LU TP 17-12
May 2017

Study of the Loss of Information from the Modified Mass-Drop Tagger in PYTHIA 8 for $Z^0 \to q\bar{q}$

Markus Ernstsson

Department of Astronomy and Theoretical Physics,
Lund University

Bachelor thesis supervised by Torbjörn Sjöstrand

LUND
UNIVERSITY

Abstract

In this thesis, the information loss from the modified Mass-Drop tagging of a $Z^0 \to q\bar{q}$ jet is studied for a pp collision at LHC centre-of-mass energies in PYTHIA 8 using FastJet 3. The jet created by the Z^0 has minimum transverse momentum of 200 GeV. This is then searched for subjets via two methods: a Cambridge/Aachen algorithm and a modified Mass-Drop Tagger with different z_{cut} values. The subjets located through these two methods are compared and specific regions of the jets are examined further.

The transverse momenta used as a measurement of information, and the amount of transverse momentum removed by the modified Mass-Drop Tagger in the different regions is used to measure the information loss. There is a sign of information loss by the modified Mass-Drop Tagger, in that it removes more transverse momentum than only the background. Indications of the removal of particles that are caused by the Z^0 decay are found and further ways of investigating this removal are proposed.

Populärvetenskaplig sammanfattning

I samband med att teknologin har utvecklats, har också vår förmåga att observera mindre och mindre objekt utvecklats. Detta är för att partiklar kan accelereras till högre och högre fart och därmed till högre energier. Utforskningen av hadroner, med partikelacceleratorer som Large Hadron Collider (LHC), har givit forskare en djupare förståelse om hur kvantkromodynamik (QCD), en av lärorna om materians innersta strukturer, fungerar, men också skapat nya forskningsområden.

Ett område som har uppkommit är studier av "underkvaststrukturer". En "kvast" inom partikelfysik är en samling av hadroner, partiklar bestående av kvarkar och gluoner, med hög energi som alla är på väg åt ungefär samma håll. Denna kvast är vitt skild från den vardagliga kvasten, förutom dess form; den är spretig från spåren av alla partiklarna inuti den. Kvastar skapas av kvarkar och gluoner som genomgår två processeser: partonkaskad och hadronisering. Den första av dessa processer kan leda till att det skapas en typ av understruktur i en kvast: flera täta samlingar av partiklar bland en större samling av partiklar. Dessa täta samlingar kallas för "underkvastar". Detta leder till en underkvaststruktur som är av stort intresse.

Underkvaststrukturer hittas av speciella algoritmer som trimmar, beskär eller märker ut vad som definieras som underkvastar. Detta kan, dock, leda till att viss information om kvasten och därmed händelsen kastas bort och det är detta som denna avhandling studerar. Försvinner information, och i så fall, vad och hur mycket?

Contents

List of Figures

1 Introduction

Over the years, luminosities at hadron colliders have increased. The Large Hadron Collider (LHC) at CERN collides protons at energies far above the electroweak scale. These protons are bundled up in intense bunches which are a few cm long. This means that for each bunch crossing, several collisions could occur at the same time. However, one wishes to view each collision as a single, unique event, which means that particles from one event should not be mixed up with particles from a different event. This is not difficult if the particle collisions occur far away from each other. If the particle collisions/interactions happen very close to each other, however, they can be mixed up. Charged particles are easy to distinguish since their trajectories can be measured in trackers. The issue arises with neutral particles. They can only be measured in calorimeters, which are less accurate than trackers [1]. It can therefore be difficult to separate which neutral particles are from which event.

A "hard" parton, of high transverse momentum, will go through two processes: a parton shower and hadronisation. These effects will create hadrons that travel in, roughly, the same direction as the original parton, and soft hadrons that are more dispersed. Such a collection of final-state particles, created by partons, is known as a "jet". Furthermore, heavy particles can decay to several partons, and each parton goes through these processes resulting in several collections of high activity — that is, several jets.

The energies involved at LHC, and similar colliders, can produce electroweak bosons and top quarks that have a large boost transverse to the beam axis, thereby, possibly, bundling up the hadronic decays into a single large jet. The coloured decay products of heavy-particle decays can create jets of their own, within the large jet. This will lead the large jet to have substructures with regions of more activity and regions with less. If a region of high activity fulfils conditions determined by certain algorithms, that region is a "subjet". The subjets and the regions around them also contain background particles, that are created from the interaction of the softer partons as well as other events.

In order to find heavy particles via their coloured decay products, the study and sorting of jets and their substructure is important. Since hadron colliders cannot detect what happened at intermediate steps but only the end product, the sorting of the clusters of particles — separating the different events — that are detected is of great interest. There are also background particles detected that are not part of the jet, increasing the difficulty of finding specific decays. In order to clean up jets, different types of tagging methods are used. These methods work by taking a large cluster of particles, breaking it up into smaller clusters, applying some parameters that define what a subjet should be, and finally remove the smaller clusters that do not fit the classification. However, when these tagging methods are used, one runs the risk of removing a cluster that is actually the result of a specific decay or from a strong emission, thereby effectively losing information about the event.

This thesis starts off in section 2 with going through the background of jets, what they are and why event generators are used in these kinds of studies. Section 3 then goes on to explain how jets are discovered and the different ways one can do it depending on what kind of jets one is searching for. How jets are tagged and two common ways of doing this is explained. In section 4 it is explained how this loss of information is studied. Using PYTHIA 8.2 [2], a $p\bar{p}$-collision with the hardest event being $q\bar{q}/qg \to Z^0 g/q$ is simulated. A $Z^0 \to q\bar{q}$ jet is then located with the anti-k_t algorithm and the subjets from the $q\bar{q}$ are searched for using the modified mass-drop tagger from FastJet 3.2 [3] and the Cambridge/Aachen algorithm. Finally the results of the study are presented, and a conclusion based on these results is drawn.

2 Defining jets

The complexity of hadron collisions is difficult to describe, due to the many processes involved. However, in order to first grasp jets and their importance, one can start off with a simple case and then gradually make it more complex.

2.1 Partons interacting

Start off with a simple parton scattering, illustrated in fig. 1, where two quarks interact via a gluon and then the quarks are scattered with a high transverse momenta. Focusing on only one of the quarks, it will exhibit two processes as it travels: First, during small times t, the quark will experience a parton shower. Then, for larger t, the quark will go through a hadronisation process. Both of these processes occur at very short time scales. In order to understand this one can start out with just a hadronisation and then add a parton shower.

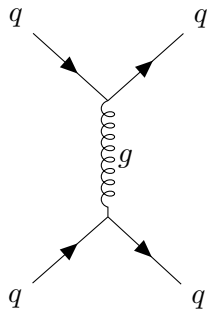


Figure 1: Two quarks interacting via a gluon.

Due to colour confinement, quarks and gluons can not exist individually; they must combine to make colour-neutral hadrons. This means that the partons themselves cannot be detected, but the hadrons that are collinearly emitted can. This collinear emission is important, since otherwise the hadrons would be spread out and it would be difficult to know anything about the partons. Partons are therefore detected as clusters of hadrons and other particles with similar rapidity and azimuthal angle; these clusters, produced by the hadronisation of the scattered quarks, are known as jets.

The transverse (relative to the hard parton) momenta of produced hadrons is small, $\langle p_\perp \rangle \approx 0.4$ GeV. If the quark has $E \sim 100$ GeV and the energy for each emission is split 50/50, the first hadron will be emitted at an angle $\theta \sim \frac{\langle p_\perp \rangle}{E/2} = \frac{0.4}{50} = 0.008$ rad. This is a very small angle, but for each emission, the available energy will decrease and the angle will increase. For this scenario, the fifth emitted hadron will have an angle of $\theta \sim \frac{0.4}{100/2^5} = 0.128$ rad. This is illustrated in (a) in fig. 2. In real life, the hadrons can be emitted with varying energy fractions and may therefore not be ordered in the same way as in (a). Still, the last hadrons to be emitted generally have larger angles than the first ones, due to there being less energy available. The whole process of hadronisation will lead to some of the hadrons being collimated and creating a tight region of hadrons; a jet core.

There is not only hadronisation at play during partonic scattering, but also an effect known as a parton shower. Since quarks can emit gluons and the gluons can branch into more gluons or quark-antiquark pairs (see (b) in fig. 2), high-energetic partons can be spread out over a wider

region, thus creating a wider cluster of final-state particles. This effect will lead to jets being wider.

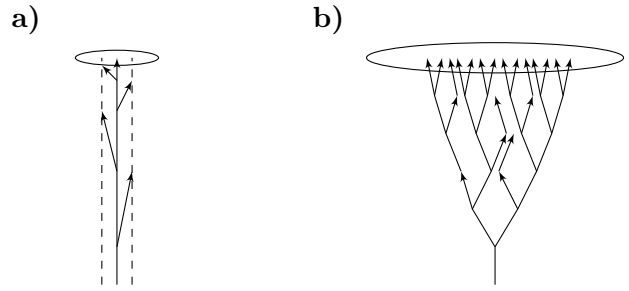


Figure 2: (a) Schematic illustration of only the hadronisation of an energetic quark, were the dashed lines indicate the transverse size. Each time a hadron is created, the available energy for the next hadronic emission is reduced, thus increasing the angle at which they are emitted. (b) Schematic illustration of a parton shower spreading out the partons, and widening the jet. The circles indicate the areas of the jets produced.

2.2 Jet substructure

The widening from parton showers gives rise to substructures in the jet, since there will be areas that have more activity than others. As an example, one can start off by studying the $Z^0 \to q\overline{q}$ decay in its rest frame, (a) in fig. 3. Here the quark and antiquark travel in opposite directions and go through hadronisation, creating hadrons travelling in, roughly, the same two directions. This will lead to two narrow jets. If the system is boosted, (b), the quark and antiquark will instead be much closer and the soft particles travelling in the boost direction will become harder. For large boosts, this can lead to a single, wide jet containing most of the particles, instead of two separate jets. However, there will be a form of structure within this jet, since the areas

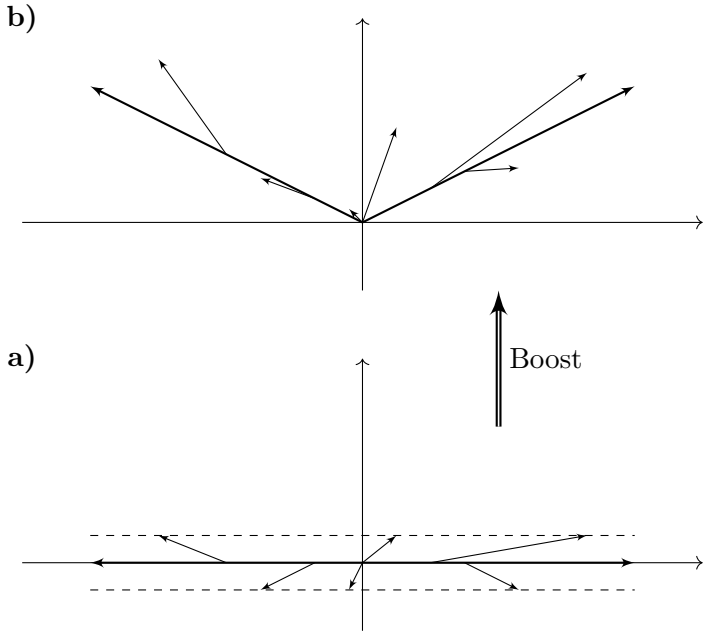


Figure 3: A $Z^0 \to q\overline{q}$ decay in both the rest frame, (a), and in a boosted frame, (b), illustrated without parton shower. The lengths correspond to the energies.

around the two quarks can be more active than other areas.

Adding the effect of a parton shower to the quark and antiquark will lead to a more complex structure within the jet. For instance, if one of them branches into a very hard gluon and a softer quark, that gluon can create a region of increased activity, similar to that of the quark. This region can therefore contain a subjet, while the softer quark may not form a subjet. The antiquark may also branch out to a antiquark and gluon with similar hardness, leading to possibly three or four or more subjets, depending on how the parton shower evolves. The jet may also contain regions that have a higher activity, or $p_\perp$, than the background. One can see in fig. 3 that the region between the quarks will have higher activity than the regions outside quarks, since the boost will suppress those particles. However, there are not necessarily any subjets created, due to it not being concentrated, but more spread out.

2.3 Event generators

It is difficult to know what happens and what generates the final states observed at particle colliders such as the LHC. This is partly due to the several hundreds of particles produced and the widely varying momentum of these particles. In particle physics one will need to compute complicated matrix elements, deal with the complex, and unknown, nature of QCD processes, handle divergences and near-divergences. This can be difficult to do analytically and can be time consuming. One would therefore like to let a computer handle it. The entirety of a hard collision can, however, be difficult to simulate.

One can however, split the hard collisions up into three different regimes based on the momentum transfer: one for low scales, one for high scales and one for everything in between. These three regimes are all suited for simulation via Monte Carlo techniques. One can thereby create a Monte Carlo event generator with wide capabilities of simulating LHC processes. This event generator can be used to find new physics among the Standard Model background, design new experiments and compare predictions to data. If one wishes to dig deeper, there is a thorough review of event generators in [4].

As such, one can test methods of filtering the events created by particle colliders in these event generators and tailor them to yield the most effective filtering methods. In the case of jets, one can test different jet algorithms and see what kinds of jets they produce and their efficiency as well as different methods for sorting and tagging these jets. This is useful, given that for particle colliders, one does not know what it was that interacted or how everything decayed, only the final-state particles. For an event generator it is all known, so one can check how well an algorithm fares when applied to known generator events and use this to optimise the algorithm. For a given process a generator can also predict how many jets to expect and with what properties, which is most helpful for experimental studies.

3 Jet algorithms

In order to find a heavy particle that decays to partons, or check partonic scattering angles, etc, certain algorithms are needed to specify what a jet is — so called jet algorithms.

The earliest jet algorithms were cone algorithms that worked by selecting a certain particle as a starting seed, around which a cone is then centred and the momenta, mass, etc.[1] of the particles

[1]The charge, spin, etc. quantities of the particles are not combined.

within this cone are combined. This will shift the cone's centre, which can then be used as the new seed to combine more particles etc. until a stable cone is located. Cone algorithms are however not used very often today, due to their infrared and collinear unsafety. More on cone algorithms can be found in [5]

3.1 Infrared and collinear safety

Infrared and collinear (IRC) safety is the property that a set of hard jets should remain unchanged if an event is modified by a collinear splitting or a soft emission. Both collinear splitting and soft emission are illustrated in fig. 4. For an algorithm to be collinear safe, it needs to produce the same amount of jets in configuration (b) as in configuration (a). In other words, a general algorithm may use the hardest particle as a seed to then combine with other particles within a certain distance Δy, as such this general algorithm should therefore only find one jet in configuration (a), given that the other two particles are within a distance Δy to the hardest particle. In configuration (b), however, the hardest particle is the leftmost particle. This will therefore provide the seed, but the jet produced from this seed may not contain the rightmost particle. In that case, the rightmost particle will then be the new seed, since it is the hardest of the ones left, and create another jet, thereby leaving configuration (b) with two jets, making the algorithm collinear unsafe.

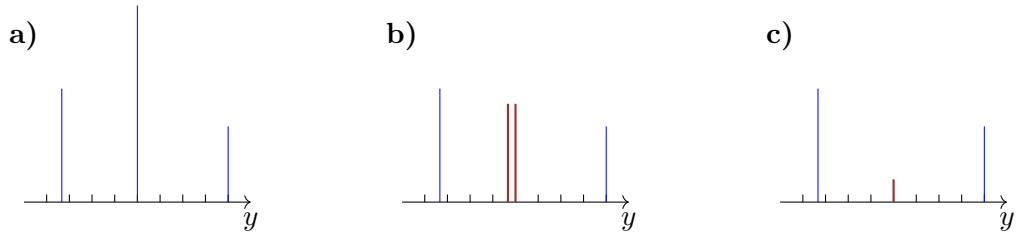


Figure 4: (a) and (b): An illustration of a collinear splitting. (c): Illustration of a soft emission. The partons are vertical lines, their height is proportional to their transverse momentum, $p_\perp$, and the horizontal axis indicates rapidity/pseudorapidity, y/η. They are assumed to have the same, or very similar azimuthal angle ϕ.

If a general algorithm instead uses all particles as possible seeds, or iterates and tries to find all the stable cones, one can see that configuration (c) may yield problems. If there was no soft emission i.e. no middle particle, and the other particles were separated with a distance $> \Delta y$ but $< 2\Delta y$, it would result in two jets. However, if there is a soft emission, as illustrated, the stable cone produced from the middle particle would be the hardest one and the algorithm would result in only one jet. This would make the algorithm infrared unsafe.

Due to this, cone algorithms have since the start of the LHC been replaced by sequential recombination jet algorithms. If one wishes to read more about cone algorithms and IRC safety, this can be done in [5].

3.2 Sequential recombination jet algorithms

Instead of dealing with cones, one can compare the distance between each particle i and j, for all particles, and if the smallest distance is below a threshold distance, then those particles are

recombined into a single new particle and the process is repeated until there is no distance that is below the threshold. This distance can be defined as

$$\begin{aligned} d_{ij} &= \min(p_{\perp i}^{2l}, p_{\perp j}^{2l}) \frac{\Delta R_{ij}^2}{R^2}, \quad \Delta R_{ij}^2 = (y_i - y_j)^2 + (\phi_i - \phi_j)^2, \\ d_{iB} &= p_{\perp i}^{2l}, \end{aligned} \tag{3.1}$$

where d_{ij} is the distance between particle i and j, d_{iB} is the distance from particle i to the beam, R is a parameter determining the size of the jet. The rapidity $y = \frac{1}{2}\ln\left(\frac{E+p_z}{E-p_z}\right)$ $\left(\text{or the pseudorapidity } \eta = \frac{1}{2}\ln\left(\frac{|\vec{p}|+p_z}{|\vec{p}|-p_z}\right)\right)$ and the azimuthal angle ϕ.

These coordinates are used in high-energy physics to describe particles and compare them. The transverse momentum $p_\perp$ is used as a measurement of how hard a particle or event is and is Lorentz invariant under boosts along the beam axis. The azimuthal angle ϕ is used because of the symmetry that occurs during particle collisions. The rapidity y has the property that Δy is Lorentz invariant[2] under boosts along the beam axis. This means that if one measures the rapidity difference between two particles, that measurement is not dependent on the boost of the lab frame along the beam axis. Also, since the four-momentum of a particle can be written as

$$p = p_\perp(\cosh y, \cos\phi, \sin\phi, \sinh y) \tag{3.2}$$

one gets that the invariant mass of two particles, $m_{12}^2 = (p_1 + p_2)^2$, is

$$m_{12}^2 = 2p_{\perp 1}p_{\perp 2}\left(\cosh(y_1 - y_2) - \cos(\phi_1 - \phi_2)\right) \approx p_{\perp 1}p_{\perp 2}\Delta R_{12}^2, \tag{3.3}$$

which is invariant for rotations of the particles around each other, with preserved ΔR_{12}.

The parameter l determines the distance measure used: $l = 1$ gives the k_t algorithm, $l = 0$ gives the Cambridge/Aachen (C/A) algorithm and $l = -1$ gives the anti-k_t algorithm. They all operate using the same procedure:

1. Determine all d_{ij} and d_{iB} from eq. (3.1)
2. Find the smallest of all d_{ij} and d_{iB}.
3. If it is d_{ij}, recombine i and j into a single new particle and return to step 1.
4. If it is instead d_{iB}, i is added to the list of final-state jets, if it is above a certain threshold $p_{\perp,\min}$; it is, in either case, removed from the list of particles. Then return to step 1.
5. If no particles remain, stop.

The combination of this procedure with the distance measurement eq. (3.3) yields a so called inclusive algorithm which includes all particles in final-state jets. There is also an exclusive algorithm which is very similar to the inclusive one. The exclusive algorithm does not include all particles in their own final-state jet, but instead combines the very soft particles into a beam jet.

[2]The difference in pseudorapidity is only invariant if the particles are massless. However, if they are highly energetic, the pseudorapidity can be approximated to be the rapidity and therefore $\Delta\eta$ is basically invariant for boosts along the beam axis.

k_t algorithm

The k_t algorithm is IRC safe and therefore good from a theorist's point of view. It yields irregular jets due to the clustering of soft particles. Irregular jets are jets that have jet "areas" of irregular shapes.

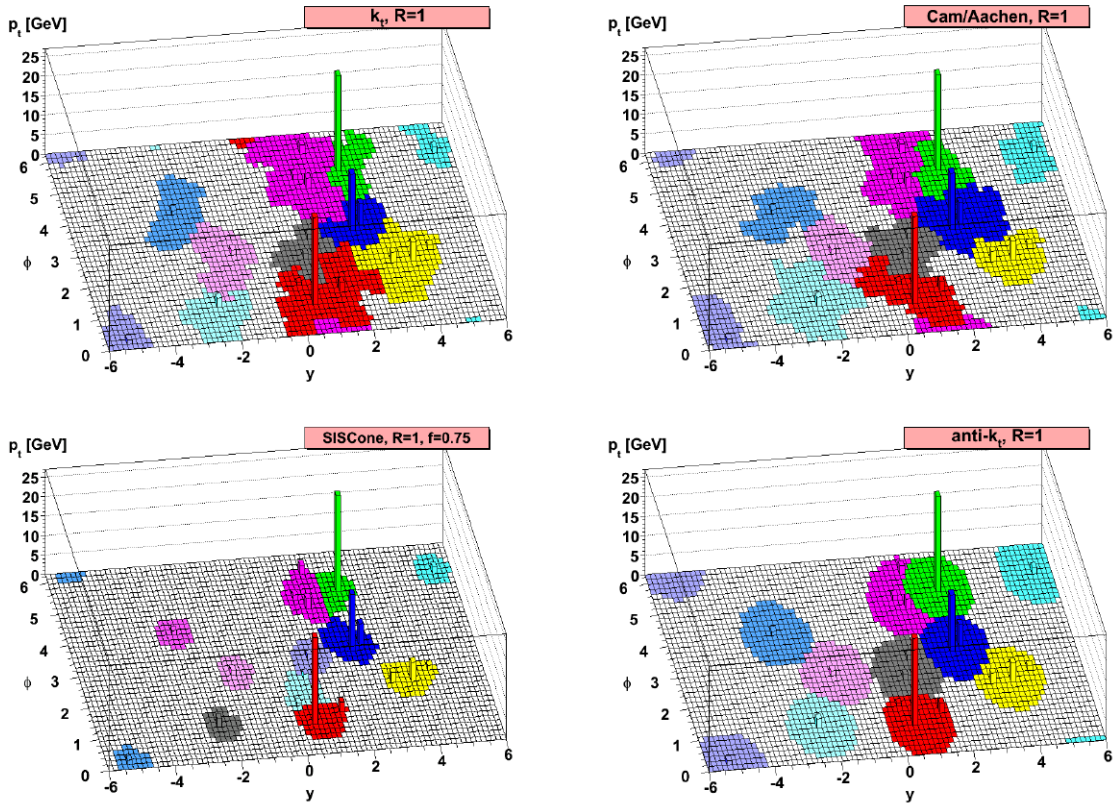


Figure 5: A sample parton-level event, together with a random set of soft "ghost" particles, clustered using four different algorithms. Illustrates irregular jet areas as well as how the different algorithms work. Image taken from [5], with permission.

A jet's area, displayed in fig. 5, is built up differently depending on how the specific jet algorithm applied behaves. This area is a type of measurement of a jet's susceptibility to contamination from soft radiation. If a particle is within a jet's area, it will be clustered with the jet. A more precise definition of jet area can be found in [5].

Cambridge/Aachen algorithm

The C/A algorithm clusters particles based on their ΔR distance and not energy, meaning that particles with very different energies but close to each other can be considered to be part of the same jet. This makes it good for finding subjets. It is IRC safe, but it can give somewhat irregular jets, like the k_t algorithm.

Anti-k_t algorithm

Since the anti-k_t algorithm has $l = -1$ in eq. (3.1), it forms clusters that grow outwards from hard particles. It resembles cone algorithms in that it yields circular hard jets, but it is also IRC safe, meaning that it is better than the cone algorithms.

3.3 Tagging jets/Limiting jets

In colliders, events can occur at several different locations at the same time. As explained in section 1, this poses no issue for the distinction of different charged particles, but the same is not true for neutral particles. This means that it is more difficult to pinpoint the event from which the neutral particles originated. In each event there are also a lot of particles created from the interaction of the softer partons that are parts of the hadron that can then also go through the parton shower and hadronisation steps.

Since the protons, and other hadrons, consist of several partons, each time a proton collides with another proton at high energy, it can be seen as the partons inside the protons colliding and interacting. This can lead to background radiation and noise, caused by the interaction of the softer partons, that can blend in with the jet, effectively making it more difficult to know what caused the jet. So in order to "clean up" the jet and find out which particles belong in the set of final-state particles and which do not, different methods of grooming, filtering and tagging jets are used.

The focus will lie on jet tagging methods, which are algorithms that determine which particles, in a collection of final-state particles, likely are the result of boosted heavy-particle decays and which are just QCD background radiation. This is done by "declustering" a cluster of particles and discarding soft clusters until a hard structure has been discovered.

3.3.1 Mass-drop tagger

A Mass-Drop Tagger (MDT) [6] is a tagging method which is used to search for a generic boosted heavy-particle decay. It works by first finding a jet using an anti-k_t algorithm and then using a C/A algorithm on the constituent particles of the jet. This will yield a jet i with a radius R similar to the original. Then the following procedure is performed on it:

1. Break the jet i into two subjets by undoing its last stage of clustering. Label them i_1, i_2 such that $m_{i_1} > m_{i_2}$.
2. If there was a significant mass drop, $m_{i_1} < \mu m_i$, and the splitting is not too asymmetric, $y = \dfrac{\min\left(p^2_{\perp i_1}, p^2_{\perp i_2}\right)}{m_i^2}\Delta R^2_{i_1,i_2} > y_{\text{cut}}$, then deem i to be the jet that contains the decay products and exit the loop.
3. Otherwise redefine i to be equal to i_1 and go back to step 1.

The two dimensionless parameters μ and y_{cut} are typically chosen to be $\mu \approx 2/3$ and $y_{\text{cut}} \in [0.09, 0.15]$.

3.3.2 Modified MDT

As the name suggests, a modified MDT (mMDT) [7] is a modified version of the MDT and works almost the same as the MDT, except it removes the mass dependence, μ, from the MDT and instead focuses on the transverse momentum. The mass dependence in the MDT can be removed to create the mMDT, if μ is chosen to be relatively large, $\gtrsim 1/2$. Given that μ usually is larger than this, the μ dependence disappears and one only needs a symmetry cut, z_{cut}, in order to tag a jet. So the algorithm for the mMDT instead becomes:

1. Break the jet i into two subjets by undoing its last stage of clustering and label them i_1, i_2 such that $p_{\perp i_1} > p_{\perp i_2}$.
2. If the splitting is not too asymmetric, $z = \dfrac{p_{\perp i_2}}{p_{\perp i_1} + p_{\perp i_2}} > z_{\text{cut}}$, then deem i to be the jet that contains the decay products and exit the loop.
3. Otherwise redefine i to be equal to i_1 and go back to step 1.

z_{cut} is normally chosen to be around 0.1 [7].

4 Jet study

In order to determine if the mMDT throws away clusters of particles containing information about the event, a PYTHIA simulation is used together with FastJet.

4.1 Finding the subjets with mMDT

A pp collision is set-up to simulate LHC centre-of-mass energies, $E_{\text{cm}} = 14$ TeV. Hard processes of $\text{q}\overline{\text{q}} \to \text{Z}^0\text{g}$ and $\text{qg} \to \text{Z}^0\text{q}$ are generated with a minimum $p_\perp$ of 190 GeV and only the quark decay products of the Z^0 are generated. An anti-k_t algorithm is used to find the jet formed from the decay of the Z^0 into two quarks. The anti-k_t algorithm is set up with $R = 1$ and $p_{\perp,\text{min}} = 200$ GeV.

The reason for this value of R is due to both the fact that the value of R is usually taken to be around 1, as well as the minimum angle between the jets created by the daughter particles. Since the Z^0 boson has a mass of ~ 90 GeV, the quark and antiquark it decays into will split this energy. As such, if the quarks each have 45 GeV, and the Z^0 is then boosted to the minimum $p_\perp$ of 200 GeV, illustrated in fig. 6, one will end up with a minimum angle $\theta_{\text{min}} \approx 0.85$ rad. This minimum angle will change depending on the $p_\perp$ that the Z^0 has; larger $p_\perp$ yield smaller minimum angles.

If the orientation of the decay relative to the boost direction instead is not symmetrical, then one will end up with an angle larger than the minimum angle. If, for instance, the quark has 3/4 of the Z^0's $p_\perp$ and the antiquark has the rest, the angle between them will increase and the resulting jet will be widened. In this example, the quark will have an angle of ≈ 0.29 rad and the antiquark will have an angle of ≈ 0.73 rad, relative to the Z^0, resulting in a jet width of ≈ 1 rad. A jet created from two coloured particles will always have a minimum angle between the two subjets, and, therefore, a minimum jet width. The width of the jet will increase depending on how asymmetrical the orientation of the decay is relative to the boost direction. For a jet with $p_\perp \geq 200$ GeV, the subjets created by the quarks are expected to be within a region of

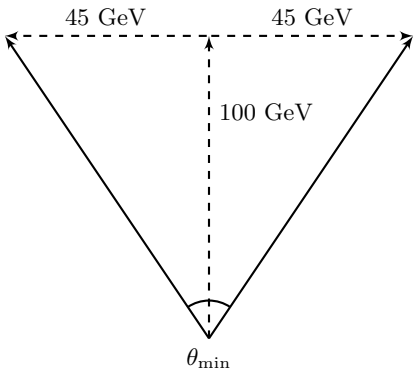


Figure 6: Schematic illustration of the minimum angle between two quarks, and the jets they create, with $\theta_{\min} \approx 0.85$ rad in this configuration.

$R = 1$ of the jet centre. This is however not always true, and as such a further constraint of finding more than one subjet in the jet is added to the simulation.

In order to find the jet from the Z^0, the distance from the centre of the anti-k_t jets are compared to the true value, i.e. the value from the event record, of the Z^0 boson. If the distance $\Delta R_{Z^0,\text{jet}}$ is less than 1, the jet is deemed to be caused by the Z^0-boson and not the recoiling gluon or quark. Once it is located, a C/A algorithm is used on the constituent particles of the jet — with a very large R so that all particles are clustered in order to perform mMDT. When all of the particles have been clustered, the mMDT is used on the clustered jet with varying z_{cut} values and the two subjets from the mMDT are located.

4.2 Finding regions

To compare the mMDT and see how much information it discards, a C/A algorithm with $R_{\text{sub}} = 0.2$ and $p_{\perp\min} = 10$ GeV is used to find the "true" subjets, which, presumably, corresponds to the two quarks produced by the decay. The C/A algorithm is used due to its "non-biased" way of forming jets, by only using the distance between particles to determine whether two particles are part of the same jet, even if they have different $p_\perp$. From these subjets, different regions are defined for further examination. In fig. 7 the different regions, and the symmetry between them, are illustrated. As explained above in section 2.2, the area in between the subjets is of interest, due to it being more active, but the region next to the hard subjet is also compared to the area next to the soft subjet. The areas perpendicular to the subjets are also examined and distinguished between the ones that are close to the midpoint and the ones that are far from this midpoint.

A more precise explanation of all of the different regions is as follows: The "true" hard subjet is located at a position (y_h, ϕ_h) and the "true" soft subjet is at position (y_s, ϕ_s). The region in between them is then centred on the point $(y_c, \phi_c) = \left(\frac{y_h+y_s}{2}, \frac{\phi_h+\phi_s}{2}\right)$. The regions next to the hard subjet and the soft subjet are centred on the points $(y_{h2}, \phi_{h2}) = (2y_h - y_c, 2\phi_h - \phi_c)$ and $(y_{s2}, \phi_{s2}) = (2y_s - y_c, 2\phi_s - \phi_c)$ respectively. The perpendicular regions close to the midpoint are centred on the points $(y_c - (\phi_h - \phi_{h2}), \phi_c + (y_h - y_{h2}))$ and $(y_c - (\phi_s - \phi_{s2}), \phi_c + (y_s - y_{s2}))$. The perpendicular regions far away from the midpoint are centred on the points $(y_c - 2(\phi_h - \phi_{h2}), \phi_c + 2(y_h - y_{h2}))$ and $(y_c - 2(\phi_s - \phi_{s2}), \phi_c + 2(y_s - y_{s2}))$. Around all of these points, except for the subjets themselves, a circle with radius R_{sub} is used as the regions. The regions then all

have the same area of 0.04π.

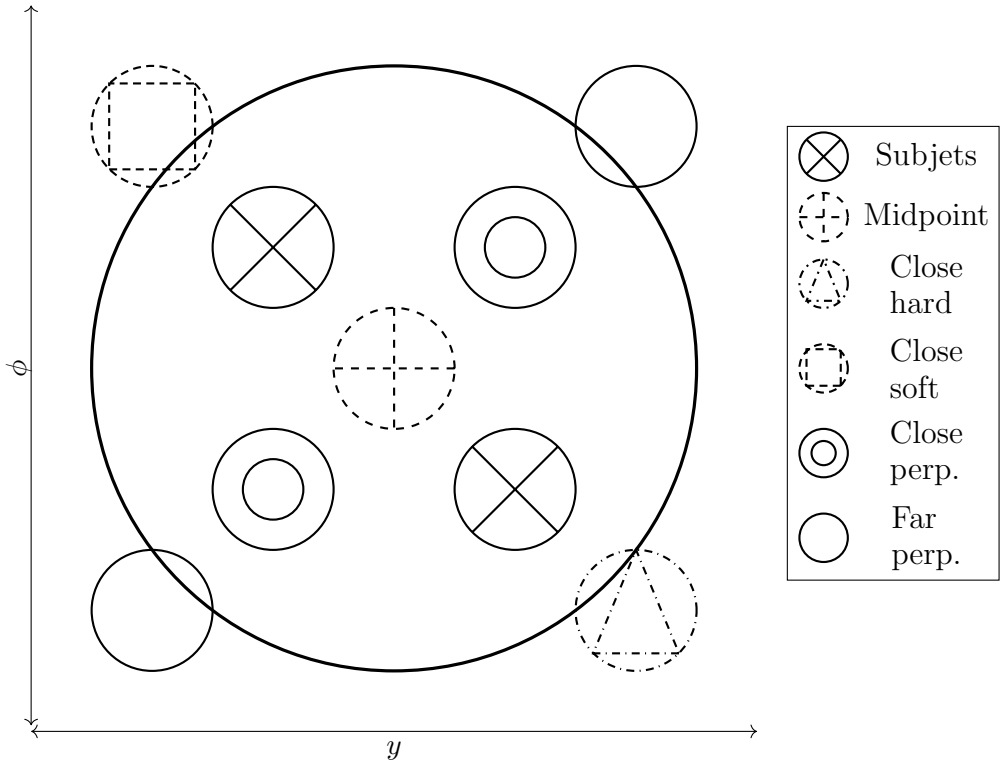


Figure 7: A schematic illustration of the jet — the largest circle — the two hardest subjets and the jet areas examined. Note that the subjets are often not located symmetrically around the centre of the jet; the hardest of the two subjets is usually located closer to the middle of the jet.

As illustrated in fig. 7, some regions may be outside or partially outside the jet and could therefore contain particles that are not considered to be part of the jet. This is taken into consideration when the results are studied. The different regions' $p_\perp$ are summed and used as a measurement of information. The discarded particles from the mMDT are checked to see if they are in the examined regions, and their $p_\perp$ is summed in order to compare how much information was lost from the mMDT.

4.2.1 Limiting the study

The different constraints placed on each event generated will impact the number of events that are not examined. The constraint of finding a jet with 200 GeV with $\Delta R^2_{\mathrm{Z}^0,\mathrm{jet}} < 1$ from a hard event with a minimum $p_\perp = 190$ GeV yields a loss of about 20%. Adding the constraints of finding more than one subjet in the jet and that the two hardest subjets are located a minimum distance of $4R_{\mathrm{sub}}$ from each other — so that there is room in between the two subjets for further studies — brings the loss up to 70-80%.

4.3 Finding the z_{cut}

In order to find the z_{cut} that yields the most accurate subjets, the z for the "true" subjets located through C/A is compared to the z for the subjets found through mMDT for approximately 2000 subjets, for different z_{cut} values. These scatter plots are shown below in fig. 8.

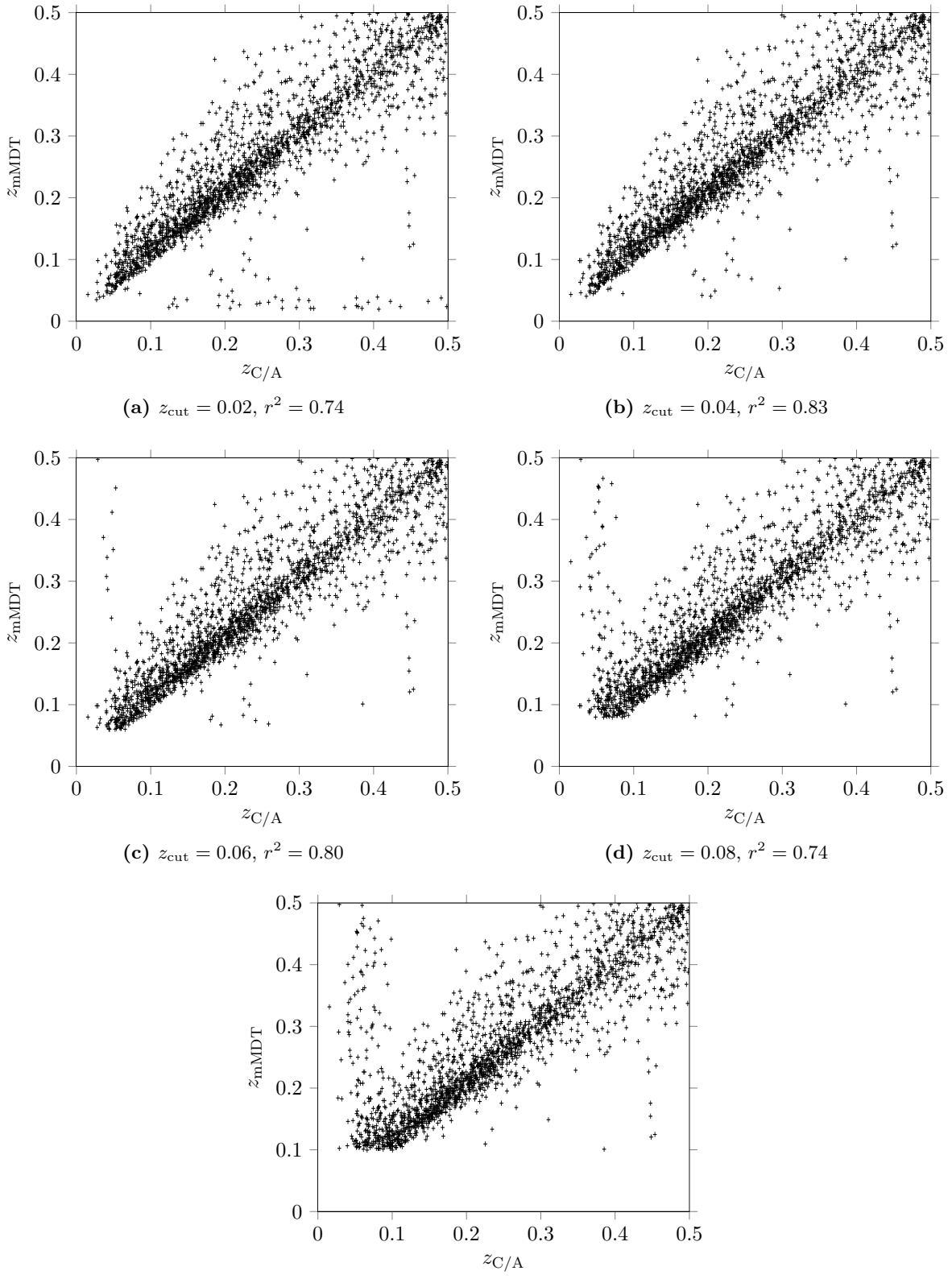
For $z_{\text{cut}} = 0.02$, subfig. (a), there are some outliers at $z_{\text{mMDT}} \approx 0.02$. What one can see is that the z for the mMDT subjet is much lower than the z for the C/A subjet. This could be because when the mMDT "declusters", i.e. goes backwards in the clustering sequence, one of the final clusterings happen between a very hard subjet and a very soft subjet. The hard subjet may contain both the "true" hard and the "true" soft subjets. Since the z_{cut} is so low, the final clusterings of very soft clusters with the hard subjets may fulfil the symmetry condition and therefore be accepted as the subjets of the jet.

Given that this behaviour disappears as the value for z_{cut} increases, the hypothesis put forth is validated further. However, for $z_{\text{cut}} = 0.06, 0.08$ and 0.10, subfigs. (c), (d) and (e), a different behaviour starts to appear. For low $z_{\text{C/A}}$, high z_{mMDT} are found. This action can be explained with the reasoning that the value of z is quite low for the "true" subjets; the distribution of the transverse momentum is uneven. The value of $z_{\text{C/A}}$ is lower than z_{cut} and the "true" soft subjet is therefore removed. This is confirmed by the fact that the number of outliers in the region $z_{\text{C/A}} \in (0, 0.1)$ increases as z_{cut} increases.

The region around $z_{\text{cut}} = 0.04$ seems to yield the most accurate subjets; the correlation r^2 is the highest for this region. The correlation does, however, differ quite substantially over the different z_{cut} values in fig. 8. The region around $z_{\text{cut}} = 0.04$ is, therefore, examined further in fig. 9, in order to see if there is a better z_{cut} value. One can see here that the correlation is at its highest for $z_{\text{cut}} \approx 0.04$, due to the fact the subfig. 9(c) and subfig. 8(b) yield a similar correlation, as well as very similar plots. Therefore the different z_{cut} in fig. 8 are sufficient to demonstrate the mMDT behaviour for different z_{cut} values.

The z-values are not the only values that should be the same for the C/A ("true") subjets and the mMDT subjets, the distance between the centre of the hard C/A subjet and the centre of the hard mMDT subjet (and similarly for the soft subjets) should be 0. The ΔR distance between the C/A subjets and the subjets located through mMDT for different z_{cut} values are plotted below in fig. 10. Here one can see that there is a gradual decrease up until around $\Delta R \approx 0.6$. There it turns around and peaks at $\Delta R \approx 0.8$. It then continues to drop at a slower pace, until it reaches zero at the end. This peak is exhibited for all the different values of z_{cut}; it is the largest for $z_{\text{cut}} = 0.10$ at $\sim 3\%$ and the smallest for $z_{\text{cut}} = 0.04, 0.06$ at $\sim 2\%$.

This peak could be because the quark and antiquark subjets do not fit in the width used for the anti-k_t algorithm, but the C/A algorithm and the mMDT still find, at least, two subjets in the jet. These could originate from the parton shower of daughter particles which is within the jet radius. However, when searching for jets with higher $p_\perp$, 400 GeV, the peak does not shift. This indicates that the peak may not be due to the two daughter particles not fitting within the jet, but instead something else. There is a possibility that a hard subjet is compared to a soft subjet, though everything is sorted in decreasing order of hardness, so it seems unlikely that that would happen. What it could otherwise be is not examined due to a lack of time.



(a) $z_{\mathrm{cut}} = 0.02$, $r^2 = 0.74$

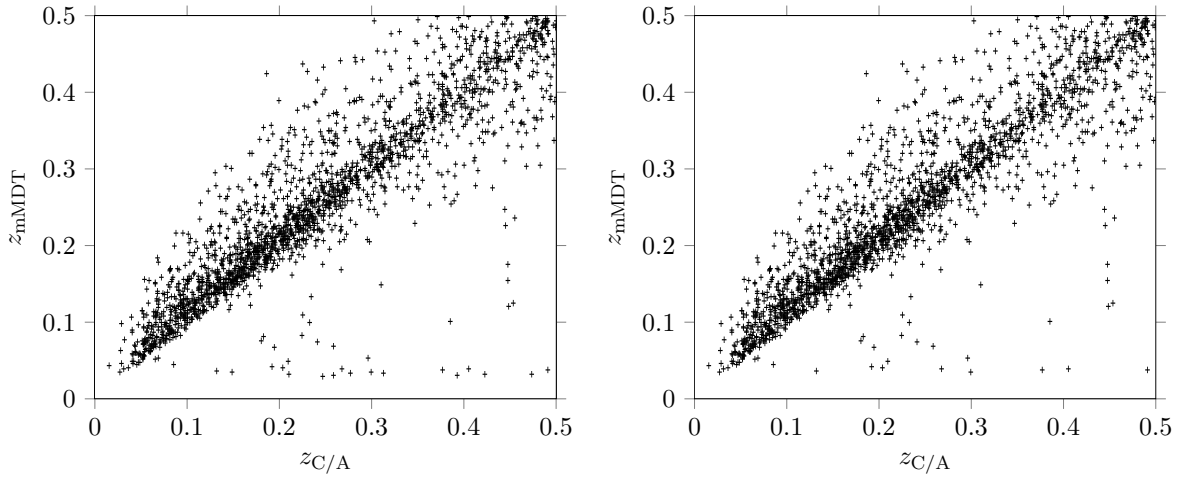
(b) $z_{\mathrm{cut}} = 0.04$, $r^2 = 0.83$

(c) $z_{\mathrm{cut}} = 0.06$, $r^2 = 0.80$

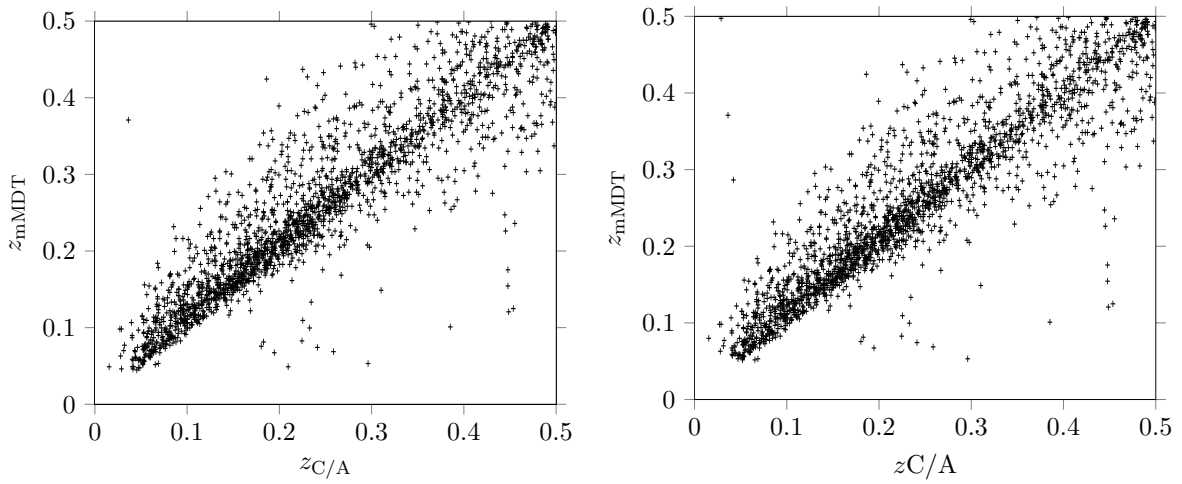
(d) $z_{\mathrm{cut}} = 0.08$, $r^2 = 0.74$

(e) $z_{\mathrm{cut}} = 0.10$, $r^2 = 0.66$

Figure 8: The z_{mMDT} plotted as a function of $z_{\mathrm{C/A}}$ for different $z_{\mathrm{cut}} \in [0.02, 0.10]$. The linear correlation r^2 is provided. Roughly 2000 subjets in 1000 jets are used.



(a) $z_{\mathrm{cut}} = 0.03$, $r^2 = 0.78$ **(b)** $z_{\mathrm{cut}} = 0.035$, $r^2 = 0.81$



(c) $z_{\mathrm{cut}} = 0.045$, $r^2 = 0.83$ **(d)** $z_{\mathrm{cut}} = 0.05$, $r^2 = 0.82$

Figure 9: The z_{mMDT} as a function of $z_{\mathrm{C/A}}$ for $z_{\mathrm{cut}} \in [0.03, 0.05]$. The linear correlation r^2 is provided. Roughly 2000 subjets in 1000 jets are used.

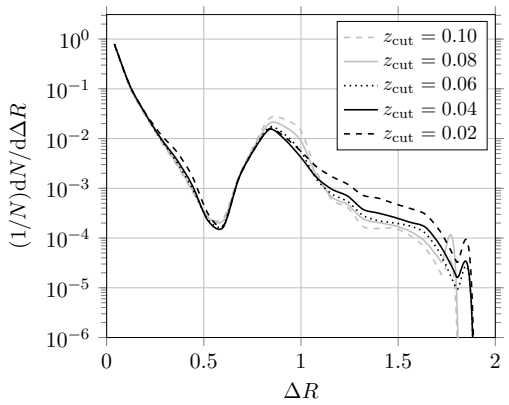


Figure 10: The distance ΔR between the hard C/A subjets and the hard mMDT subjets, and similarly for the soft subjets, for $z_{\mathrm{cut}} \in [0.02, 0.10]$ for roughly $4 \cdot 10^5$ subjets.

In table 1 the mean distance distance ΔR between the C/A ("true") and the mMDT subjets is given. Studying this, one will find that the mean ΔR is the lowest for $z_{\text{cut}} = 0.04$. This, together with fig. 8 indicate that $z_{\text{cut}} = 0.04$ may yield the most accurate subjets, compared to the "true" subjets. The difference between the different z_{cut} is, however, quite small, both in figs. 8 and 9 as well as in table 1. This indicate that a z_{cut} value of 0.04 may not actually yield the most accurate subjets. In this study however, this value is assumed to yield the most accurate subjets.

Table 1: The mean distance ΔR between C/A and mMDT for $z_{\text{cut}} \in [0.02, 0.10]$.

z_{cut}	0.02	0.04	0.06	0.08	0.10
Mean	0.089	0.080	0.084	0.094	0.109

4.4 Finding removed transverse momentum

In order to find out how much and from where the mMDT discards particles, one first needs to know how much transverse momentum is contained within the different regions, defined in fig. 7.

This is plotted below in fig. 11 with the mean $p_\perp$ and the overflow (i.e. the amount of jets with $p_\perp > 30$ GeV) listed in table 2. In these one can see that the midpoint region has the highest $p_\perp$ of all the regions, with a mean region $p_\perp$ of 4.59 GeV. This is expected, due to the substructure of the jet, as explained in section 2.2. Then the close perpendicular regions follow with a mean of 1.37 GeV, which is also expected due to symmetry. Thereafter the region next to the soft subjet continues with a mean $p_\perp$ of 0.85 GeV, and below that, the region next to the hard subjet at 0.76 GeV. This is not expected, since one would assume that the region next to the hard subjet has more activity than the one close to the soft subjet, due to it being harder. This could be because the structure shown in fig. 7 is not the structure that is most often observed. The hard subjet is usually located closer to the centre of the jet and the soft subjet is closer to the edge. The region next to the soft subjet could therefore contain final-state particles with large $p_\perp$ that are not part of the anti-k_t jet. This region could thereby have slightly higher $p_\perp$ than the region next to the hard one. Lastly the perpendicular regions that are the farthest away from the midpoint have the lowest amount of $p_\perp$ with a mean of 0.61 GeV.

The mean $p_\perp$ and overflow for the different regions of the jet is shown in table 2. Here one can also see that there is a slight overflow, with the largest one being for the midpoint region at 1.64%. This means that the activity in the midpoint region is sometimes higher than 30 GeV. This is quite a substantial amount, considering that the softest of the two subjets usually is below 100 GeV. Since all the other overflows are very low, $< 0.4\%$, the rest of the regions contain much less activity than the midpoint region.

Finally moving on to what this thesis is aiming at, in fig. 12 the removed transverse momenta (the summed $p_\perp$ of all the removed particles) for the different regions is plotted and in table 3 the mean $p_\perp$ and the overflow (the amount of jets with removed $p_\perp > 5$ GeV) for the respective plots is shown. Here one can see that for all the studied z_{cut} values, the summed $p_\perp$ for the particles removed by the mMDT from the different regions, including the subjets discovered through C/A, is very low, below 10%. This means that most of the subjets located through the mMDT are the last two to be clustered together.

However, focusing on the particles that are discarded by the mMDT, one can see that for $z_{\text{cut}} = 0.02$ and 0.04 the perpendicular regions that are located far away from the midpoint

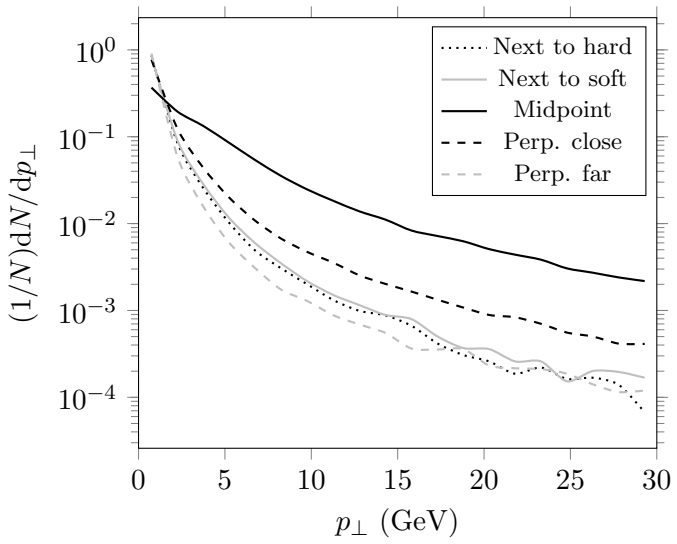


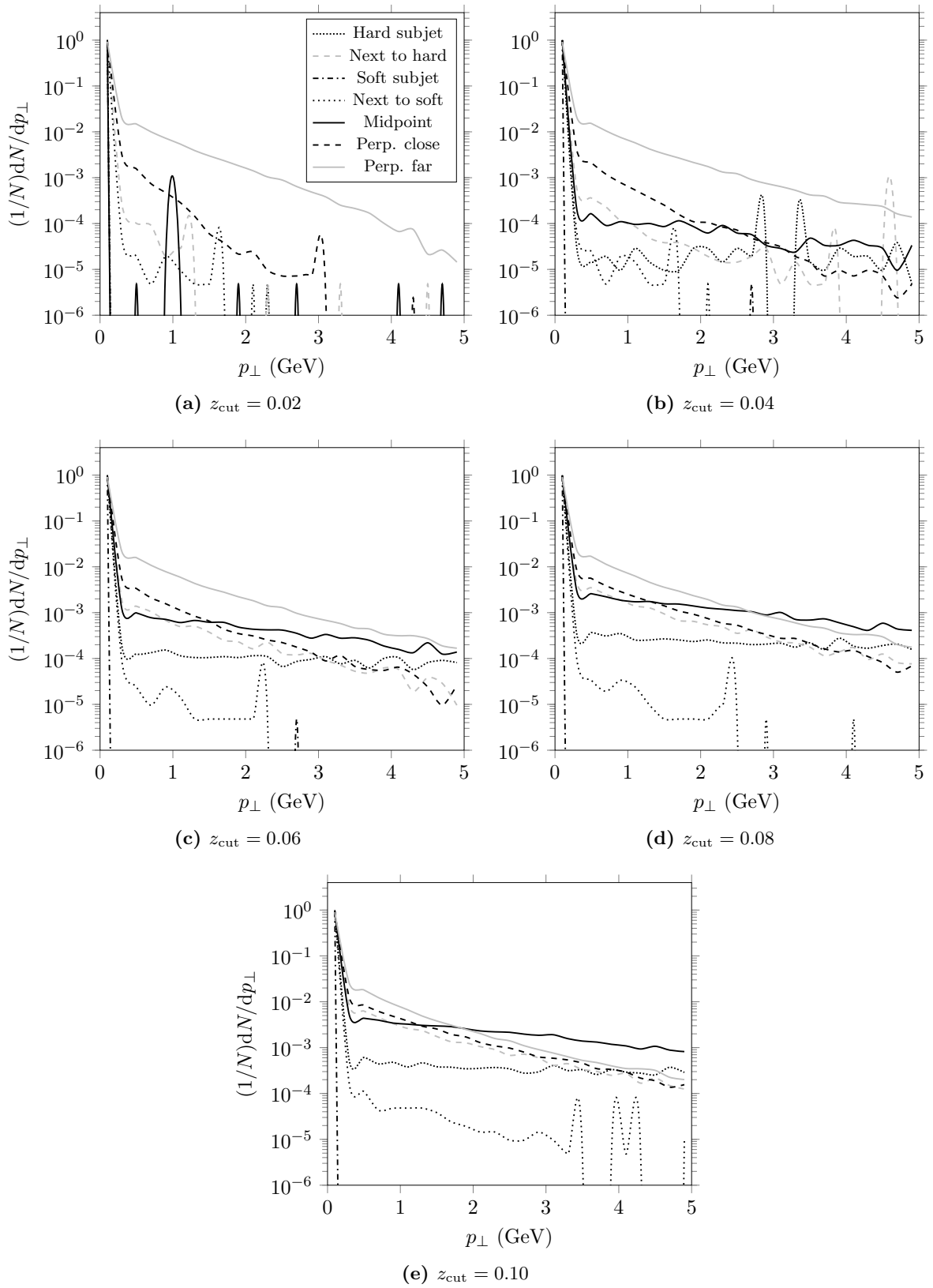
Figure 11: The $p_\perp$ for the different regions defined according to fig. 7 for approximately $2 \cdot 10^5$ jets.

Table 2: The mean $p_\perp$ calculated for the different regions, defined in fig. 7, and the overflow (the amount of jets with $p_\perp > 30$ GeV).

Region	Mean (GeV)	Overflow (%)
Midpoint	4.59	1.64
Next to hard	0.76	0.0604
Next to soft	0.85	0.112
Perp. close	1.37	0.309
Perp. far	0.61	0.127

are the main zones for discarding particles. This is good, since these regions are expected to be populated by mostly background particles, due to the symmetry of the $\mathrm{Z}^0 \to \mathrm{q\overline{q}}$ decay, and should therefore not be parts of the jet substructure. For the other z_{cut} values, the region perpendicular and far away from the midpoint seem to be no longer be the main discarding zones. The main zone for discarding particles is instead the midpoint, as can be seen in table 3. This is a problem since the midpoint region should not contain a large amount of irrelevant, or background particles, unlike the previously mentioned regions. It should instead contain hard particles caused by the parton shower and hadronisation; this can be seen in fig. 11 and table 2.

If one imagines that the background is uniformly distributed, the amount of $p_\perp$ that is discarded in the perpendicular regions far away should work as a reference; the amount that is discarded in these regions should be the amount of background, due to symmetry. As such, the other regions should never discard any more than these. If they do, then the tagging method removes information about the event and that is not sought after. If there is less $p_\perp$ removed than what is discarded in the farthest perpendicular regions, then that is also an issue. Given that the farthest perpendicular regions contain mostly background, they also indicate the amount of $p_\perp$ that should be removed from all the regions. Therefore, if there is not as much transverse momentum removed as in the perpendicular regions far away, then the tagging method has not removed all of the background. This is not as bad as removing too much, but it is still not good.



(a) $z_{\text{cut}} = 0.02$ **(b)** $z_{\text{cut}} = 0.04$

(c) $z_{\text{cut}} = 0.06$ **(d)** $z_{\text{cut}} = 0.08$

(e) $z_{\text{cut}} = 0.10$

Figure 12: The removed $p_\perp$ from the different regions specified in fig. 7 from $2 \cdot 10^5$ jets for $z_{\text{cut}} \in [0.02, 0.10]$.

However, the background is most likely not uniform, so one can not explicitly use this as a way to measure how good the method is. One can, however, use it to indicate how well it performs. Studying table 3, one can see that the amount removed from the perpendicular regions far away increases with the z_{cut}. This increase is small compared to the increase of the other regions, though. One can therefore say that the farthest perpendicular regions indicate the amount of background.

What is then seen, is that for $z_{\text{cut}} = 0.02, 0.04$ the amount of discarded $p_\perp$ in the other regions is smaller than what the supposed background $p_\perp$ is. For the other z_{cut} values, on the other hand, the amount of discarded transverse momentum is larger than the background. This would indicate that the mMDT does not operate perfectly, by discarding information about the event.

Moving on to the perpendicular regions close to the midpoint, one can see in fig. 12 that they are also major discard zones, although if one also studies table 3 there are not that many hard particles discarded; both the mean and the overflow is quite small compared to the other regions. This is quite good, given that the soft particles in this region are most likely background particles, while the harder ones are more likely to be due to the parton shower from the decay products. However, the regions discards less than the perpendicular regions farthest from the midpoint, which is not good.

There is also an interesting difference between the region next to the hard subjet and the region next to the soft subjet. The particles in the region next to the soft subjet are barely discarded at all. In the region next to the hard subjet there is a low amount of particles discarded, the second smallest next to the previously mentioned region. One would expect that the particles in the region next to the soft subjet would be discarded at a higher rate than those in the region next to the hard subjet, given that there should be more soft background in it. However, as it was revealed in fig. 11 and table 2, the expected effect of these regions are the opposite of what one finds. This could be because of the same reason, that the structure of the jet is not like fig. 7, but instead one where the hardest subjet is located closer to centre of the jet and the softer subjet is closer to the edge. This can cause the region next to the soft subjet to contain particles which are not part of the jet. Since the mMDT only applies to the particles within the jet, the region next to the soft subjet may be ignored by it. Thus the mMDT may discard more particles from the area next to the hard subjet than the area next to the soft subjet.

Lastly, there exists an interesting dynamic between the regions of the hard and soft subjet found through C/A. As is clear in fig. 12, soft particles are discarded from the hard region and the amount increases with z_{cut}, in the domain $p_\perp \in [0, 5]$ GeV, while no particles are discarded from the soft region in this domain. First of all, the fact that there are any discarded particles in these regions is not good, since they are supposed to be the subjet regions of the jet caused by the partonic decay of the Z^0. Secondly, studying table 3 one can see that the overflow for the hard and soft regions are the highest out of all the regions, growing all the way up to 6.7% and 8.4% respectively in subtable (e).

However, the mean $p_\perp$ that is discarded is much lower for the soft subjet than for the hard subjet. From $z_{\text{cut}} = 0.06$ and onwards, the region of the hard subjet is the second largest zone for discarding particles. This is a major issue and is most likely caused by the fact that when the subjets have a very low z, the mMDT throws the softer one away, due to it being below z_{cut}. Therefore, the mMDT goes on to "decluster" the hard subjet until it finds a z larger than z_{cut} In this process, both the particles of the soft subjet are discarded and particles that are part of the hard subjet are also discarded. This explanation is strengthened by the fact that the mean

$p_\perp$ and overflow in table 3 increases so drastically for the increasing value of z_{cut}.

Table 3: Tables of the mean $p_\perp$ values and the overflows (the amount of jets with removed $p_\perp > 5$ GeV) for different z_{cut}.

(a) $z_{\text{cut}} = 0.02$

Region	Mean (GeV)	Overflow (‰)
Hard	0.001	0.029
Next to hard	0.000	0.000
Soft	0.000	0.053
Next to soft	0.000	0.000
Midpoint	0.001	0.000
Perp. close	0.004	0.002
Perp. far	0.068	0.050

(b) $z_{\text{cut}} = 0.04$

Region	Mean (GeV)	Overflow (‰)
Hard	0.033	1.69
Next to hard	0.002	0.034
Soft	0.004	2.64
Next to soft	0.000	0.000
Midpoint	0.038	0.153
Perp. close	0.008	0.031
Perp. far	0.088	0.867

(c) $z_{\text{cut}} = 0.06$

Region	Mean (GeV)	Overflow (‰)
Hard	0.239	12.1
Next to hard	0.011	0.254
Soft	0.030	17.1
Next to soft	0.000	0.005
Midpoint	0.296	1.49
Perp. close	0.018	0.242
Perp. far	0.099	1.71

(d) $z_{\text{cut}} = 0.08$

Region	Mean (GeV)	Overflow (‰)
Hard	0.706	34.1
Next to hard	0.031	0.858
Soft	0.088	45.3
Next to soft	0.000	0.010
Midpoint	0.963	4.43
Perp. close	0.039	0.678
Perp. far	0.111	2.34

(e) $z_{\text{cut}} = 0.10$

Region	Mean (GeV)	Overflow (‰)
Hard	1.478	66.6
Next to hard	0.059	1.69
Soft	0.180	83.5
Next to soft	0.001	0.024
Midpoint	2.133	9.57
Perp. close	0.073	1.61
Perp. far	0.124	2.94

5 Conclusion

To summarise the discussion above, for this set-up, the most optimal value of z_{cut} for the mMDT is 0.04, as this yields the most accurate subjets. Nonetheless, even if this value yields the most accurate subjets, fig. 12 and table 3 reveal that the mMDT still discards particles from regions of high activity and not only background particles. This could mean that the mMDT does throw away information about the event. Adding to that, the value that is often chosen for the mMDT is around 0.10. This selection could result in an even larger loss of information, according to fig. 12(e).

However, in this study only one type of jet is studied and the Z^0 bosons that are created can have a very broad spectrum of energy; there are no limits on the energy. The Z^0 bosons are only limited to having a minimum $p_\perp$ of 190 GeV. This can result in the Z^0 having a very high energy, ~ 1 TeV, and then decay into a quark and antiquark with a very disproportionate split

of the energy. This type of event could cause the mMDT to yield incorrect results, due to the extreme nature of it. This is an extremely unlikely event, though, and is very rare in particle colliders and it should therefore not pose any major problem.

Another fact is that the regions that are examined, fig. 7, do not encompass the entire jet region. As such, particles may very well be removed from regions that are not examined. Which would explain the low mean in table 3. This could also result in incorrect interpretations from fig. 12.

In order to complement the results of this thesis one could in the future be studying QCD jets, which are jets from single partons, for instance in partonic scatterings. One could also compare $Z^0 \to q\overline{q}$ jets and QCD jets and see what differences the mMDT method yields for each case. A follow-up for this study could involve comparing Z^0 with different $p_\perp$. One could also expand the methods examined and compare different tagging methods to see if they throw away similar amounts or if they result in completely different subjets. One could also study the role of colour flow in the formation of jets by studying the pattern of radiation as a function of it.

References

[1] M. Thomson, "Modern particle physics," 1st ed. Cambridge: University of Cambridge; 2013.

[2] T. Sjöstrand *et al.*, "An Introduction to PYTHIA 8.2," Comput. Phys. Commun. **191** (2015) 159 [arXiv:1410.3012 [hep-ph]].

[3] M. Cacciari, G. P. Salam and G. Soyez, "FastJet User Manual," Eur. Phys. J. C **72** (2012) 1896 [arXiv:1111.6097 [hep-ph]].

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

[5] G. P. Salam, "Towards Jetography," Eur. Phys. J. C **67** (2010) 637 [arXiv:0906.1833 [hep-ph]].

[6] 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 [hep-ph]].

[7] M. Dasgupta, A. Fregoso, S. Marzani and G. P. Salam, "Towards an understanding of jet substructure," JHEP **1309** (2013) 029 [arXiv:1307.0007 [hep-ph]].