Alma Mater Studiorum · Università di Bologna

Scuola di Scienze
Dipartimento di Fisica e Astronomia
Corso di Laurea in Fisica

Study of jets produced at LHC and measurements of their substructures

Relatore:
Prof. Alberto Cervelli

Presentata da:
Nicola Forti

Anno Accademico 2020/2021

Abstract

Jets are produced by the hadronization shower of partons produced in high energy hard scattering, and they are fundamental observables to understand the physics processes at high energy colliders. In the past few years the study of their internal structure has become an hot topic. Considering the large center of mass energy experienced at LHC, many decaying particles can produce almost collinear partons. The understanding of the internal jet structure allows us to comprehend the underlying interactions and the dynamics leading to the production of, in particular, large R jets. In this thesis will be firstly provided a description of what are jets and how they are reconstructed with different algorithms, such as with the anti-k_t algorithm which is one of the most used. Then will be described the jet substructures which allow us to discriminate the various type of process we are considering, for example to distinguish three-prong events from the two- or one-prong processes. The study continues with the analysis of a simulated samples of events generated with Monte Carlo in which, through a series of selection on the variable τ_{32} and D_2, has been shown the validity of these two variables that can allow to distinguish boson decays from the top ones.

Sommario

I jets sono prodotti dalla pioggia di adronizzazione di partoni prodotti negli hard scattering ad alta energia e sono osservabili fondamentali per comprendere i processi fisici nei collisori ad alta energia. Negli ultimi anni lo studio della loro struttura interna è diventato un tema scottante. Considerando il grande centro di energia di massa a LHC, molte particelle che decadono possono produrre partoni quasi collineari. La comprensione della struttura interna dei jets ci permette di comprendere le interazioni e le dinamiche che portano alla produzione, in particolare, di large-R jets. In questa tesi verrà innanzitutto fornita una descrizione di cosa sono i jet e di come vengono ricostruiti con diversi algoritmi, come per esempio con l'algoritmo anti-k_t, che è uno dei più utilizzati. Verranno poi descritte le variabili di sottostruttura dei jets che ci consentono di discriminare i vari tipi di processo che stiamo considerando, ad esempio per distinguere gli eventi a tre vertici dai processi a due od uno. Lo studio prosegue con l'analisi di un campione simulato di eventi generati con Monte Carlo in cui, attraverso una serie di selezioni sulle variabili τ_{32} e D_2, è stata dimostrata la validità di queste due variabili che possono permettere di distinguere i decadimenti bosonici da quelli del quark top.

Contents

Introduction

At present the Large hadron collider (LHC), operating at CERN, is the largest proton proton collider ever made. Its collision center of mass energy is 13 TeV, this energy, together with the high luminosity of the LCH, allows us to make precise measurement of rare SM processes and search for physics beyond the standard model.
The primary proton-proton interaction produce hadrons and particles decaying in hadrons. Such particles have high momenta due to the large CM energy, and may produce showers of large numbers of particles, reconstructed as jets, and in case of intermediate particles decaying in hadrons, the initial partons may be very close to each other due to relativistic boost. In the latter case it is not possible to resolve the shower produced by the hadronization of the two different partons, and we can rely on the study of the structure of these showers, the jet substructure variables, to understand the event of interest.
In the first chapter of this thesis it is described how a jet can be defined and how the jet signals are obtained in the LHC, with the use of the ATLAS detector and its calorimeters. In this part it is also shown the way in which the jets constituents are constructed using calorimeter information and the description of the jet algorithms which are fundamentals in the reconstruction of the jets.
The second chapter contains the description of the jet substructures, that are fundamental in the characterization of the internal structure of the jets, and of the jet taggers which are used to discriminate the different types of particles from where jets are originated.
In the third chapter is contained the analysis of the thesis, where is described how the simulated data have been generated and how they have been elaborated to study the different processes considered. Moreover in this section is shown what happened if some constraints are imposed to the data and what conclusions we can achieve.

Chapter 1

Jets and jet algorithms

1.1 Hadron collider kinematics

If consider the LHC case, a proton-proton collider, the interactions are indeed parton-parton interactions. The two colliding partons carry respectively a fraction x_1 and x_2 of the initial proton's momentum, as x_1 and x_2 are different, the centre of mass of the hard interaction is longitudinally boosted compared to the lab frame. Since there is no balance in the longitudinal axis of the event, the best choice is to consider the transverse plane for reconstructing the four-momenta of the particles. Instead of using energy and polar angles, transverse momentum x_t, rapidity y and azimuthal angle ϕ are used [2]. Considering a four-vector (E, p_x, p_y, p_z), p_t and ϕ are defined as the modulus and azimuthal angle in the transverse plane (p_x; p_y), i.e. we have

$$p_t = \sqrt{p_x^2 + p_y^2}\,,$$

while rapidity is defined as

$$y = \frac{1}{2}\log\Big(\frac{E + p_z}{E - p_z}\Big).$$

A four-vector of mass m, therefore can be represented as

$$p^\mu \equiv (m_t \cosh y, p_t \cos\phi, p_t \sin\phi, m_t \sinh y),$$

with $m_t = \sqrt{p_t^2 + m^2}$ often referred as transverse mass.The distance between two particles in the (y - ϕ) plane can be calculated with the following equation:

$$\Delta R_{12} = \sqrt{\Delta y_{12}^2 + \Delta\phi_{12}^2}.$$

In an experimental context, it is often used the *pseudo-rapidity* η instead of rapidity. The former is directly defined either in terms of the magnitude $|\vec{p}|$ of the 3-momentum,

or in terms of the polar angle θ between the direction of the particle and the beam:

$$\eta = \frac{1}{2}\log\left(\frac{|\vec{p}| + p_z}{|\vec{p}| - p_z}\right) = -\log\left(\tan\frac{\theta}{2}\right) = \operatorname{arctanh}\left(\frac{p_z}{|\vec{p}|}\right).$$

Contrary to rapidity differences, the pseudo-rapidity ones are generally not invariant under longitudinal boosts. On the other hand for massless particles $y = \eta$, so it holds in case of high energy partons. By the definition of η follows that if the particle has a direction perpendicular to the beam, the pseudo-rapidity will be equal to zero, while it will tend toward infinity when aligned with the z axis, as shown in the figure 1.1.

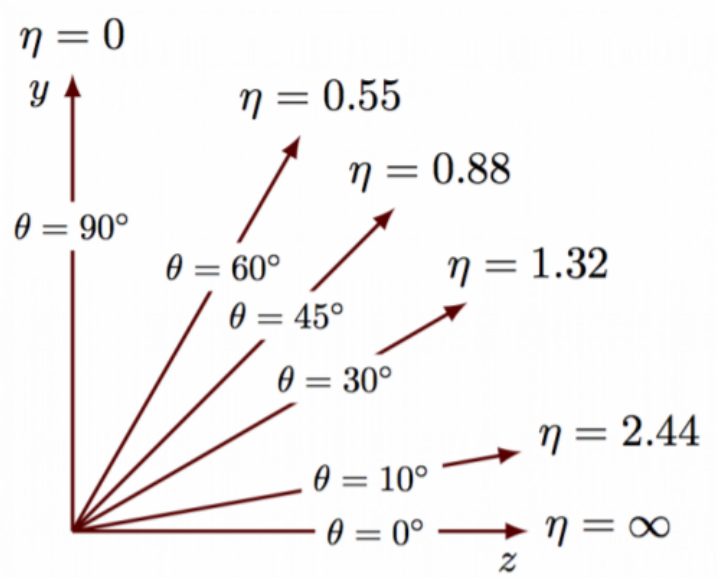


Figure 1.1: Example of some values of pseudo-rapidity η and his corresponding θ angle.

1.2 Jet definition

In high energy collisions, such as the ones occurring at LHC, hard QCD partons may be produced. These high-energy quarks and gluons are not observed directly in the final states due to the QCD confinements, as it prevents the existence of isolated color-charged particles. All partons go through an hadronization process, radiating new partons and forming colorless hadrons, resulting in a collimated spray of particles which in the reconstruction of the events we define as jets.
While the parton definition is well defined in the theory, the same cannot be said for jets, as they are a product of a particular way of reconstructing the final state of the collision. So we need to give a definition for the jets. The decision of reconstructing two, or more, different particles as a single jet has some degree of arbitrariness. A jet definition can be seen as made of a few essential parts: the *jet algorithm* and a set of parameters in the algorithm. A characteristic parameter, which is used in almost every jet definitions used in hadron colliders, is the *jet radius*. This parameter determines the limit distance in the polar and azimuthal above which two particles should not be considered collinear and hence no longer reconstructed as a single jet.
The other ingredient for the Jet reconstruction is the definition of the magnitude of its momentum, and its momentum direction. To do so we make use of what are called recombination scheme. One of the most used recombination scheme is the *"E-scheme"*, which simply sums the components of the four-vector of the particles considered. Multiple jet substructure applications use the *winner-take-all* (WTA) [7] recombination scheme which guarantees that the jet direction will coincide with one of the reconstructed particles inside the jet. In the context of a pairwise clustering algorithm, the recombination scheme determines how two pseudo-jets p_1 and p_2 will be merged to form a new pseudo-jet p_r. In the WTA scheme, the transverse momentum of p_r is obtained by the sum of the two pseudo-jets, while the direction of p_r is determined by the hardest one:

$$p_{T_r} = p_{T_1} + p_{T_2} \quad \hat{n}_r = \begin{cases} \hat{n}_1 & \textbf{if } \ p_{T_1} > p_{T_2}, \\ \hat{n}_2 & \textbf{if } \ p_{T_2} > p_{T_1}. \end{cases}$$

Infra-red and collinear safety (IRC) is a property for which if you modify any event with the addition of a soft emission or by splitting, for example, a parton into two collinear parton, the observable V, that has been considered initially, must satisfy the following features:

- collinear safety: $V_{m+1}(\dots, k_i, k_j, \dots) \to V_{m+1}(\dots, k_i + k_j, \dots)$ if $k_i \parallel k_j$,
- infra-red safety: $V_{m+1}(\dots, k_i, \dots) \to V_{m+1}(\dots, k_{i-1}, k_{i+1}, \dots)$ if $k_i \parallel k_j$.

IRC safety [10] is important for many reasons, for example during the fragmentation processes of an hard parton, since it will go through many collinear splittings, or when

there is soft emission of soft particle in QCD events. Another important case is when experimental detectors are involved, in fact they provide some regularisation of IRC unsafety, but this depends on the combination of tracking or calorimeters that have been used. Therefore this can complicate the connection between the experimental results obtained with an IRC unsafe algorithms and the expectation at hadron-level. If utilized with an infrared/collinear safe clustering measure, such as jet algorithms explained in the section 1.5, the winner-take-all scheme is also IRC safe. Since the direction of the jet in the winner-take-all scheme is aligned along one of the input particles, which is often the hardest one, the final set of recombined jet directions is much smaller than the number of hadrons in the final state.

1.3 The ATLAS detector

The Large Hadron Collider (LHC) is a proton-proton particle collider located at CERN in Geneva with a radius of 27 km at the a mean depth of 100 meters [2]. The particle beams in the LHC are allowed to interact at four sites which are respectively two multi-purpose detectors (ATLAS and CMS) and two specialised experiment (ALICE and LHC-b). The ATLAS detector is a particle detector designed for high precision measurements with a lenght of 44 meters, a diameter of 25 meters and a weight of around 7000 tons.
In the center of the detector is located the proton beam interaction point (IP) with a cylinder axis parallel to the beam pipe around which there are several layers of concentric detector cylinders. The particles, after passing through the IP, pass through the Inner Detector, which has the function of measuring the direction and momentum of the charged particles. The calorimeters are placed at larger radius than the inner tracker, and are used to absorb and measure the energy of photons and electrons (in the EM calorimeter), and hadrons (in the hadronic calorimeters). The muons will escape the calorimeters and their trajectories are measured again in the muon spectrometer. Figure 1.2 shows a schematic model of the ATLAS detector.

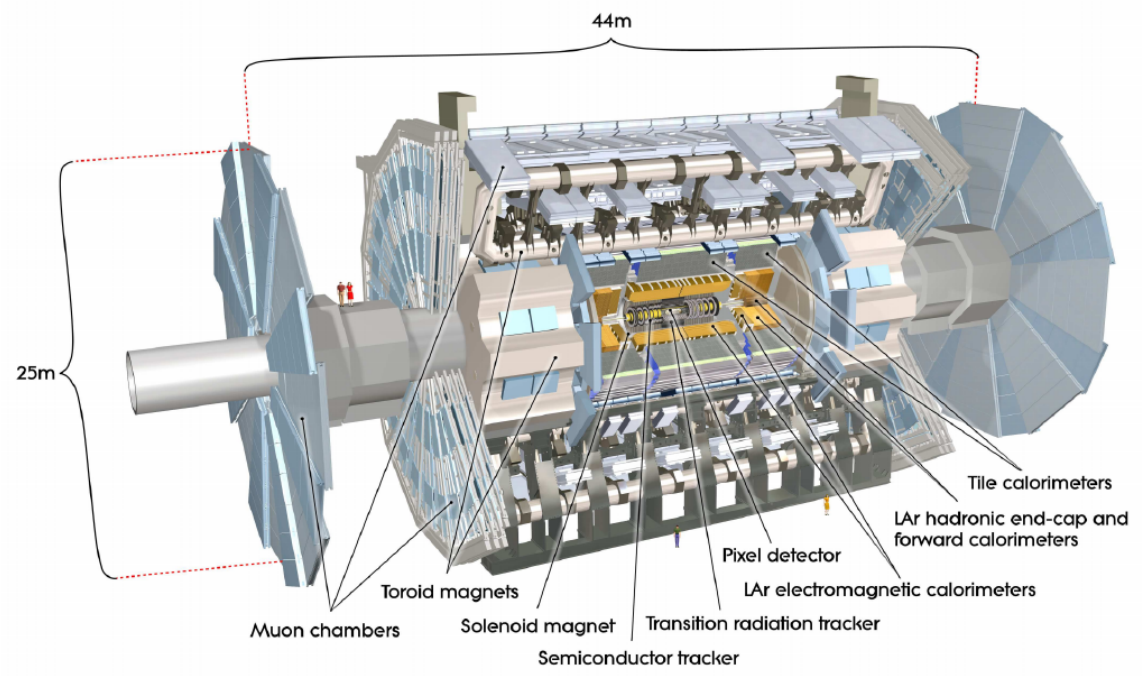


Figure 1.2: A model of the ATLAS detector which show all the sub-detector.

1.3.1 Calorimeters

The purpose of the calorimeters is to measure the energy of particles (except for the muons and neutrinos) by absorbing the particle showers produced in their material. The ATLAS calorimeter system is composed by an electromagnetic calorimeter nested inside of the hadronic one. This is due to the fact that electrons and photons interact electromagnetically with matter, which is different from the hadronic interaction of hadrons and therefore two detector are necessary.

Electromagnetic calorimeter

The electromagnetic calorimeter (EM) consists of a barrel section which covers a range equals to $|\eta| < 1.475$ and two end-caps components ($1.375 < |\eta| < 3.2$), where the active material used is liquid argon (LAr) while the absorption material is the lead. The granularity of the calorimeter changes with the value of the pseudorapidity $|\eta|$, i.e. the resolution of the end-cap region is $\Delta\eta \times \Delta\phi = 0.1 \times 0.1$, while in the central region is achieved a value of $\Delta\eta \times \Delta\phi = 0.25 \times 0.25$. In this region the calorimeter is made of three channels with varying thickness with the innermost layer having a finer granularity compared to the outer layer as you can see in the figure 1.3.

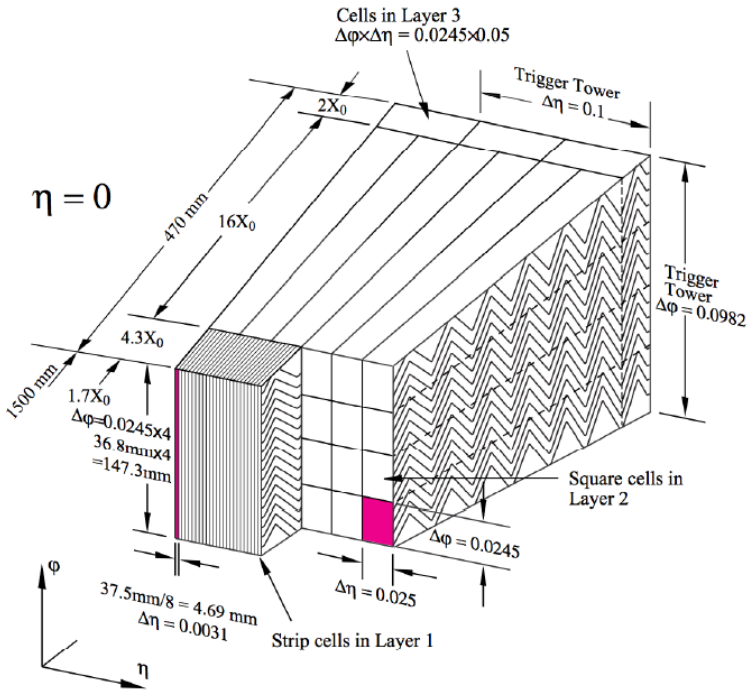


Figure 1.3: Sketch of a barrel module where are visible the different layers. The granularity of the cells of each of the three layers is also shown.

Hadronic Calorimeter

The hadronic calorimeter (HCAL) has the purpose to absorb the particles that pass through the electromagnetic calorimeter and that interact via the strong force. The HCAL can be divided into the following three parts:

- **Tile Calorimeter**. The tile calorimeter is a hadronic calorimeter placed outside the EM calorimeter which as iron as absorbing material and scintillating tiles as active material. It consists of a tile barrel in the central region which covers a range of $|\eta| < 1.0$ and has a granularity of $\Delta\eta \times \Delta\phi = 0.1 \times 0.1$ and an extended barrel for $0.8 < |\eta| < 1.7$ with the same resolution. This barrels are divided azhimuthally into 64 modules.

- **LAr hadronic end-cap calorimeter**. This calorimeter (HEC) consists of two independent wheels per end-cap which is built from 32 identical modules. Each of these wheels is formed by two segments in depth, for a total of four layers. The HEC covers $1.5 < |\eta| < 3.2$ with a granularity of $\Delta\eta \times \Delta\phi = 0.2 \times 0.2$. Here the active material is the liquid argon like in the EM calorimeter while the passive material is the lead.

- **LAr forward calorimeter**. The FCal is both an electromagnetic and hadronic calorimeter with a depth of 10 interaction lengths and consists of 3 modules in each end-cap; the first is for electromagnetic measurements and has copper as passive material, while the other two are for hadronic interactions and have tungsten as passive material. The active material, as the name says, is the liquid argon and the calorimeter covers a range between $3.1 < |\eta| < 4.9$.

The calorimeter system is the most essential sub-detector for jet physics since, without the calorimeters, there would not be any signals that could be use by the jet algorithm to reconstruct jet's history.

1.4 Topological cluster cells at ATLAS

The experimental input to the jet algorithms is reconstructed from energy deposits of particles within the different detector components. The method of the reconstruction differs between each of the four LHC experiments. In this section will be explained how calorimetric information is used by ATLAS to construct jet constituents.
The signal extraction is guided by reconstructing three-dimensional "energy zones" from particle showers in the calorimeter volume. Depending on the incoming particle types, energies, spatial separations and cell signal formation, individual topo-clusters represent the response, or part of it, to a single particle, the merged response of several particles, or a combination of merged full and partial showers.

1.4.1 Topo-cluster formation

The collection of calorimeter cell signals into topo-cluster has spatial signal-significance patterns determined by particle showers. The basic observable related to this formation is the cell signal significance $\zeta^{\mathrm{EM}}_{\mathrm{cell}}$, that is defined as the ratio of the cell signal to the average noise $\sigma^{\mathrm{em}}_{\mathrm{noise,cell}}$ in this cell.

$$\zeta^{\mathrm{EM}}_{\mathrm{cell}} = \frac{E^{\mathrm{EM}}_{\mathrm{cell}}}{\sigma^{\mathrm{EM}}_{\mathrm{noise,cell}}} \tag{1.1}$$

Either the cell signal $\sigma^{\mathrm{EM}}_{\mathrm{noise,cell}}$ or $\sigma^{\mathrm{EM}}_{\mathrm{noise,cell}}$ are evaluated on the EM energy scale, which reconstructs the energy deposited by electrons and photons correctly but does not include any corrections in case of loss of signal for hadrons due to the non-compensating characteristic of the ATLAS calorimeters.
Topo-clusters are composed by a growing-volume algorithm which start from a calorimeter cell with a highly significant seed signal.In this algorithm the seeding, growth and boundary features of topo-clusters are controlled by the three parameters (S,N,P), that

define the signal thresholds in terms of $\sigma^{\mathrm{EM}}_{\mathrm{noise,cell}}$ and so apply the previous selections from Eq. (1.1)

$$|E^{\mathrm{EM}}_{\mathrm{cell}}| > S\sigma^{\mathrm{EM}}_{\mathrm{noise,cell}} \Rightarrow |\zeta^{\mathrm{EM}}_{\mathrm{cell}}| > S \quad \text{(primary seed threshold, default } S = 4); \tag{1.2}$$

$$|E^{\mathrm{EM}}_{\mathrm{cell}}| > N\sigma^{\mathrm{EM}}_{\mathrm{noise,cell}} \Rightarrow |\zeta^{\mathrm{EM}}_{\mathrm{cell}}| > N \quad \text{(threshold for growth control, default } N = 2); \tag{1.3}$$

$$|E^{\mathrm{EM}}_{\mathrm{cell}}| > P\sigma^{\mathrm{EM}}_{\mathrm{noise,cell}} \Rightarrow |\zeta^{\mathrm{EM}}_{\mathrm{cell}}| > P \quad \text{(principal cell filter, default } P = 0). \tag{1.4}$$

1.4.2 Topo-clusters collecting algorithm

Topo-cluster formation is a sequence of seed and collect steps, which are repeated until all cells pass the criteria given in Eqs. (1.2) and (1.3) and their direct neighbours satisfy the condition in Eq. (1.4). The algorithm starts by selecting all cells, which later will be ordered in decreasing $\zeta^{\mathrm{EM}}_{\mathrm{cell}}$, with signal significances $\zeta^{\mathrm{EM}}_{\mathrm{cell}}$ passing the threshold defined by S in Eq. (1.2) from calorimeter regions that are allowed to seed the clusters.
Therefore each seed cell forms a *proto-cluster* and the cells neighbouring a seed which also satisfy Eq. (1.3) or Eq. (1.4) are collected into the corresponding proto-cluster. In this procedure Two cells are considered neighboring if they are directly adjacent in a given sampling layer, or if in adjacent layers with at least a partial overlap in the (η - ϕ) plane. This means that the cell collection for topo-clusters can include modules within the same calorimeter as well as calorimeter sub-detector transition regions. If a neighbouring cell has a signal significance that exceeds the threshold defined by the parameter N in Eq. (1.3), its neighbours are also collected into the proto-cluster. The proto-clusters are merged if a particular neighbour is a seed cell, which pass the threshold S defined in Eq. (1.2) or if a neighbouring cell is attached to two different proto-clusters and its signal significance is above the threshold defined by N. This procedure is applied to further neighbours until the last set of neighbouring cells with significances which exceed the threshold defined by P in Eq. (1.4), but not the one in Eq. (1.3), is collected. When this point is reached in the algorithm, the formation stops.

The final proto-cluster is characterised by a core of cells with highly significant signals, surrounded by an envelope of cells with less significant signals. The configuration optimised for ATLAS hadronic final-state reconstruction uses $S = 4$, $N = 2$, and $P = 0$, as parameters. In this particular configuration with $P = 0$, any cell neighbouring a cell with signal significance that overcomes the threshold set by N in Eq. (1.3) is collected into a proto-cluster, independent of its signal. The use of correlations between the energies in adjacent cells thus allows the retention of cells with signals close to the noise levels while preserving the noise suppression feature of the algorithm. Figure 1.4 shows an example of topo-clusters generated by an MC simulated jet in the first module of the ATLAS forward calorimeter under 2010 run conditions.
An additional calibration is applied with the use of the local cell weighting (LCW) scheme to form clusters in which their energy is calibrated at the correct particle-level scale. This

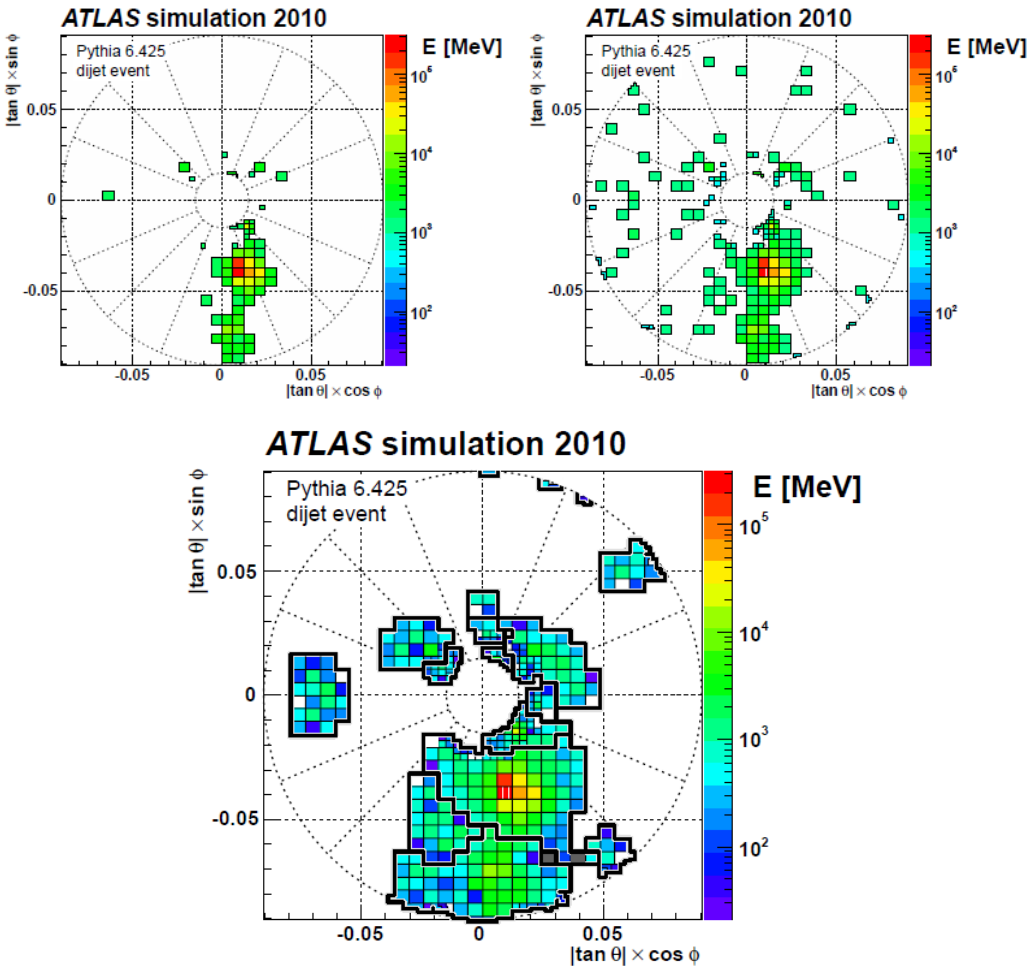


Figure 1.4: Different phases of topo-cluster formation in the FCAL calorimeter for a simulated dijet event with at least one jet entering this calorimeter. In the top-left image are shown cells with signal significance $\zeta_{\text{cell}}^{\text{EM}} > 4$ that can seed topo-clusters, in the top-right image cells with $\zeta_{\text{cell}}^{\text{EM}} > 2$ controlling the topo-cluster growth, and in the bottom one all clustered cells and the outline of topo-clusters and topo-cluster fragments in this module.

weighting scheme allows to classify the energy depositions as either electromagnetic- or hadronic-like using a variety of cluster momenta and accounts for the non-compensation of the calorimeter.
At the end of the algorithm, the angular coordinates (η and ϕ) of topoclusters are re-calculated with respect to the reconstructed primary vertex of the event, instead of the geometric centre of the ATLAS detector.

1.5 Jet algorithms

There are two different categories of jet algorithm [8]: the cone algorithms and the sequential-recombination algorithms. The latter are the most used and are based on the concept that, from a perturbative QCD viewpoint, jets are the product of successive parton branchings. Therefore the main purpose of these algorithms is to invert this process by successively recombining two particles into one.

Generalised-k_t algorithm. A large part of the recombination algorithms used in the hadronic collisions studies belong to the family of the *generalised-k_t algorithm* in which jets are clustered as follows:

1. The particles in the event are taken as the initial list of objects.
2. Two set of distances are built from the initial list of objects: an inter-particle distance
$$d_{ij} = \min(p_{t,i}^{2p}, p_{t,j}^{2p}) \frac{\Delta R_{ij}^2}{R^2} ,$$
in which p is a free parameter and ΔR_{ij} is the geometric distance in the rapidity-azimuthal angle plane, and a *beam distance*
$$d_{iB} = p_{t,i}^{2p} ,$$
3. Then iteratively find the smallest distance among all the d_{ij} and d_{iB}
   - If the smallest distance is a d_{ij} then objects i and j are cancelled from the list and recombined into a new object k (using the recombination scheme chosen) that is itself added to the list.
   - If the smallest is a d_{iB}, object i is called a jet and removed from the list.

   The algorithm proceeds going back to step 2 until all the objects in the list have been examined.

If two objects are close in the (y - ϕ) plane, like after a collinear parton splitting, the distance d_{ij} becomes small and is more possible that the two objects recombine, while when the inter-particle distances is larger than the jet radius ($\Delta R_{ij} > R$), the beam distance becomes smaller than the former and the objects are no longer recombined.

Cambridge/Aachen algorithm. A specific cases of the generalised-k_t algorithm is the Cambridge/Aachen algorithm which is obtained by setting the parameter $p = 0$. The distance evaluated in this algorithm becomes geometrical and suffers less from the contamination generated from the soft backgrounds than other algorithm, such as the k_t algorithm.

k_t **algorithm**. The best known algorithm in the generalised-k_t family is the k_t *algorithm*, in which the parameter p is set equal to 1 in the above description. In this case a soft emission would be associated with cluster close to it and therefore recombined in the early stages of the recombination process. Because of this sensitivity to soft emissions, the jets reconstructed with the k_t algorithm become more sensitive to extra soft radiation in the event, becoming inefficient in environments with high track density, and large soft emissions, such as the ones experienced in underlying (pile-up) events at LHC.

Anti-k_t algorithm. The most used jet algorithm in the context of LHC physics is the anti-k_t one, which corresponds to the generalised-k_t algorithm with $p = -1$, this choice, differently from the k_t algorithm favours the initial reclustering of hard particles. A hard jet will grow by collecting soft particles around it until the jet has a distance from the jet axes that is equal to R, which means that hard jets will be insensitive to soft radiation and have a circular shape, characteristic of this algorithm, in the (y - ϕ) plane. This feature of the anti-kt algorithm simplifies the calorimeter energy calibration, and the reconstruction of the jet axis, and has been adopted by both ATLAS and CMS collaborations as the standard clustering for their jet reconstruction. The figure 1.5 shows a step-by-step example of a clustering sequence with the anti-k_t algorithm on a set of particles.

The differences between the three algorithms are easily observed in the momentum weighting. For the k_t algorithm, the weighting $(\min(p_{t,i}^2, p_{t,j}^2))$ is done in such a way as to merge constituents with low transverse momentum with respect to their nearest neighbours, while in the anti-k_t algorithm, the weighting $(\min(1/p_{t,i}^2, 1/p_{t,j}^2))$ is conceived to merge constituents which have high transverse momentum with respect to their neighbors.
The approach of the anti-k_t algorithm is different from the k_t approach, indeed the jets clustered with the former are roughly circular in the (y-ϕ) plane. Instead the C/A algorithm relies only on distance weighting without k_t weighting. The differences between these algorithms can be appreciated in the figure 1.6. All the algorithms considered above have the property to be infra-red collinear safe.

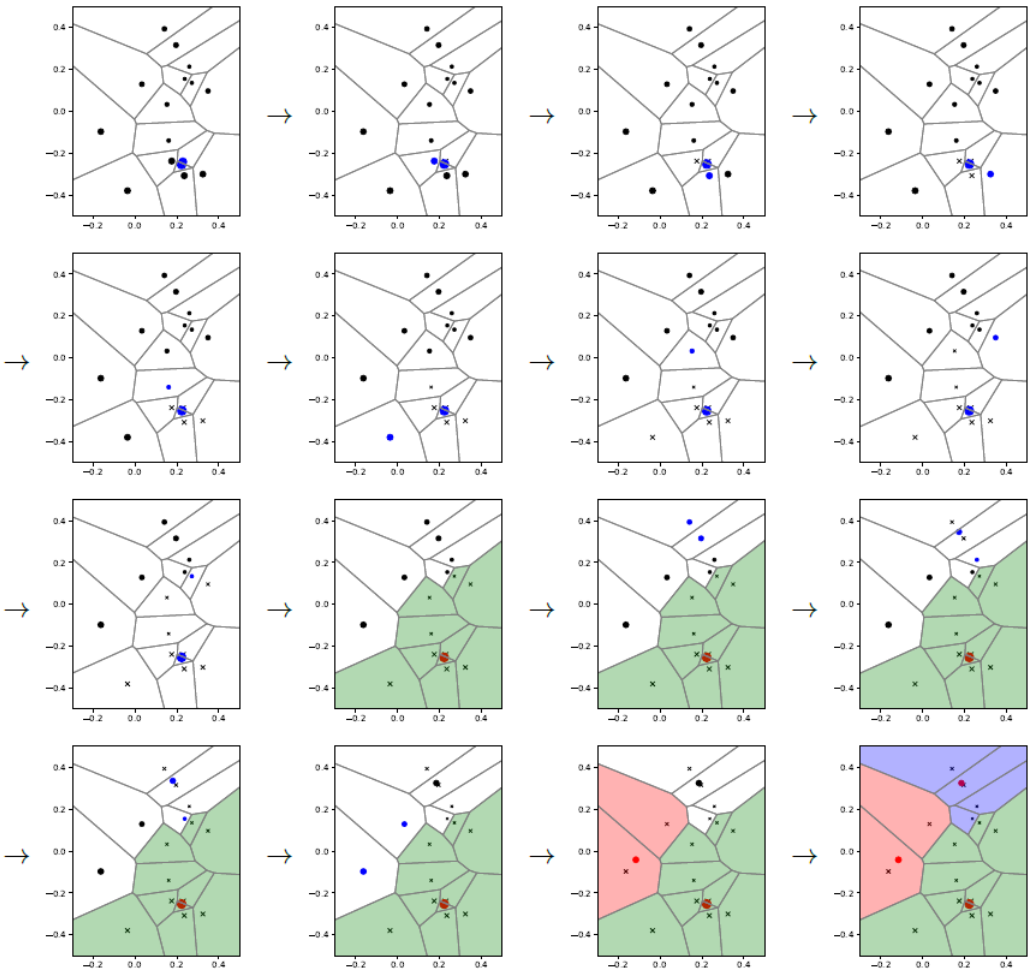


Figure 1.5: Description step-by-step of clustering using the anti-k_t algorithm with $R = 0.4$. The axes of each plot are rapidity and azimuthal angle. Each particle is represented by a cross with his size increasing with the transverse momentum of the particle, while every panel corresponds to one step of the clustering. The dots instead represent the objects that are left for clustering. Pairwise clusterings are marked with a blue pair of dots, while red dots correspond to final jets. The shaded areas show the cells included in each of the three final jets.

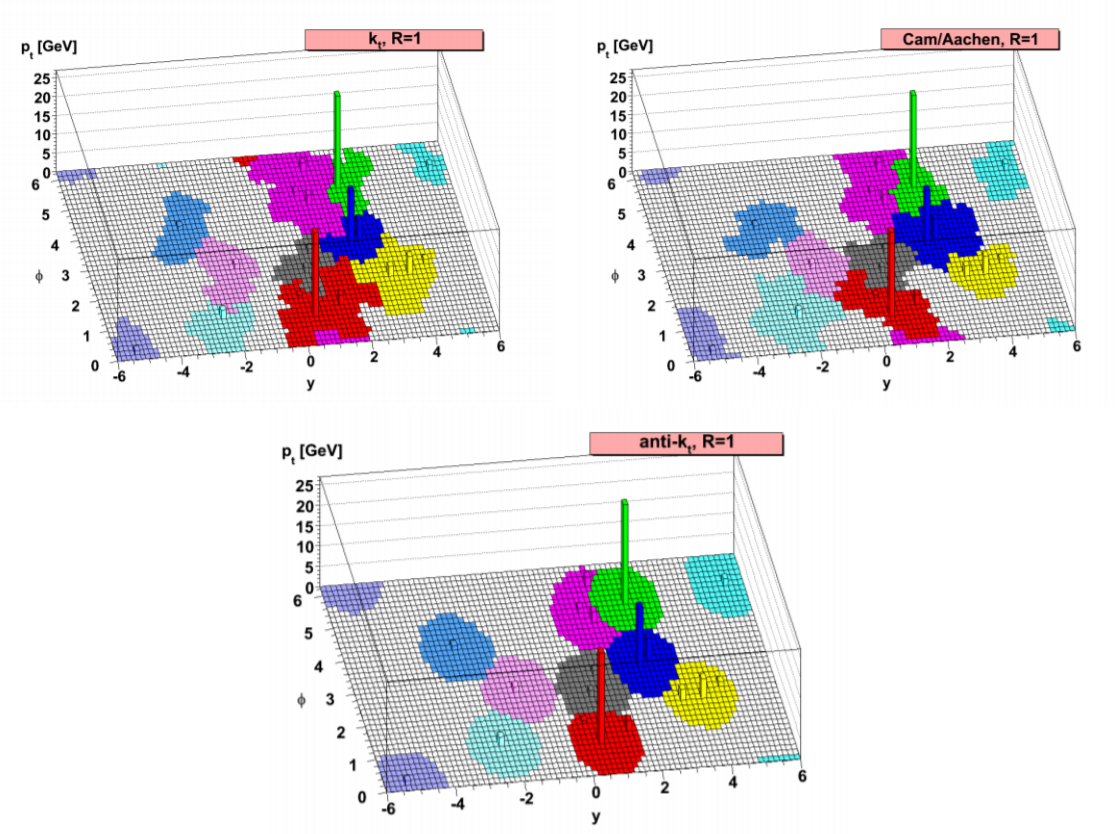


Figure 1.6: Jets obtained with the k_t (top-left), Cambridge/Aachen (top-right) and anti-k_t (bottom) algorithms with jet radius R = 1. The coloured regions correspond to the catchment area of each jet.In this picture is shown that the jets obtained with the Cambridge/Aachen and k_t algorithm have complex boundaries, while the hard jets obtained with anti-kt clustering are almost perfectly circular.

1.6 Pile-up

Pile-up are originated from simultaneous proton-proton (pp) collisions that occur in addition to a hard scattering collision of interest [6]. The hard scattering event of interest is referred to as the Primary Vertex (PV). The pile-up products are uncorrelated with the PV and typically consist of a mixture of inelastic, elastic and diffractive pp processes, which are separated in the longitudinal direction. These products can interfere by overlapping with the products produced by the collision of interest, causing an additional correction to be applied, or can generate generate new clusters contributing to the PV jets signal. Subtraction methods are used to mitigate the contributions of the

pile-up to the jets, these algorithms consist in estimate the average amount generated by pile-up products using MC simulation, and then subtract this value to the jet energy. This procedure depends by a number of factors such as the number of interactions per bunch crossing μ, the number of primary vertices N_{PV} and pseudo-rapidity η, from the reconstructed jet pt. One of the method used is the pile-up energy subtraction which is characterized by the following scheme:

$$p_t^{corr} = p_t^{jet} - \mathcal{O}(\langle\mu\rangle, N_{PV}, \eta) = p_t^{jet} - \rho \times A^{jet},$$

where ρ is the estimated pile-up p_t density defined as

$$\rho = \text{median}\left\{ \frac{p_{t,i}^{jet}}{A_i^{jet}} \right\},$$

in which $p_{t,i}^{jet}$ is the transverse momentum of each k_t jet and A_i^{jet} is his corresponding area; each jet is defined with an nominal radius parameter $R_{k_t} = 0.4$. An extension to this method is the pile-up shape subtraction, which determines the sensitivity of the jet shape observables to pile-up by considering how it is affected by adding depositions of infinitesimally soft particles. This variation is calculated for each jet in each event and then extrapolated to zero to obtain the correction. The method uses a uniform distribution of infinitesimally soft particles, called *ghosts*, which are added to the event. A number density ν_g per unit is associated to these ghosts in the $y - \phi$ space and its inverse is the individual ghost area A_g. The four-momentum vector of the i^{th} ghost is defined as

$$g_{\mu,i} = (g_t \cos\phi_i, g_t \sin\phi_i, g_t \sinh y_i, \cosh y_i),$$

where g_t is transverse momentum of the ghosts, which are set to have zero mass. This definition creates a uniform ghost density equal to g_t/A_g, that is used as a proxy to evaluate the pile-up contribution. A different correction value can be obtained by changing the pt density of the ghosts and checking again the shape change in the structure variables of the jets. Due to the introduction of these ghosts, a jet shape variable $\mathcal{V}$ become also a function of the ghost g_t transverse momentum. The reconstructed, but uncorrected, jet shape is then $\mathcal{V}(g_t = 0)$, while the corrected one can be calculated by extrapolating to the value of $g_t = -\rho \times A_g$, that cancels the pile-up density contribution. The corrected jet shape variable can be calculated by using the Taylor expansion:

$$\mathcal{V} = \sum_{k=0}^{\infty} (-\rho \cdot A_g)^k \left. \frac{\partial^k \mathcal{V}(\rho, G_t)}{\partial g_t^k} \right|_{g_t=0}.$$

The derivatives are calculated by using different values of $\mathcal{V}(g_t)$ for $g_t \geq 0$.

Chapter 2

Jet substructures

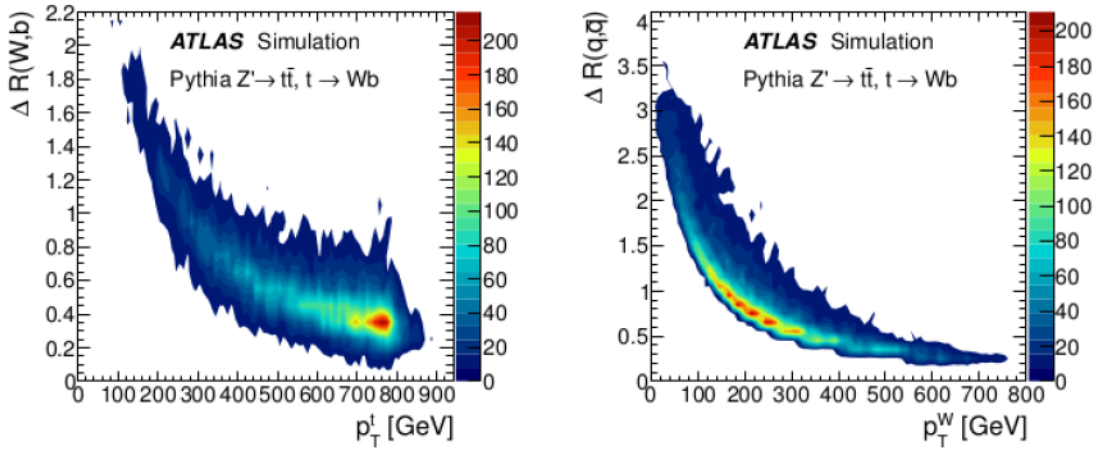
The jets algorithms descibed in the previous chapter allow to reconstruct the jet's recombination history, but these methods are not able to distinguish whether a jet is originated by a decaying boosted electroweak (EW) resonance or by a QCD parton. This task is up to the jet substructure methods that study the internal structure of jets and allow to distinguish a jet originated from a boosted massive particles (see section 2.1), and hence produced by the superposition of two almost correlated partons, from jets produced by a single parton.

2.1 Boosted objects

LHC center of mass energy is such that intermediate bosons (W/Z) may be produced with momenta which are much larger than their masses. At these energies, heavy particles of SM, i.e. W, Z and H bosons and top quarks, are produced with large transverse momentum (boosted particles) and as a consequence their decay products have large Lorentz boost [5]. The property of highly boosted particles is that their decay product may be collimated to the momentum direction of the mother particle. In the detector rest frame the angular separation between the decay product may be as small as

$$\Delta R \simeq 2m/p_t \ ,$$

where ΔR is calculated as in the section 1.1, while m is the mass of decaying particle and p_t is the transverse momentum of the particle. In the figure 2.1 is shown the angular separation between the W and b decay products of a top quark in simulated $Z' \to t\bar{t}$, as well as the separation between the light quarks of the subsequent W decay.



(a) $t \to Wb$ (b) $W \to q\bar{q}$

Figure 2.1: In the left plot is represented the angle between the W and b in top decay $t \to Wb$ as function the top transverse momentum in simulated Pythia $Z' \to t\bar{t}$ event. On the right the angle of $W \to q\bar{q}$ system from $t \to Wb$ decays.

It has been observed that the capability to resolve the decay products using standard jet algorithms begin to deteriorate for $p_T^W > 200$Gev, while when $p_T^t > 200$, the top quark decays products have a separation $\Delta R < 1.0$.
A direct consequence of this phenomenom is that the traditional reconstruction algorithms start to lose efficiency and it is therefore appropriate the use of *large-R jet* for the reconstruction of these objects as shown in the figure 2.2.

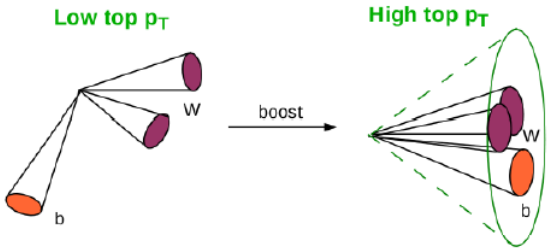


Figure 2.2: Jet produced in top quark hadronic decays, in case of low (left) values of top p_t and high (right) value of transverse momentum using a large-R jet reconstruction.

Obviously a single large-R jet, that contains all of the decay products of a heavy particle, will have properties that differ from a single large-R jet with the same p_T which originates from a quark or gluon. In order to use these jets it is needed to understand the sub-

structure of the large-R objects. The most interest parameters for this characterization are discussed in the following subsections.

2.2 Jet substructure methods

The aim of jet substructure is to study the kinematic properties of high-p_t jet and being able to distinguish between jets produced by one or more than one partons. Three wide categories can be identified according on which physical observation they rely on [8].

Prong finders. Observables in this category use the fact that partons, which decays from highly boosted particles, carry a sizeable fraction of the initial jet transverse momentum and the result is a large amount of multiple hard cores in the jet. Therefore prong finders look for these cores in a jet, reducing the contamination from "standard" QCD jets that are characterized by single-core jets due to their radiation of soft gluons. For this reason boosted jets are labelled in terms of their "pronginess", i.e. to their expected number of hard cores: QCD jets should be one-prong objects, Z/W/H jets would be two-pronged and boosted top jets would be three-pronged.

Groomers. Jet groomers are procedures conceived to "clean" part of the soft and large-angle radiation present in a fat jet. These type of jets indeed are particularly sensitive to soft backgrounds such as the underlying event, i.e. what is detected in a event that is not coming purely from the primary hard scattering process,and pile-up. The distinction between groomers and prong finder is not always clear, and one same procedure can be considered a prong finder or a groomer depending on many factors, such as how they are combined with other methods or even the choice of parameters.
Radiation constraints. Another difference between 1-parton and multi-parton jets is their colour structure, which means that they will display different soft-gluon radiation patterns. Jet shapes are functions of the constituents of the jet and explore the fact that different type of jets are conceived in such a way that they have larger values for QCD jets and smaller values for multi-parton jets.

2.3 Prong-finders and groomers

Prong-finding and grooming can be performed with many different techniques, each depending on different parameters fulfilling the particular need of the algorithm in use. The three main algorithms used by ATLAS and CMS are trimming, pruning and mass-drop filtering.

2.3.1 Mass-drop filtering

The mass drop filtering procedure aims to isolate concentrations of energy within a jet by finding and tagging symmetric subjets with a smaller mass than the original jet [5]. This method use the C/A algorithm because it provides an angular ordered shower history which begins with the widest combinations while the cluster sequence is reversed. Mass-drop filtering method has two stages:

- *Mass-drop and symmetry.* In this stage, jet costituents of the fat jets are reclustered with C/A algorithm and then we undo the last step of the clustering. The jet splits into two subjets, j_1 and j_2, in such a way that the mass of the first jet is larger ($m^{j_1} > m^{j_2}$). The mass-drop demands that the difference between the original jet mass m^{jet} and the m^{j_1} after the splitting respect the following relation:

$$m^{j_1}/m^{jet} < \mu_{frac} \ , \tag{2.1}$$

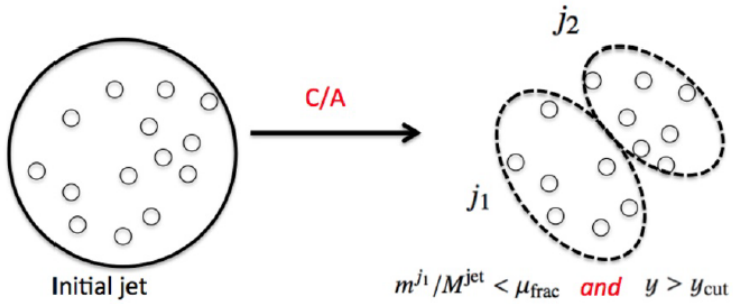
  where μ_{frac} is a parameter of the algorithm. The above splitting is also expected to symmetric with the subsequent requirement:

$$\frac{\min[(p_t^{j1})^2, (p_t^{j2})^2]}{(m^{jet})^2} \Delta R^2_{j_1,j_2} > y_{cut} \ , \tag{2.2}$$

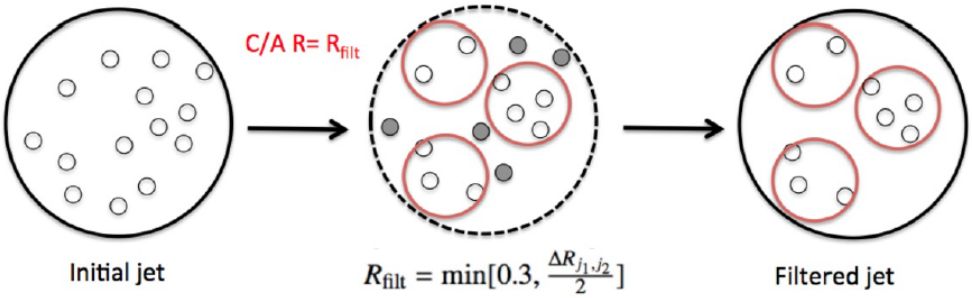
  where $\Delta R_{j_1,j_2}$ is the value of the opening angle between j_1 and j_2, while *cut* is a parameter that define how the energy is shared between the subjets considered in the original jet. When both criteria are verified, "$i + j$" is kept as the result of the mass-drop procedure, while if these are not satisfied, the jet is discarded.

- *Filtering.* The constituents of the subjets j_1 and j_2 are reclustered with the use of C/A algorithm with a radius parameter $R_{filt} = \min[0.3, \Delta R_{j_1,j_2}/2]$, with $R_{filt} < \Delta R_{j_1,j_2}$. The jet is therefore filtered using this criterion and all the constituents outside the three hardest subjets, which number is chosen to allow one more radiation from a two body decay to be taken, are discarded.

The figure 2.3 illustrates how the mass-drop filtering procedure works, showing the two different stages.



(a) The mass-drop and symmetric splitting criteria.



(b) Filtering.

Figure 2.3: Representation of the two stages of the mas-drop filtering.

2.3.2 Trimming

The trimming algorithm exploits the fact that the contamination from underlying events, pile-up, initial state radiation (ISR), and the multiple parton interactions (MPI) are much softer than the partons generated in the hard-scattering events and in the final-state radiation (FSR). The trimming method uses a k_t algorithm to build subjets with a radius R_{sub} from the costituents of the original jet. If these subjets do not satisfy the relation $p_{t_i}/p_t^{jet} < f_{cut}$ are removed, in which p_{t_i} is the transverse momentum of the of the i^{th} subjet and f_{cut} is a parameter of the algorithm. The smaller-radius jets with a momentum fraction $f < f_{cut}$ are removed, while the remaining form the trimmed jet, as it is shown in the figure 2.4.

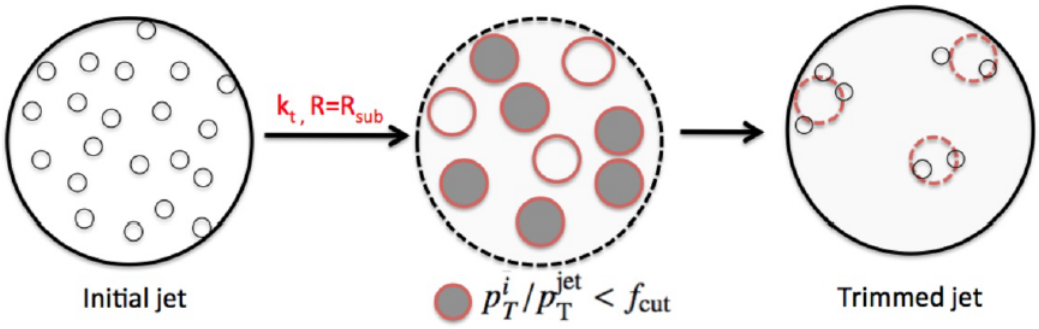


Figure 2.4: Diagram of the trimming procedure of a jet.

2.3.3 Pruning

The pruning algorithm is similar to the trimming one as that it removes constituents with a small relative p_t, but with an additional condition on the wide-angle radiation. This procedure is applied at each recombination step of the jet algorithm, that can be either the k_t or the Cambridge/Aachen, and it is based on the addition of the constituent considered if they satisfy fixed criteria. The method, which is presented in the figure 2.5, run as follows:

1. The k_t or C/A recombination jet algorithm operates on the reconstructed constituents.

2. During every recombination step of the constituents j_1 and j_2, in which the transverse momentum of j_1 is larger than the other, the following condition must be satisfied:

$$\frac{p_t^{j_2}}{p_t^{j_1+j_2}} > z_{cut},$$

$$\Delta R_{j_1,j_2} < R_{cut} \times \left(\frac{2m^{jet}}{p_t^{jet}}\right),$$

   where z_{cut} and R_{cut} are parameters of the algorithm.

3. j_2 with j_1 are merged if one or both of the conditions above are satisfied, otherwise j_2 is discarded and the algorithm goes on.

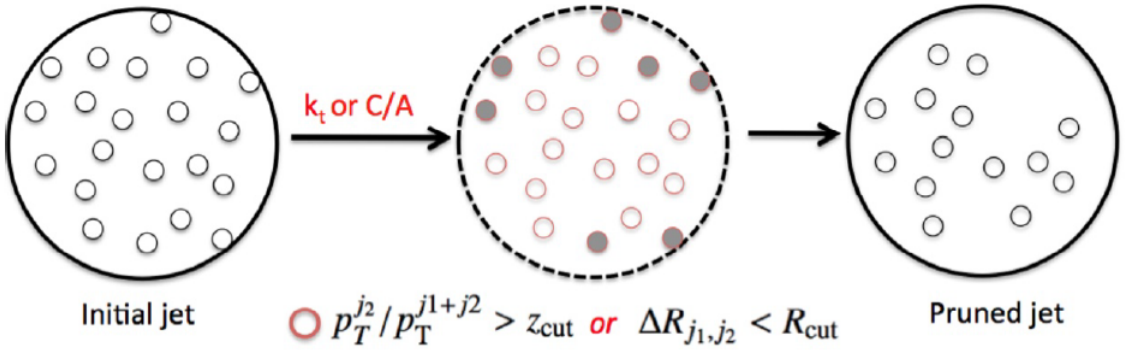


Figure 2.5: Diagram of the pruning procedure of a jet.

2.4 Radiation constraints

The standard approach to constrain radiation inside a jet is to impose a cut on the observables of the studied jet. Jet shapes [8] are function of the constituents of the jet which differ in case of one-parton or multi-parton jets. Two of the most used algorithms are described in the following.

2.4.1 *N*-subjettiness

N-subjettiness is a jet shape variable accounting for the number of subjets constituting the observed jet. In order to obtain the subjettiness. A new set of axes $a_1, ..., a_n$ is introduced and then the variable $\tau_n^{(\beta)}$, which is the jet shape, is defined as follows:

$$\tau_n^{(\beta)} = \sum_{i \in jet} p_{ti} \min(\Delta R_{ia_1}^{\beta}, ..., R_{ia_n}^{\beta});$$

where β is a free parameter, while the axis a_i can be chosen in various ways, as shown below:

- **k_t axes**. In this case the jet is reclustered with the k_t algorithm, as the subsection suggests, and the axes a_i are taken N exclusive jets.

- **WTA k_t axes**. Here the jet is reclustered using the k_t algorithm, with the *winner-takes-all* recombination scheme. The a_i axes have the same properties of the case before, but the use of the WTA scheme guarantees a recoil-free observable.

- **generalised-k_t axes**. This case is defined as above but now one uses the exclusive jets obtained with the generalised-k_t algorithm. It is easier to set the p parameter

of the algorithm to $1/\beta$, in this way the distance measure used for the clustering will be matched with the one used to compute τ_n; if $\beta < 1$ one would use the WTA generalised-k_t axes.

- **minimal axes**. The axis a_i are chosen to minimize the value of the variable τ_n.

Therefore τ_n is a measure of the radiation around the N axes chosen $a_1, ..., a_n$. Depending on the N prongs of a jet, it is expected that $\tau_1, ..., \tau_{n-1}$ will be large while the τ_n with $n > N$ will be small. The value of τ_n will also be larger when the prongs are gluons, so for this reason the N-subjettiness ratio

$$\tau_{n,n-1}^{(\beta)} = \frac{\tau_n^{(\beta)}}{\tau_{n-1}^{(\beta)}}$$

is defined to better distinguish the N-prong signal jets from the QCD backgrounds. For this reason, a cut is imposed on the ratio $\tau_{21}^{(\beta)} < \tau_{cut}$ to discriminate W/Z/H jets against QCD jets and on $\tau_{32}^{(\beta)} < \tau_{cut}$ to discriminate top jets against QCD jets. In the most common use of N-subjettiness the parameter β is taken equal to one.

2.4.2 Angularities

Another group of jet shapes are the *generalised angularities* which are defined as

$$\lambda_\beta^k = \sum_{i \in jet} z_i^k \left(\frac{\Delta R_{i,jet}}{R} \right)^\beta,$$

where z_i is the transverse momentum fraction of the i_{th} constituents of the jet and $\Delta R_{i,jet}$ its distance to the jet axis:

$$z_i = \frac{p_{t,i}}{\sum_{j \in jet} p_{t,j}} \qquad \text{and} \qquad \Delta R_{i,jet}^2 = \Delta\phi_{i,jet} + \Delta y_{i,jet}. \tag{2.3}$$

These jet shapes, however, are not collinear safe, except for the case in which $k = 1$ and IRC safety is respected. In this condition they are called *angularities* with $\lambda_\beta \equiv \lambda_\beta^{(k=1)}$ and when $\beta = 1$ is the case is referred to as width. Following the definition above, the more radiation is contained in a jet, the larger the generalised angularities are; therefore both definition give the measure of how much QCD radiation there is around the axis, for example they are used for the quark-gluon discrimination.

2.4.3 Energy-Correlation Functions

Energy-correlation functions (EFCs) are essentially similar to the N-subjettiness, but with the difference that here is not required the selection of the N reference axes. They are defined as

$$e_2^\beta = \sum_{i<j\in jet} z_i z_j \Delta R_{ij}^\beta, \tag{2.4}$$

$$e_3^\beta = \sum_{i<j\in jet} z_i z_j z_k \Delta R_{ij}^\beta \Delta R_{jk}^\beta \Delta R_{ik}^\beta, \tag{2.5}$$

$$\vdots \tag{2.6}$$

$$e_N^\beta = \sum_{i<\ldots<i_n\in jet} \Big(\prod_{j=1}^{N} z_{ij} \Big) \Big(\prod_{k<l=1}^{N} \Delta R_{i_k i_l}^\beta \Big), \tag{2.7}$$

with $z_i = p_{t.i} / \sum_j p_{t,i}$. Energy-correlation function are insensitive to recoil for all values of the angular exponent β and this allow to have easier calculations as in the case of the WTA axes.
There are more general versions of the EFCs which involve p_t weighted sums over pairs, triplets, ... of particles but built from other angular combinations:

$${}_1e_2^{(\beta)} \equiv e_2, \tag{2.8}$$

$${}_3e_3^{(\beta)} \equiv e_3, \tag{2.9}$$

$${}_2e_3^{(\beta)} = \sum_{1<j<k\in jet} z_i z_j z_k \, \min(\Delta R_{ij}^\beta \Delta R_{ik}^\beta, \Delta R_{ij}^\beta \Delta R_{jk}^\beta, \Delta R_{ik}^\beta \Delta R_{jk}^\beta), \tag{2.10}$$

$${}_1e_3^{(\beta)} = \sum_{1<j<k\in jet} z_i z_j z_k \, \min(\Delta R_{ij}^\beta, \Delta R_{jk}^\beta, \Delta R_{ik}^\beta), \tag{2.11}$$

$$\vdots \tag{2.12}$$

$${}_ke_N^\beta = \sum_{i<\ldots<i_n\in jet} \Big(\prod_{j=1}^{N} z_{ij} \Big) \Big(\prod_{l=1}^{k} \min_{u<v\in i_1,\ldots,1_N}^{l} \Delta R_{uv}^\beta \Big), \tag{2.13}$$

with $\min^{l}$ denoting the l^{th} smallest number.
In order to discriminate boosted massive particles from background QCD jets, it is useful to introduce ratios of generalised-EFCs, like it has been made for the N-subjettiness.

Here there are some examples of ratios for two-prong taggers

$$C_2^{(\beta)} = \frac{{}_3e_3^{(\beta)}}{\left({}_1e_2^{(\beta)}\right)^2} \equiv \frac{e_3^{(\beta)}}{\left(e_2^{(\beta)}\right)^2}, \qquad D_2^{(\beta)} = \frac{e_3^{(\beta)}}{\left(e_2^{(\beta)}\right)^3}, \tag{2.14}$$

$$N_2^{(\beta)} = \frac{{}_2e_3^{(\beta)}}{\left(e_2^{(\beta)}\right)^2}, \qquad M_2^{(\beta)} = \frac{{}_1e_3^{(\beta)}}{e_2^{(\beta)}}, \tag{2.15}$$

while the following values are introduced for the three-prong taggers

$$C_3^{(\beta)} = \frac{e_4^{(\beta)}e_2^{(\beta)}}{\left(e_3^{(\beta)}\right)^2}, \qquad N_3 = \frac{{}_2e_4^{(\beta)}}{\left({}_1e_3^{(\beta)}\right)^2}, \qquad M_3 = \frac{{}_1e_4^{(\beta)}}{{}_1e_3^{(\beta)}};$$

$$D_3^{(\alpha,\beta,\gamma)} = \frac{e_4^{(\beta)}\left(e_2^{(\alpha)}\right)^{\frac{3\gamma}{\alpha}}}{\left(e_3^{(\beta)}\right)^{\frac{3\gamma}{\beta}}} + k_1\left(\frac{p_t^2}{m^2}\right)^{\frac{\alpha\gamma}{\beta}-\frac{\alpha}{2}} \frac{e_4^{(\gamma)}\left(e_2^{(\alpha)}\right)^{\frac{2\gamma}{\alpha}-1}}{\left(e_3^{(\beta)}\right)^{\frac{2\gamma}{\beta}}} + k_2\left(\frac{p_t^2}{m^2}\right)^{\frac{5\gamma}{2}-2\beta} \frac{e_4^{(\gamma)}\left(e_2^{(\alpha)}\right)^{\frac{2\beta}{\alpha}-\frac{\gamma}{\alpha}}}{\left(e_3^{(\beta)}\right)^2},$$

where k_1 and k_2 are $\mathcal{O}(1)$ constants. In this series, the D family has typically a larger discriminating power, at the expense of being more sensitive to model-dependent soft contamination in the jet like the UE or pileup. The N family is closer to N-subjettiness, and the M family is less discriminating but more resilient against soft contamination in the jet.

2.5 Jet Tagging

Particle identification is one of the goals when using detector such as ATLAS and CMS, therefore a large number of substructure variables have been developed to identify the particle origin of jets. The term tagger indicates though the use of one or more of the jet substructure variables to discriminate jets coming from different types of particles. For example, in the quark/gluon discrimination, jet shapes like angularities and energy correlation functions are used since it has been observed that gluons tend to radiate more than quarks and these variables, as explained in the previous section, are a measure of this radiation, while in the two- and three-prong tagging there is a combination of jet shapes, such as ECFs and N-subjettiness, and groomers.

2.5.1 Vector Boson Tagging

The two-prong decays from a electroweak vector bosons (H/W/Z) tend to have distinct pattern of radiation if compared to high p_t gluons and quarks, i.e. boosted bosons usually have two subjects with similar momentum while quark and jets have mostly one single prong and if they have two, the second is generally softer [4].
Two-prong taggers tend to combine a prong finder, which can act as a groomer, and a cut on a jet shape for radiation constraint. Discriminating jets with one prong versus two prongs requires the comparison between the value of different N-subjettiness or EFCs, for example if $\tau_1^{(\beta)}$ is larger than $\tau_2^{(\beta)}$ it means that the radiation in the jet is localized about two hard direction, which implies that the jet is two pronged. Usually the variables used in bosons tagging are:

- $\mathbf{C_2}$: A ratio of two- and three-point energy correlation functions, which are sums over the pT-weighted angular separations of the pairwise and tripletwise combinations of jet constituents defined as

$$C_2 = \frac{e_3}{(e_2)^2},$$

  where the exponent has been removed and set to $\beta = 1$ in the Eq.(2.14).

- $\mathbf{D_2}$: A variation on the ratio of energy correlations that optimises the separation between one-prong and two-prong decays and is described by the following equation

$$D_2 = \frac{e_3}{(e_2)^3},$$

  in which has been adopted the same consideration done above.

- $\boldsymbol{\tau_{21}}$: The N-subjettiness ratio choosen with the winner-take-all method of defining internal axes and calculated as

$$\tau_{21} = \frac{\tau_2}{\tau_1}$$

  with β set equal to one.

However, even if the individual N-subjettiness or energy correlation function observables are IRC safe, $\tau_{21}^{(\beta)}$, $D_2^{(\beta)}$ and $C_2^{(\beta)}$ are not generically IRC safe. The variables will become safe if is applied a cut on the jet mass, which acts as a cut on the denominator.

Both ATLAS and CMS use these jet substructure variables combined with the jet mass to create taggers. For example one of the standard ATLAS bosons tagger for Run 2 was the so called "R2D2", where the jet shape considered is the ratio C_2 and the jet mass m^{comb}, which is defined as

$$m^{comb} = \left(\frac{\sigma_{calo}^{-2}}{\sigma_{calo}^{-2} + \sigma_{TA}^{-2}}\right) m^{calo} + \left(\frac{\sigma_{TA}^{-2}}{\sigma_{TA}^{-2} + \sigma_{calo}^{-2}}\right) m^{TA}, \tag{2.16}$$

where σ_{calo} and σ_{TA} are the calorimeter-based jet mass resolution and the track-assisted mass resolution, while m_{calo} and m_{TA} are their corresponding masses. The jet mass is trimmed with the anti-k_t algorithm with a $f_{cut} = 0.05$ and $R = 1.0$ jets with a $R_{sub} = 0.2$, that is identified with the acronyms R2.
Another jet substructure variable adopted is the N-subjettiness dichroic ratio. The dichroic ratios use a variable like τ_{21} using a ratio between loose grooming τ_2 and tight groomed τ_1. This choice is driven by the fact that the shape measurement for single prong is expected to have better performance in non-groomed jets; on the other hand large-R jets would suffer more of UE and pile-up contribution, so in this case the grooming would help in rooting out the contributions not related to the actual partons forming the jet. The dichroic ratio is defined as:

$$\tau_{21}^{(\beta,dichroic)} = \frac{\tau_2^{(\beta,loose\ grooming)}}{\tau_1^{(\beta,tight\ grooming)}},$$

in which the denominator (τ_1), that is sensitive to two hard prongs, is computed on the result of the groomer jet, while the numerator (τ_2) is computed on a larger jet, which can be either the plain jet itself or just lightly-groomed.

2.5.2 Top Tagging

The three-prong decays of boosted top quarks in the hadronic channel has more phenomenology for their identification than the two-prongs decays of vector bosons [3]. Top tagging must operate in a moderate boosted regime in which the decay products could not be contained all inside a single jet with $R < 1.0$ because the mass of the top quark is heavier than the electroweak bosons.
A top-tagging algorithm is formed by the use of two substructure-related variables which are respectively the jet mass calculated as in the Eq.(2.16) and the N-subjettiness ratio τ_{32} defined as follow:

$$\tau_{32} = \frac{\tau_3}{\tau_2},$$

where the parameter β has been set equal to one and the N-subjettiness variables τ_2 and τ_3 has been reconstructed with a winner-take-all recombination scheme. The ratio τ_{32} allows the discrimination between jet with a three-prong structure and jet with a two-prong structure, depending on the cut on the jet shape.
The jets used in this algorithm are reconstructed with the anti-k_t trimming algorithm with the radius parameter $R = 1.0$, the subjets radius parameter set to 0.2 and the transverse momentum fraction $f_{cut} = 0.05$.

Another algorithm used is the HEPTopTagger (HTT) [8] which was firstly created to reconstruct boosted top quarks with medium pt, for example for the reconstruciton of of top quarks in the process $pp \to t\bar{t}h$, where the decays product are semi-leptonic. The algorithm proceeds as follows:

1. Firtsly the fat jet is defined using the Cambridge/Aachen algorithm with jet radius parameter $R = 1.5$,

2. then for a given large-R jet, one undoes the last step of the clustering, i.e. declustering the jet j into the subjets j_1 and j_2 that have masses m_2 and m_2 correlated by the formula $m_{j1} > m_{j1}$, until a mass drop $m_{j1} < 0.8m_j$ is observed. If this condition is not satisfied, the declustering procedure continue with j_1.

3. After the drop condition is met the subjets is further decomposed into smaller subjets if the initial subjet has a mass greater than 30 Gev.

4. In this phase is applied a filtering radius $R_{filt} = \min(0.3, \Delta R_{ij})$; a third hard subjet is added to all the considered pairs of subjets, then the filter is used on the three subjets keeping the 5 hardest pieces that will necessary to evaluate the jet mass. Amongst all the triplets of the original hard subjets, the used combination is the one with the resulting jet mass closer to the top mass and that lie in the window of the true top mass (150-200 Gev).

5. From the 5 filtered pieces, only three j_1, j_2, j_3, ordered in p_t, are extracted to from a subset and accepted to be a top candidate if the masses satisfy at least one of the following criteria:

$$
\begin{aligned}
0.2 < \arctan\left(\frac{m_{13}}{m_{12}} < 1.3\right) \qquad & and \qquad R_{min} < \frac{m_{23}}{m_{123}} < R_{max} \\
R^2_{min}\left(1 + \frac{m^2_{13}}{m^2_{123}}\right) < 1 - \frac{m^2_{23}}{m^2_{123}} < R^2_{max}\left(1 + \frac{m^2_{13}}{m^2_{123}}\right) \qquad & and \qquad \frac{m_{23}}{m_{123}} > 0.35 \\
R^2_{min}\left(1 + \frac{m^2_{12}}{m^2_{123}}\right) < 1 - \frac{m^2_{23}}{m^2_{123}} < R^2_{max}\left(1 + \frac{m^2_{12}}{m^2_{123}}\right) \qquad & and \qquad \frac{m_{23}}{m_{123}} > 0.35
\end{aligned}
$$

6. As last step, the combined pt of the 3 subjets of the previous phase is imposed to be at least 200 Gev.

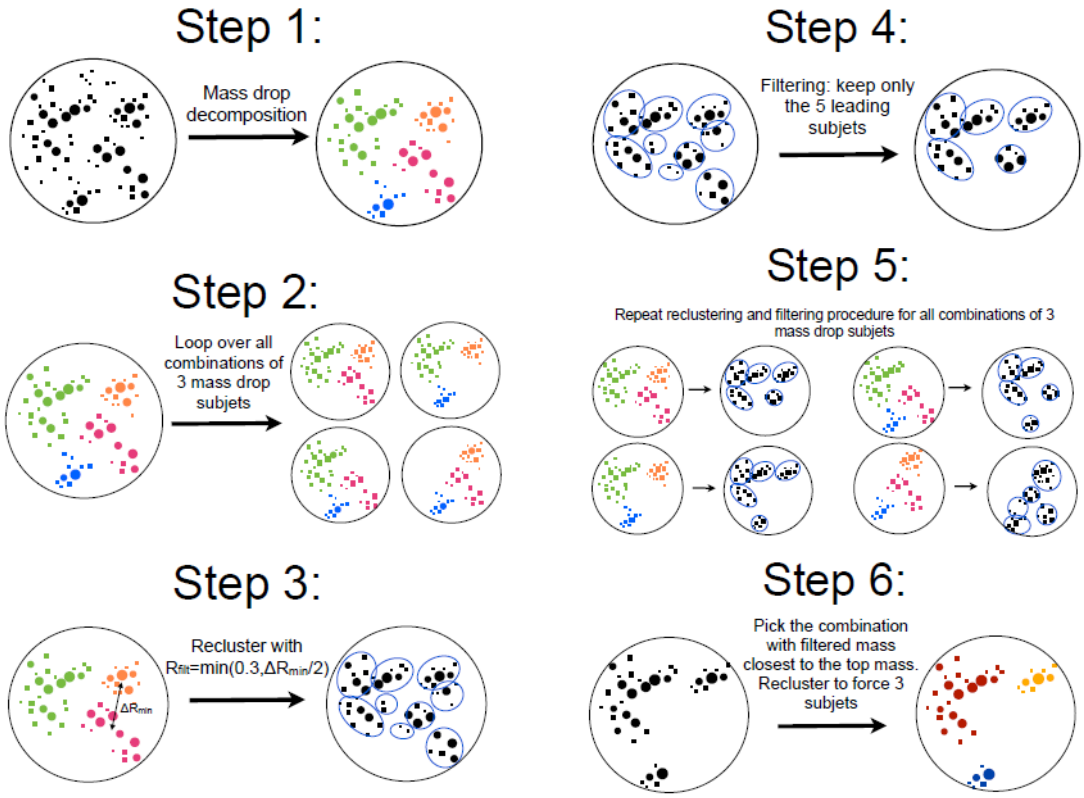


Figure 2.6: Diagram of the HEPTop tagger algorithm.

In the figure 2.6 it is visually summarized the entire algorithm explained above. The first three step decompose a massive object (a fat jet) into his hard partons while the filtering step has the task of cleaning the contamination from the Underlying Event like in the grooming algorithm. The criteria described in the fifth step form a cut on the three subjet, which acts like a 3-parton system, to match the kinematics of a top decay and to remove the QCD background.

Chapter 3

Dataset and Analysis

In this chapter will be discussed the data used in the following analysis, i.e. how they have been generated, and the way these data can be studied with the use of the jet substructure, introduced in the section 2. In particular, the next part will focus on the N-subjettiness τ_{32}, which helps in the discrimination between three-prong jets and two-prong ones, and on the D_2 variable which is effective in the vector boson tagging.

3.1 Simulated events

The physics case examined is the collision of a proton-proton couple (pp) in the LHC with an energy in the center of mass of 13 TeV, where the events that can be considered are the following: $t\bar{t}$ pair production, single top production (s-channel,t-channel and Wt-channel mode), ttV production (with V being an electroweak boson), W/Z+jets production, diboson and multi-bosons production and both the Higgs boson production (tth and Vh). All this processes have been simulated through a Monte-Carlo generation (MC), with different generator depending on the production considered. For example `Sherpa 2.2.1` [1] has been used in the case of diboson, multibosons and W/Z+jets vents even for the parton shower and hadronisation samples, while in all the other cases the generator for the production is `Powheg-Box` [9] and the one for the hadronisation process is `Pythia` [11].

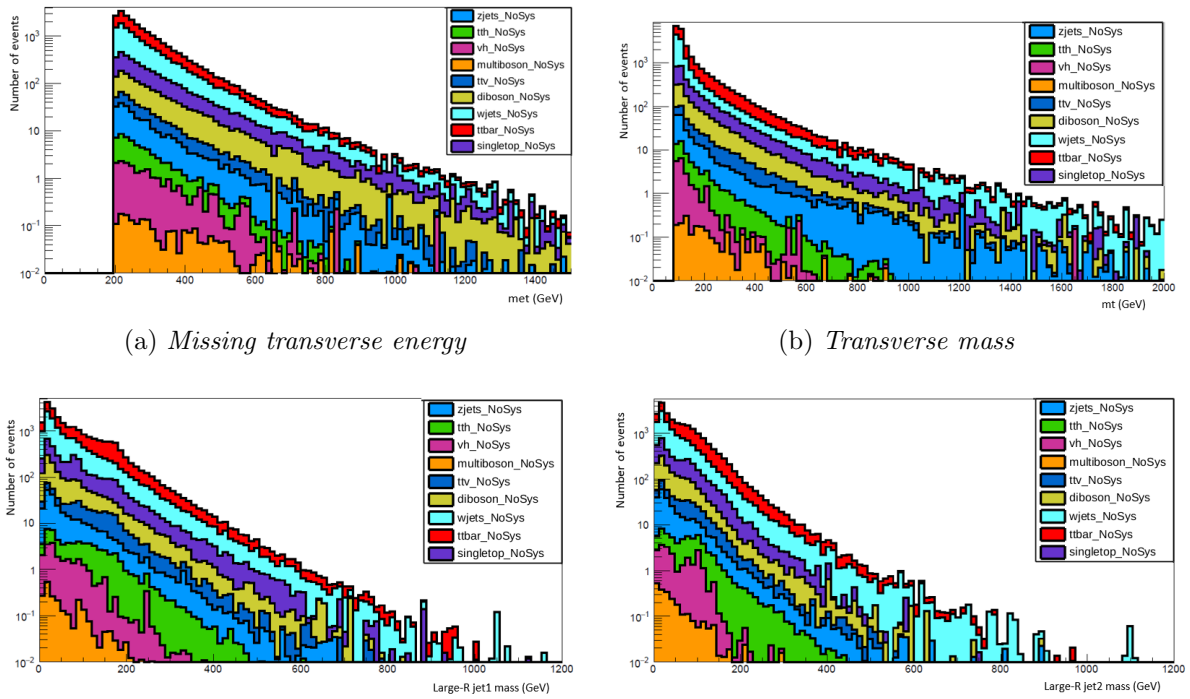
In the following samples including at least 1 large R jet are considered. The large-R jets selected are built using the anti-k_t algorithm with a radius parameter R equal to 1.0, which then have been trimmed with the ATLAS configuration known as "R2", where the parameters are $R_{sub} = 0.3$ and $f_{cut} = 0.05$. R_{sub} defines the radius of the subjets of the large-R jet considered and f_{cut} sets the minimum value of the ratio between the p_t of the i^{th} subjet and the p_t of the large-R jet in order that the considered subjet is not excluded from the jet. In this analysis to simplify the study of the generated data, the

systematic uncertainties have not been considered.
On the dateset has been made a preselection with the following characteristic:

- `N_Jet` > 1, which represents the number of the jets with transverse momentum larger than 30 GeV allowed to be present in the events,
- `met` > 200 GeV, that correspond to the missing transverse energy. This variable measures the energy imbalance in the transverse plane: since the total sum of the particles produced in a given event should be 0 in the transverse plane, `met` is defined as $\sum_i p_{t,i}$ of the reconstructed particles. Met represent the pt of the undetected particles (such as neutrinos) produced in the event of interest,
- `mt` > 50 GeV, which is the transverse mass defined as $m_t = \sqrt{p_t^2 + m^2}$.

3.2 Minimal analysis of jet substructures

For the analysis in this thesis the number of the large-R jets, produced by the decays of the different events presented above, has been set equal to two, respectively j_1 and j_2.
In the figure 3.1 are presented, for all the processes considered, the distributions of the missing transverse energy, the transverse mass and the mass related to the two large-R jets studied.
In the figure 3.2 are shown the jet substructures that have been chosen to be analyzed, which are respectively the energy correlation function ratio D_2 and the N-subjettiness ratio τ_{32}, calculated for both the large-R jets considered. As explained in the previous chapter the D_2 variable is used to discriminate two-prong decays from the one-prong decays, while the ratio τ_{32} allows to identify three-prong processes from other type of decay.



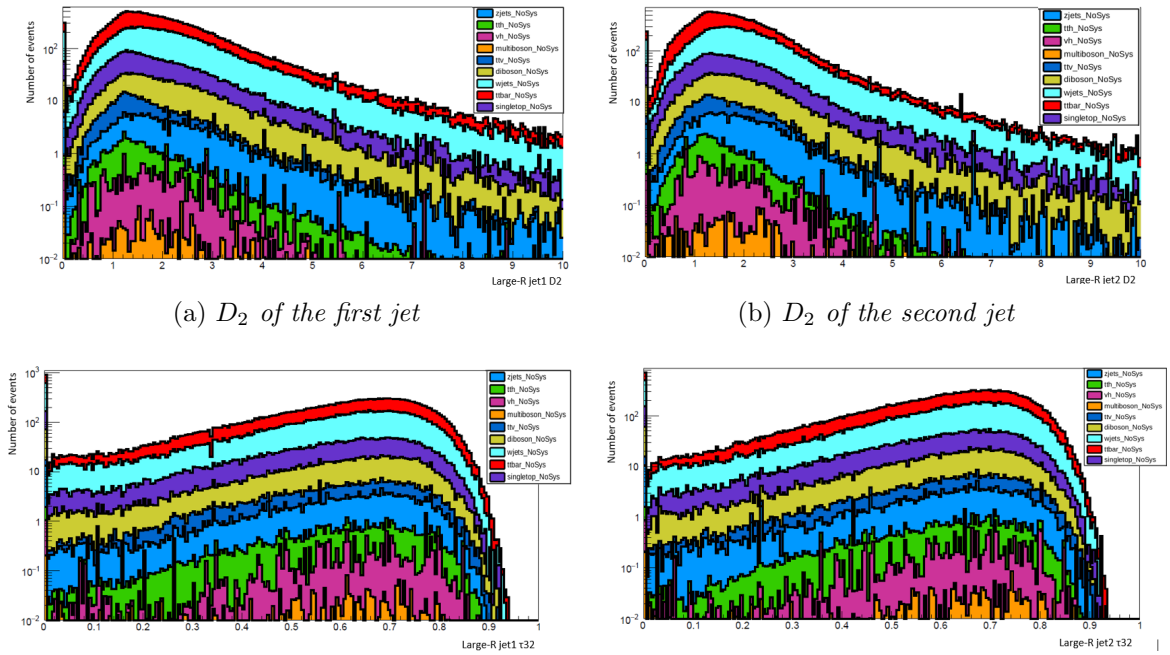
(a) *Missing transverse energy*

(b) *Transverse mass*

(c) *Large-R jet* m_1

(d) *Large-R jet* m_2

Figure 3.1: Comparison between the distributions of the met (a), m_t (b) and the masses of the two large-R jets (c-d), for different processes.



(a) D_2 *of the first jet* (b) D_2 *of the second jet*

(c) τ_{32} *of the first jet* (d) τ_{32} *of the second jet*

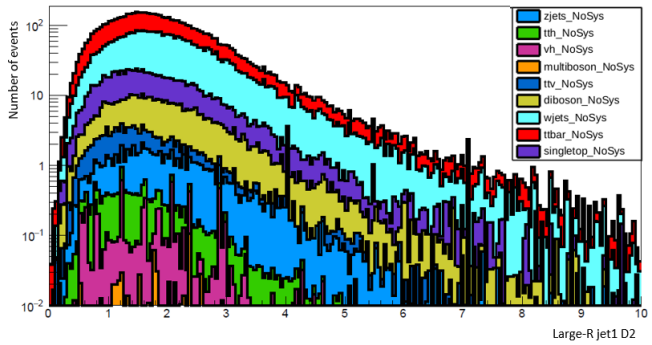
Figure 3.2: Distributions of the jet substructure D_2 variable for j_1 (a) and for j_2 (b) and comparison between the graphs of the variable τ_{32} of both the jets.

3.2.1 τ_{32} Effects

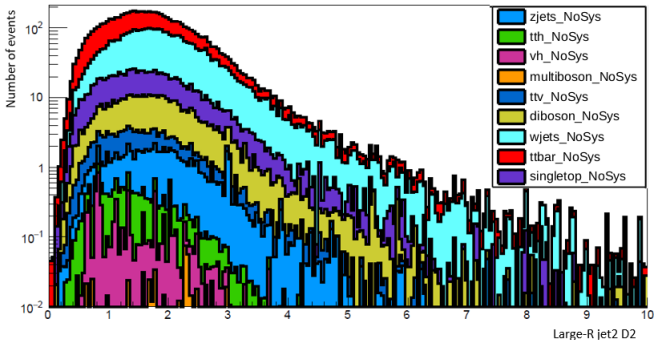
In this section will be shown what happens to the data and distributions described above if we impose a cut on the jet shape τ_{32} and what this means in the interpretation of the dataset.
The cut has been made by requiring a new selection of the data, which had to satisfy the following equation:

$$\tau_{32} = \frac{\tau_3}{\tau_2} > 0.6,$$

in which the N-subjettiness ratio, defined in section 2.4.1, can have a value between 0 and 1.
In the figure 3.3 are presented the D_2 jet shape for both the jets considered, after the cut on τ_{32} has been imposed.



(a) D_2 *of the first jet*



(b) D_2 *of the second jet*

Figure 3.3: Distributions of the jet substructure D_2 variable for j_1 (a) and for j_2 (b) after the constrain on the jet shape variable τ_{32} has been applied.

The effects of this selection in term of event yields is reported in table 3.1. From the yields we can observe a reduction of a factor 0.74 in diboson. This is due to the fact that diboson events with reconstructed fat jets are typically two-pronged fat jets, as they are produced by the hadronic decays of intermediate bosons. On the other hand events containing top decays have a much smaller rejection factor, as top decays goes as $t \to Wb$ and in case the W decays hadronically, as it does the 29% of the time it produces two more prongs resulting in a large number of 3-prong like jets.

We can see from this simple selection that the τ_{32} variable can, even by itself, separate the contributions from bosons and top decays, with good accuracy.

Type of event	Number of events before cut	Number of events after cut
dibosons	1056	278
multibosons	2.25	0.6
Vh	20.1	6.6
$t\bar{t}$	96442	2774
ttV	217	59.3
tth	42.3	11.2
single-top	2161.1	536
Z+jets	198.2	54.2
W+jets	9714	2583

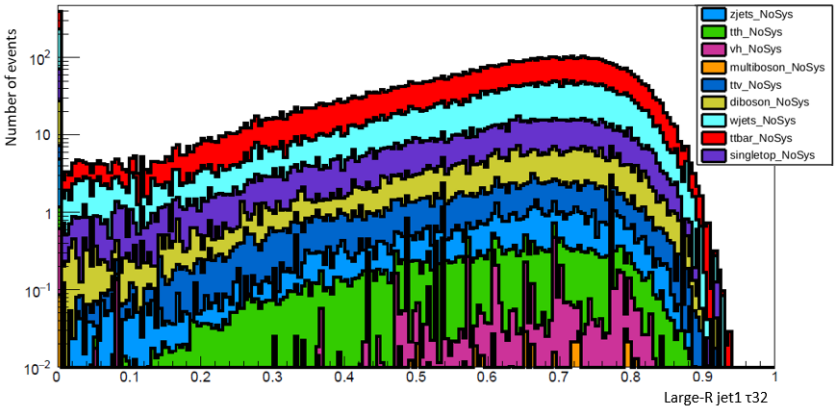
Table 3.1: Table of the number of simulated events processed before and after the cut on τ_{32}.

3.2.2 D_2 Effects

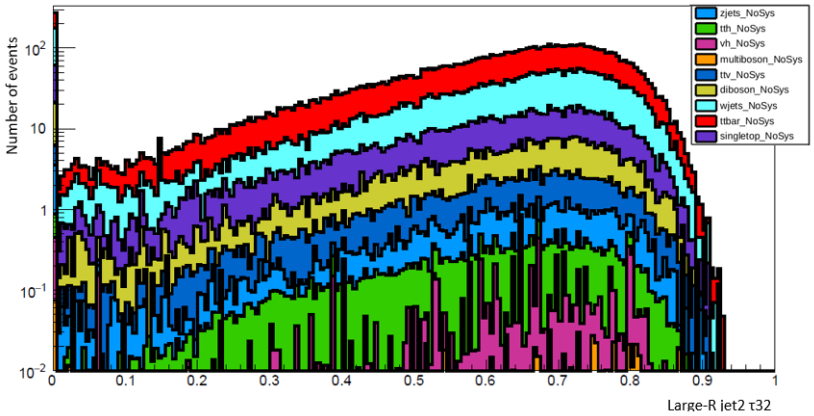
In this part the dataset will be analyzed with a constrain on the jet shape D_2 and will be described how the data have change and what implications this cut have involved.
As said before, the date will undergo an other selection that has to respect the following law:

$$D_2 = \frac{e_3}{(e_2)^3} < 2,$$

where e_3 and e_2 are the energy-correlation functions calculated as in the section 2.4.3. In the figure are represented the plots of the distributions of the N-subjettiness ratio τ_{32} in the case that the cut on the variable D_2 is applied.



(a) τ_{32} *of the first jet*



(b) τ_{32} *of the second jet*

Figure 3.4: Distributions of the jet substructure τ_{32} variable for j_1 (a) and for j_2 (b) when the cut on D_2 is imposed.

The way the data have changed can be seen in the table . It can be observed that the events which have usually two-pronged decays, such as the diboson and multiboson ones, have been heavily reduced of a factor about 0.73 and 0.7, while other type of events, mostly the one that contained top decays, such has *tth* events, have a smaller reduction factor which is around 0.48 as can be calculated from the table below.

With this selection it has been shown that even the jet shape variable D_2 can be contribute to the identification of two-prong processes from the three prong ones.

Type of event	Number of events before cut	Number of events after cut
dibosons	1056	285
multibosons	2.25	0.7
Vh	20.1	7.6
$t\bar{t}$	96442	3879.2
ttV	217	114
tth	42.3	23.4
single-top	2161.1	731.3
Z+jets	198.2	46.1
W+jets	9714	2198

Table 3.2: Table of the number of simulated events processed before and after the cut on D_2.

Conclusions

After describing the most common jet algorithm reconstruction, and the properties of large-R jets, this thesis presented a study on simulated LHC proton-proton events recorded by the ATLAS detector at a center of mass energy of $s = \sqrt{13}$ TeV.
The aim of this study is to study the effect of the substructure variables as selection variables for different physics processes. The study focused mostly on on the correlation energy function ratio D_2 and the N-subjettiness ratio τ_{32}, which have been compared in the distribution of the different processes for the two large-R jets considered. The analysis continues with the introduction of a cut on the variable τ_{32}, which has the aim of discriminate three-prong decays from the two-prong ones. Indeed the distributions achieved after the cut show that the number of the events is now reduced of a factor which depends on the type of process. For the diboson we can see that the reduction factor is about 0.74, because the decay products are two-pronged fat jets.
As last point has been imposed a cut even on the energy-correlation functions ratio D_2 and it has been shown from the distribution of N-subjettiness ratio and from the table related that it is a good variable which can help in the discrimination of two-pronged jets. For the diboson and multibosons events, which are typically two-pronged events, has been obtained a reduction factor of about 0.73 and 0.7, while 0.48 is the factor for the reduction of three-pronged fat jets.
From this thesis it can be seen the validity of the use of jet substructures such as the D_2 and the τ_{32} for the discrimination of the N-prong type of events and it is shown that these variables can be used in other and new experiments, with more complex and deeper analysis, such us with the use of neural networks and machine learning, to allow the identification of new processes and different events.

Bibliography

[1] T. Gleisberg et al. *Event generation with SHERPA 1.1*. 2009. arXiv: 0811.4622.

[2] The ATLAS Collaboration et al. *The ATLAS Experiment at the CERN Large Hadron Collider*. 2008. DOI: 10.1088/1748-0221/3/08/S08003.

[3] The ATLAS collaboration. *Boosted hadronic top identification at ATLAS for early 13 TeV data*. ATL-PHYS-PUB-2015-053. 2015.

[4] The ATLAS collaboration. *Identification of Boosted, Hadronically-Decaying W and Z Bosons in $\sqrt{s} = 13$ TeV Monte Carlo Simulations for ATLAS*. ATLAS-PHYS-PUB-2015-033. 2015.

[5] The ATLAS collaboration. *Performance of large-R jets and jet substructure reconstruction with the ATLAS detector*. ATLAS-CONF-2012-065. 2012.

[6] The ATLAS collaboration. *Performance of pile-up mitigation techniques for jets in pp collisions at $\sqrt{s} = 8$ TeV using the ATLAS detector*. 2016. arXiv: 1510.03823.

[7] A. J. Larkoski, D. Neill, and J. Thaler. *Jet Shapes with the Broadening Axis*. 2014. arXiv: 1401.2158.

[8] S. Marzani, G. Soyez, and M. Spannowsky. *Looking inside jets: an introduction to jet substructure and boosted-object phenomenology*. 2020. arXiv: 1901.10342.

[9] C. Oleari S. Alioli P. Nason and E. Re. *A general framework for implementing NLO calculations in shower Monte Carlo programs: the POWHEG BOX*. 2010. arXiv: 1002.2581.

[10] G. P. Salam. *Towards jetography*. 2010. arXiv: 0906.1833.

[11] S. Mrenna T. Sjöstrand and P. Z. Skands. *A brief introduction to PYTHIA 8.1*. 2008. arXiv: 0710.3820.