, with the quantities defined in equations (8) and (9).

Other than the lines - drawn fairly transparent - resulting from the two random, independent and uniformly distributed elements of the convex hull, we draw lines across the plots axis, which correspond, to ignoring one of the samples completely - these are drawn with no transparency. The bottom line (the one close to the Dijets axis) corresponds to our two topic and two sample case and as we can see in figure 6, it has a yellow colour, which is telling us that: 1) the topic is reconstructed with very decent errors; 2) the topic errors enclose at one sigma the reference points; and 3) the topic points are at an average reference sigma bellow three, which we considered to be good in the previous results.

We also draw the lines corresponding to taking averages of the input distributions with one distribution common to both averages. These are the lines forming the inverted triangle inside the plot, also drawn with no transparency. We note that for the multiplicity, having the γ + jet sample as the common in the averages, would be the best possibility if we had considered only averages in the algorithm. Still, this is worse than our initial results, having worse errors. To finish the discussion on the plot for the multiplicity, we note that there are indeed possibilities that improves the results from using two samples. In particular, there are a series of red lines that use an almost pure sample of dijets and a near symmetric mixture of the remaining inputs, with lower fractions for the Z + jet sample. These lines can be due to the actual shape of the distributions or merely an effect of increased statistics. In any case, we can conclude that including a third sample improves the results of the algorithm

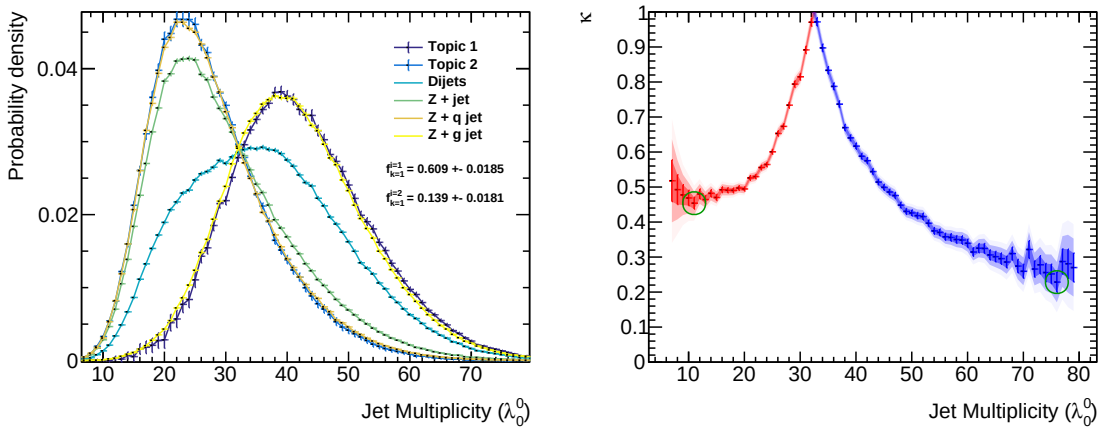

Figure 5: (right) Demix applied to the jet's multiplicity from Z + jet and dijet samples (in green and aquamarine). The extraction of pure quark and gluon initiated jet distributions is attempted and references are drawn in yellow and orange. The obtained distributions are drawn in purple and blue. (left) Plot of the reducibility factors when smaller than one.

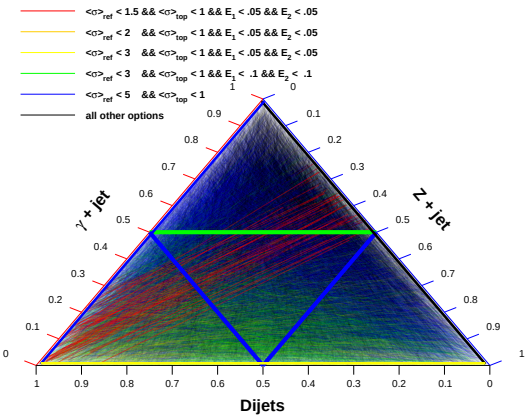

Figure 6: Ternary plot for the problem of figure 5 with an additional $\gamma + jet$ sample. Lines correspond to two elements of this space, needed by the Demix algorithm to extract two topics. Each possible color of the drawn lines correspond to a categorical class constructed with the quantities defined by equations (8,9) and are presented at the top of the figure.

but only for certain combinations of elements of the convex hull. Why this is so should be investigated in further work. Also, lines originating from the upper corner, corresponding to a pure γ + jet distribution are black, indicating that this is a bad point.

4.2. Quark vs gluon separation with JEWEL's quenching model

The separation of quarks and gluons in vacuum has been motivated in the previous subsection. In the presence of the QGP, this separation would entail having two fundamentally different QCD probes of the medium. Furthermore, different jet substructure observables are sensitive to different aspects of QCD dynamics. If the fractions of quark and gluon initiated jets change from pp to PbPb collisions, the jet substructure observable distributions will have this modification convoluted with the actual QGP interaction. This competition between effects entails the fact that separating these two types of jets in heavy-ion collisions allows the separation of these effects, and hence more precise measurements of the QGP's properties.

The multiplicity seems to extract the quark topic reliably. The gluon topic is also compatible with our reference, but deviates considerably for low multiplicities. For the jet's mass, the topic for gluon jets is relatively well reconstructed, albeit with some deviations for higher multiplicities. The quark topic is also a bit off of the expected distribution, as is usual for this observable. Overall all observables yield fractions compatible with each other at one sigma.

4.3. Quenched vs unquenched separation

In this subsection, we present our main results, pertaining to the separation of quenched and unquenched jets in the presence of the QGP. In figures 10 and 9, we present results for the jet mass for Z + jet and dijet samples, respectively. Here we take the fraction of modified jets in vacuum as an indicator that the algorithm performed well. We know this fraction is zero in theory, but the algorithm will still return us some non-zero fraction. The lower this value is, the better the algorithm is considered to perform.

The fractions of quenched jets in the medium (the first fraction presented in both plots) is consider-

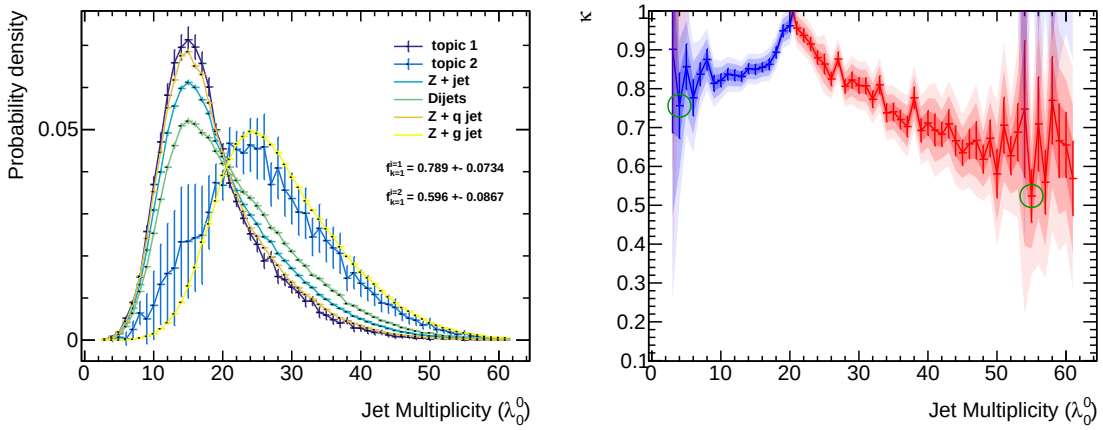

Figure 7: (right) Demix applied to the jet's multiplicity from Z + jet and dijet JEWEL 0-10% centrality samples (in green and aquamarine). The extraction of pure quark and gluon initiated jet distributions is attempted and references are drawn in yellow and orange. The obtained distributions are drawn in purple and blue. (left) Plot of the reducibility factors when smaller than one.

ably lower for the Z + jet samples, with respect to dijets. This difference is understood by the different fractions of quark and gluon initiated jets in both samples. The dijet sample is more gluon rich, and gluons, having a softer fragmentation pattern (more constituent particles with the initial parton energy more equitably distributed among them) are expected to be more modified by the QGP. This logic leads to the conclusion that the gluon rich dijet sample yields more modified jets (in population and intensity). It is good to see that the algorithm reflects this phenomenological expectation. The fractions of quenched jets in the vacuum samples are essentially the same, being compatible with each other at one sigma.

Introducing a lower centrality sample indeed improves the separation according to our measure. In figure 11, the ternary plots for the convex hulls of the jet's mass for the three samples per process are presented. The line in the left side of the triangle corresponds to the previous two samples case and is found to be in the worst class. Considering the two PbPb samples seems to yield the best results for only using two input samples. For the dijet samples, none of the tested elements yields an excellent agreement with the vacuum sample. However, for the Z + jet sample, the algorithm finds that using a pure sample of the higher centrality and an average of the other two yields the best results with acceptable error bars and within the reference's error bars at less than one and a half sigma on average.

These differences between the results for both

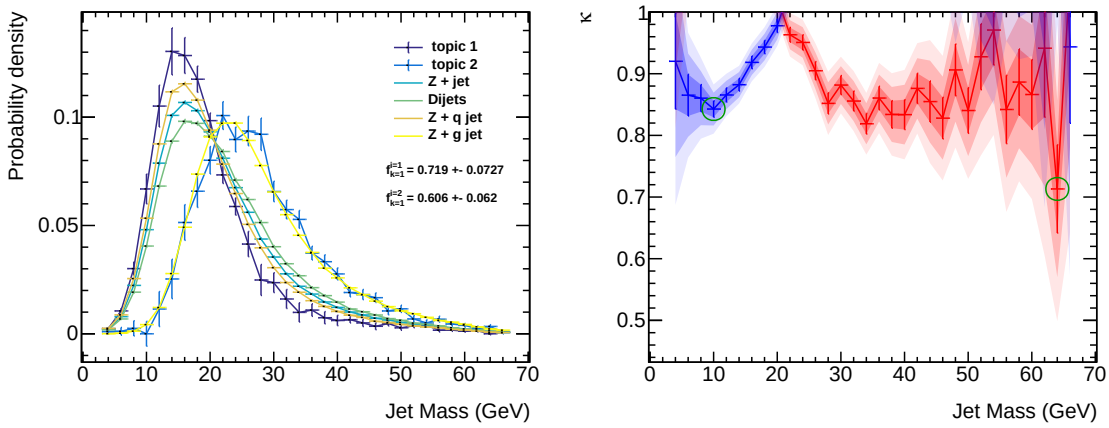

Figure 8: (right) Demix applied to the jet's mass from Z + jet and dijet JEWEL 0-10% centrality samples (in green and aquamarine). The extraction of pure quark and gluon initiated jet distributions is attempted and references are drawn in yellow and orange. The obtained distributions are drawn in purple and blue. (left) Plot of the reducibility factors when smaller than one.

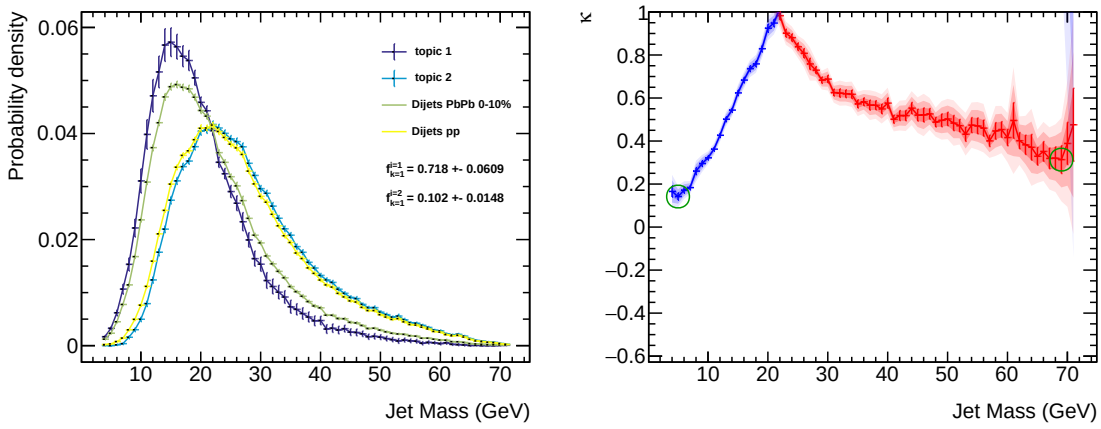

Figure 9: (right) Demix applied to the jet's mass from JEWEL's dijet 0-10% centrality and vacuum reference samples (in olive and yellow). The extraction of modified and unmodified jets distributions is attempted. The obtained distributions are drawn in purple and blue. (left) Plot of the reducibility factors when smaller than one.

processes can be explained through the argument of sample dependence. Being that dijets are a gluon rich sample and that gluon initiated jets are more likely to interact with the medium, the gluon fractions will be different for all samples of this process. Quenched and unquenched jets then are dependent on the sample for this process, where more gluon modified jets are expected to be present in the lower centrality, with respect to the higher centrality. For Z + jet the fraction of gluon jets is small, and hence yields a smaller effect of sample dependence.

5. Conclusions

In this article, we conclude that considering a third sample improves the performance of the Demix algorithm in the separation for quark and gluon initiated jets in vacuum. The separation of these types of jets with JEWEL's quenching model is performed and needs further developing with the use of a third sample (like γ + jet or trijet events).

The separation of quenched and unquenched jets with the Demix algorithm was done by applying it to the jet's mass of high centrality PbPb collisions and pp collisions. The results are shown to be consistent with theoretical and phenomenological expectations. The addition of a third centrality sample is shown to improve performance, but still, the algorithm seems to underperform for the dijet samples due to sample dependence.

A possible future directions passes by the extension to more than three samples and ways to study the performance of the algorithm based on the ini-

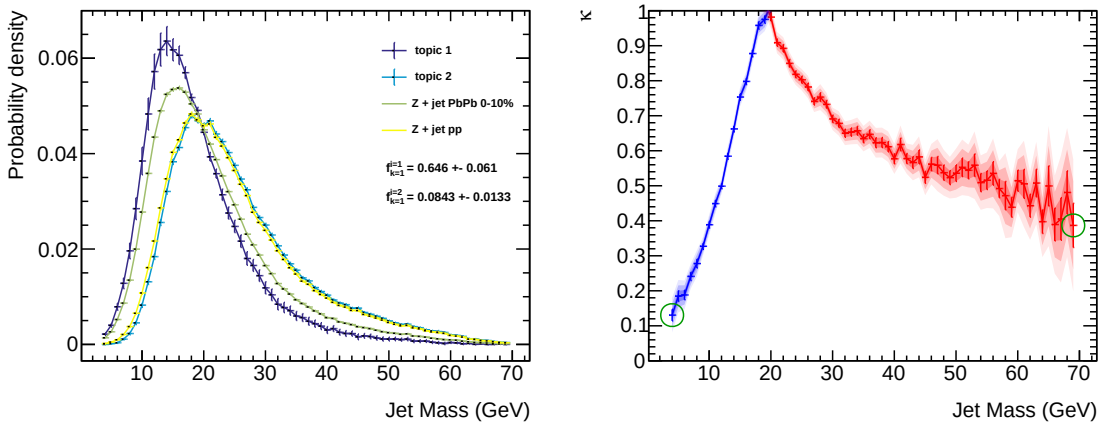

Figure 10: (right) Demix applied to the jet's mass from JEWEL's Z + jet 0-10% centrality and vacuum reference samples (in olive and yellow). The extraction of modified and unmodified jets distributions is attempted. The obtained distributions are drawn in purple and blue. (left) Plot of the reducibility factors when smaller than one.

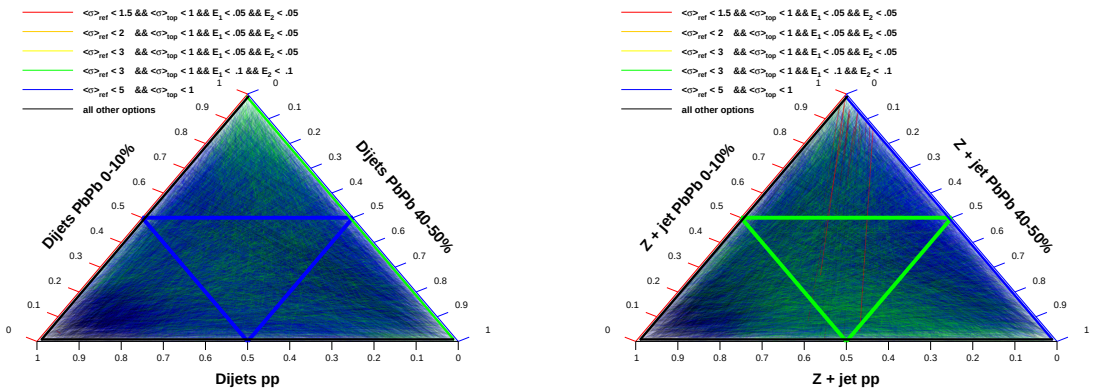

Figure 11: Ternary plots for the problem of figures 9 and 10 with an additional 40-50% centrality sample. Lines correspond to two elements of this space, needed by the Demix algorithm to extract two topics. Each possible color of the drawn lines correspond to a categorical class constructed with the quantities defined by equations (8,9) and are presented at the top of the figure.

tial convex hull elements. The characterization of the convex hull for more than four samples is a difficult task. Another exciting future direction is the inclusion in current heavy ion MCs of a way to know if a given jet was or not modified by the medium in a significant way. Furthermore, the separation of quark and gluon jets in PbPb collisions would allow for better quenched/unquenched discrimination, and in turn further constrain the models we have today.

Acknowledgements

The author would like to thank Professors Liliana Apolinário and Guilherme Milhano for supervising this work, for the invaluable contributions whether through discussions on the theoretical and phenomenological aspects whether on the implementation. The author thanks the support of Fundacao para a Ciencia e a Tecnologia through a BI in project 'Understanding Big Data in High Energy Physics: finding a needle in many haystacks' (PTDC/FIS-PAR/29147/2017).

References

[1] Korinna C. Zapp. JEWEL 2.0.0: directions for use. *Eur. Phys. J.*, C74(2):2762, 2014.

[2] Julian Katz-Samuels, Gilles Blanchard, and Clayton Scott. Decontamination of Mutual Contamination Models. *arXiv e-prints*, page arXiv:1710.01167, Sep 2017.

[3] Morad Aaboud et al. Measurement of the azimuthal anisotropy of charged particles produced in $\sqrt{s_{\mathrm{NN}}} = 5.02$ TeV Pb+Pb collisions with the ATLAS detector. *Eur. Phys. J.*, C78(12):997, 2018.

[4] Morad Aaboud et al. Measurement of the nuclear modification factor for inclusive jets in Pb+Pb collisions at $\sqrt{s_{\mathrm{NN}}} = 5.02$ TeV with the ATLAS detector. *Phys. Lett.*, B790:108–128, 2019.

[5] David M Blei. Probabilistic topic models. *Communications of the ACM*, 55(4):77–84, 2012.

[6] Eric M Metodiev and Jesse Thaler. Jet topics: disentangling quarks and gluons at colliders. *Physical review letters*, 120(24):241602, 2018.

[7] Patrick T Komiske, Eric M Metodiev, and Jesse Thaler. An operational definition of quark and gluon jets. *Journal of High Energy Physics*, 2018(11):59, 2018.

[8] Eric M. Metodiev, Benjamin Nachman, and Jesse Thaler. Classification without labels: Learning from mixed samples in high energy physics. *JHEP*, 10:174, 2017.

[9] M. A. Graham, L. M. Jones, and S. Herbin. A Neural network classification of quark and gluon jets. *Phys. Rev.*, D51:4789–4807, 1995.

[10] Philippe Gras, Stefan Höche, Deepak Kar, Andrew Larkoski, Leif Lönnblad, Simon Plätzer, Andrzej Siódmok, Peter Skands, Gregory Soyez, and Jesse Thaler. Systematics of quark/gluon tagging. *Journal of High Energy Physics*, 2017(7):91, 2017.

[11] Andrew J. Larkoski and Eric M. Metodiev. A Theory of Quark vs. Gluon Discrimination. 2019.