Improvement of the event selection for $W$-pairs in the semi-leptonic channel with a boosted hadronic $W$

by

Elizabeth Fons

Supervisors:
Dr. Sven Menke
Dr. Gustavo Otero y Garzón
Abstract

This thesis presents studies related to the improvement of the event selection of $W$-pairs in the semi-leptonic channel with a boosted hadronic $W$. The improvement of the event selection was done at two levels. One level is the offline event selection, where the boosted $W$ decaying hadronically is analyzed using a ‘fat jet’, and then looking at its substructure. Here we focus on the choice of jet variables and cuts, isolation criteria and event vetoes and the choice of topological relations between the final state objects.

The other event selection was implemented at generator level. In this case, we included filters in a Monte Carlo generator, in order to produce Monte Carlo samples which had already a preselection of boosted $W$s. This has several advantages, the most important one being that having a preselection in the generator means that we have a way of producing smaller samples that in the end will have large enough statistics for analysis. In the existing official Monte Carlo samples only 0.4% of the events survives the selection, while in our private production around 10% of the events remain.

We conclude the analysis with a review of the substructure variables and their discriminating power between signal and background. Using different sets of substructure and grooming variables we applied a multivariate analysis technique, using a neural network and boosted decision trees, in order to study which set of variables and parameters were better discriminants.
ACKNOWLEDGEMENTS

First of all I would like to thank Dr. Sven Menke for giving me the opportunity to work in the ATLAS HEC Group, for all his support during this work and for always finding the time to advice me! Special thanks to Fabian Spettel, for his guidance and generosity, this thesis wouldn’t exist without him. I had an amazing time working side by side with Fabian, through a million code bugs and Athena crashes... I would also like to thank Dr. Andreas Wildauer for looking after us, for his patience with our ROOT codes and for sharing the adventures of Lina and Manuel during lunch! And many thanks to the members of the ATLAS HEC Group and the MPI, Teresa, Martin, Andrey, Peter, Giselher, Jörg, Dennis, Horst, Stephan Stonjek and Elisabeth Träder, for making me feel welcome from the very first day of work! I enjoyed my time in Munich so much!

I would also like to thank Prof. Gustavo Otero y Garzón, for encouraging me to work in HEP, and for being my mentor ever since. Also special thanks to Prof. Ricardo Piegaia, who is always open to talk about physics, thank you for your advice and support!

Esta tesis es la conclusión de muchos años de estudio y trabajo, que no hubieran sido posibles sin muchísimas personas a las que quiero agradecer:

Gracias a mis papás, Myriam y Sergio, por su apoyo incondicional siempre. Y gracias a Xavi por su infinito cariño :) A Bob por la buena onda siempre y por revisar la tesis! A mis abuelos que ya no están pero que siempre los recuerdo.

A Ale, por su amor y paciencia, por estar siempre y ser mi compañero de aventuras.

A Naty Bisero que la quiero mucho, por ser mi amiga desde el primer día que pisé la facultad! A Fede Frith, el mejor compañero de laboratorio que alguien podría pedir (y aun mejor persona!) por hacerme llorar de risa en incontables horas de laboratorio. A Marto, Yago, Mati y Maxi S con quienes pasé infinitas horas de charlas, estudio y diversión . A los chicos de meteorología, después de tantos años me siento una de ellos! A los chicos de la facu Ger A., Mary, Ale M., Guada, Gastón, Vane, Fer y muchos más que seguro me olvido!

A Dario Rodrigues, que sabe física como nadie, por sus consejos y paciencia a lo largo de los años. A Juan Roederer, por las historias y la confianza, y al grupo del libro de electromagnetismo, Iván, Maria Emilia y Dani. Y gracias a Sabri, por su infinita buena onda (y por atemorizarme con el balde por meses!).
A Maxi Geier que es un genio y lo quiero mucho! A Denchi que siempre me lleva a comer a lugares ricos. A Jorge y Elsa por su apoyo, y por llenarnos de comida!

A Guille Mattei y Cristina Caputo por darme la oportunidad de ser divulgadora. A dos de los mejores ayudantes que tuve, Pedro G. Molinelli y Hernán Reisin. A Andrés Arazi por ser una gran persona y por darme la oportunidad de hacer labo 6 y 7 en TANDAR. Y por aceptar ser jurado de la tesis!
## Contents

Introduction ........................................... 1

1 Atlas Experiment .................................... 3
   1.1 LHC ............................................ 3
   1.2 Atlas detector .................................. 6

2 Jets and their substructure ......................... 17
   2.1 Jet definitions .................................. 17
   2.2 Jet substructure ................................ 22
   2.3 Jet grooming .................................... 27

3 Object and event selection .......................... 31
   3.1 Trigger selection ................................ 31
   3.2 Jet reconstruction ............................... 32
   3.3 Lepton Reconstruction ............................ 33
   3.4 Missing transverse momentum .................... 34
   3.5 Event selection ................................ 34

4 MC simulation and data ............................. 41
   4.1 Stages of the full simulation .................... 41
   4.2 Monte Carlo samples and Data-sets ............... 44
   4.3 Event selection at generator level ............... 45
   4.4 Validation of private MC sample ................. 53

5 Jet substructure and multivariate analysis for W tagging .............................. 59
   5.1 Jet substructure results ......................... 59
   5.2 Multivariate Analysis ............................ 63
   5.3 Training results ................................ 66

6 Conclusions ......................................... 71
Appendices

A PYTHIA plots
  A.1 Signal truth (no pile up) vs signal MC .......................... 75
  A.2 Pile up generated with pythia ..................................... 76
  A.3 W+jets with pile up added: ........................................ 78

B Herwig plots
  B.1 Validation of MC without filters ................................. 81
  B.2 Validation of MC with filters ..................................... 83

C Substructure Results

Bibliography
Introduction

The Large Hadron Collider (LHC) is the largest and highest-energy particle accelerator in the world and was built to take a step forward in our understanding of the Standard Model and new physics beyond that model. It was designed to collide bunches of $10^{11}$ protons 40 million times per second to provide 14 TeV collisions, reaching an unprecedented high energy and luminosity. One of the key experiments at the LHC is ATLAS ("A Toroidal LHC ApparatuS"), a general-purpose detector designed to fulfill precision measurements within the Standard Model and to characterize a wide set of processes covering much of the new phenomena sought to be observed at the TeV scale.

In the LHC, the production of gauge-boson pairs in $pp$ collisions, such as $W$-pairs, provides an important test of the electroweak sector of the Standard Model and represents a major background for Higgs boson production. Furthermore, deviations of the total cross section or in kinematic distributions from predictions would hint at new physics processes like anomalous triple gauge boson interactions or from new resonances decaying to vector bosons.

An electroweak gauge boson can decay to a lepton and neutrino or to quarks (except top quark). When the decay is hadronic, it produces two sprays of hadrons called jets. But for a large enough boost factor, the decay and fragmentation of such boson yields one collimated spray that normally would be reconstructed as a single jet. Therefore, a standard reconstruction method for electroweak bosons would still be possible but less efficient. This is true especially for the case where both $W$ decay hadronically, where the QCD background is largest. One possibility would be to focus on channels where the boosted object decays leptonically, but that may not be optimal for detecting new heavy resonances because such methods discard much of the original signal. This scenario of highly boosted particles in Run I becomes even more frequent for Run II of the LHC. Fortunately, in the last years there’s been a great deal of effort salvaging the boosted hadronic decay channels using jet substructure. In addition, there are jet grooming techniques, which aid in the identification of boosted objects by reducing the smearing effects of jet contamination from underlying event activity, initial state radiation and pile up. Jet substructure used together with these grooming techniques are very promising for enhancing searches for new physics in all-hadronic decay channels.
The subject of this thesis is the improvement of the event selection for $WW$ decaying in the semileptonic channel. Here, one $W$ decaying leptonically is used to tag the events and get rid of QCD while the other $W$ decaying hadronically in the boosted regime further suppresses $W+\text{jets}$ background. Once the boost is very strong the decay products can end up in one jet making the need for substructure analyses. The physics challenge here is to suppress enough of the $W+\text{jets}$ background, where one of the additional QCD jets fakes the boosted hadronic decaying $W$, without losing signal efficiency.

The improvement of the event selection was done at two levels. One level is the offline event selection, where the boosted $W$ decaying hadronically is analyzed using a ‘fat jet’, in order to combine the two collimated quark jets from the $W$, and then look at this jet substructure. Here we focus on the choice of jet variables and cuts, isolation criteria and event vetoes and the choice of topological relations between the final state objects. The other event selection implemented, which represents a large part of the work done for this thesis, was applied at generator level. In this case, we included filters in a Monte Carlo generator, in order to produce Monte Carlo samples which had already a preselection of boosted $W$s. This has several advantages, the most important one being that having a preselection in the generator means that we have a way of producing smaller samples that in the end will have large enough statistics for analysis. In the existing official Monte Carlo samples only 0.4% of the events survives the selection, while in our private production the efficiency is around 10%. We conclude the analysis with a review of the substructure variables and their discriminating power between signal and background.

The structure of the thesis is as follows. In chapter 1 we describe the LHC proton-proton collider, the ATLAS experiment and operation conditions for 2012 data taking. Chapter 2 presents how jets are reconstructed and explains the most common algorithms for this. Next, we describe jet substructure and grooming techniques, of great importance for the analysis of boosted objects. In chapter 3 we give an overview of the object reconstruction and we describe in detail the object and event selection used in the analysis of $WW$ events in the semileptonic channel. Chapter 4 explains briefly the stages for the Monte Carlo full simulation and we present studies done including event selection at generator level using two different MC generators, PYTHIA and HERWIG. In chapter 5 we show the distributions of the most relevant substructure variables obtained after the event selection and show some preliminary results from the application of a multivariate analysis technique to discriminate between signal and background using substructure and grooming variables as input. Finally, the conclusions of the thesis are presented in chapter 6.
1

ATLAS EXPERIMENT

1.1 LHC

The Large Hadron Collider (LHC) is the world’s largest particle collider and the most powerful one, reaching collisions with center of mass of 14 TeV. Originally conceived in the late 1980s and the beginning of 1990s, its construction was approved in 1994 by the European Organization for Nuclear Research (CERN). It was built between 1998 and 2008 in collaboration with over 10000 scientists and engineers from over 100 countries, as well as hundreds of universities and laboratories.

The LHC was built in a circular tunnel of 27 kilometers long and 100 meter underground located in the border of France and Switzerland, near Geneva using the original ring of the existing Large Electron–Positron (LEP) Collider. There are two general purpose experiments in the ring, ATLAS and CMS and two other dedicated experiments, LHCb to study b-quark physics and ALICE to study heavy-ion (Pb-Pb nuclei) collisions (figure 1.1).

The number of processes per second generated in the LHC collisions is given by:

\[ N_{\text{process}} = L_0 \sigma_{\text{process}} \]  

where \( \sigma_{\text{process}} \) is the cross section for a certain physical process (calculated from scattering amplitudes) and \( L_0 \) is the luminosity, that depends only on the beam parameters and can be written as:

\[ L_0 = n_c f_{\text{rev}} \frac{N_1 N_2}{4 \pi \sigma^*_x \sigma^*_y} \]

where \( n_c \) is the number of colliding bunch pairs, \( f_{\text{rev}} \) is the protons revolution frequency, \( N_1 \) and \( N_2 \) are the protons per bunch in beam 1 and beam 2 respectively, and \( \sigma^*_{x,y} \) are the transverse sizes of the beam at the interaction point. The luminosity, and therefore the
Figure 1.1: LHC layout with four interaction points at the site of the four principal experiments.

production rate, increases as bunches have more protons, more bunches crossings occur or if the beams are more collimated at the interaction point (IP) as to reduce the transverse spread or protons.

The nominal luminosity for ATLAS and CMS is $10^{34}$ cm$^{-2}$s$^{-1}$ for beams of approx. 3000 bunches of $10^{11}$ protons each. The proton beams are bunched with a bunch-to-bunch distance of 25 ns (or a multiple of 25 ns). This corresponds to a maximum bunch crossing frequency of 40 MHz [1].

1.1.1 LHC operation during 2012

The ATLAS detector has recorded a total integrated luminosity of 21.3 fb$^{-1}$ of proton-proton collisions at $\sqrt{s} = 8$ TeV during 2012, out of a total of 22.8 fb$^{-1}$ delivered by the
LHC [2]. High luminosity lead to a proportional increase of $pp$ collisions within the same bunch crossing. The average expected number of inelastic collisions is referred to as $\langle \mu \rangle$. This variable follows a Poisson distribution whose mean depends on $L_0$ as a result of equation 1.1. The different LHC conditions such as beam optics and bunch parameters at the interaction point can change across different periods of data-taking, this results in changes of the luminosity of the machine. Therefore, the underlying Poisson distribution can change towards higher or lower values of $\langle \mu \rangle$ as shown in figure 1.2 for 2011 and 2012 data.

The 2012 $pp$ program started effectively on 5 April and continued until December 17. As mentioned before the instantaneous luminosity was on average higher during 2012 than in previous years peaking at $7.7 \times 10^{33}$ cm$^{-2}$s$^{-1}$, close to the design value $1 \times 10^{34}$ cm$^{-2}$s$^{-1}$. The average number of $pp$ collisions per bunch crossing was 20.7, and peaked at more than 40 simultaneous crossings [3].

Normally there’s only one hard-scatter process in each bunch crossing that triggers the event, while the remaining inelastic interactions contribute to soft additional deposition of energy in the detector that blurs the resolution of the hard process of interest, this phenomenon is called pile-up. There are two different forms of pile-up, one comes from the multiple $pp$ collisions that occur in the same bunch crossing, this is known as in-time-pile-up. In-time-pile-up conditions can be estimated by the number of primary vertices, as measured using the tracking system in ATLAS. The other form, out-of-time-pile-up, is due
Figure 1.3: Cut-away view of the ATLAS detector. The sub-systems of the Inner Detector, the Calorimeter and the Muon Spectrometer are shown. The dimensions of the detector are 25 m in height and 44 m in length.

to additional $pp$ collisions occurring in bunch crossings just before and after the collision of interest. As a consequence of the longer electronic read-out windows at certain parts of the ATLAS detector (longer than 25 ns) the detectors are sensitive to several bunch crossings, these collisions can affect the signal of the $pp$ interaction of interest.

1.2 Atlas detector

The ATLAS detector (A Toroidal LHC Apparatus) is a multi-purpose detector that spans over 44 m length and 25 m height, with a total weight of 7000 tons. ATLAS has cylindrical symmetry around the incoming beam and consists of multiple detector sub-systems, as shown in figure 1.3. Next we present a brief description of the detector and it’s subsystems (for a detailed description follow [4]). The goal of the detector design is to achieve the identification of a large spectrum of particles, generated with different energies and with different event topologies. In the design and construction of each of its parts it was necessary to take into account the demanding conditions of operation: as it was mentioned
before, the LHC was designed to deliver bunch crossing of $10^{11}$ protons every 25 ns, with a maximum energy of 14 TeV and a luminosity of $10^{10}$ fb$^{-1}$.

The ATLAS detector coordinated system is used to describe the position of particles as they traverse these sub-detectors. The origin of coordinates $x = y = z = 0$ corresponds to the nominal point of interaction. The direction of the protons beam, coinciding with the axis of the cylinder, defines the $z$ axis, and the plane $xy$ is transverse to the beam direction. The positive $x$-axis is defined as pointing from the interaction point to the center of the LHC ring and the positive $y$-axis is defined as pointing upwards. The azimuthal angle $\phi$ is measured around the beam axis, and the polar angle $\theta$ is the angle from the beam axis. The pseudorapidity is defined as $\eta = -\ln(\tan \frac{\theta}{2})$. For massive objects like jets, the rapidity is defined as $y = \frac{1}{2} \ln \left( \frac{E + p_z}{E - p_z} \right)$. The transverse momentum $p_T$, the transverse energy $E_T$, and the missing transverse energy $E_T^{\text{miss}}$ are defined in the $xy$ plane. The distance $\Delta R$ in the pseudorapidity-azimuthal angle space is defined as $\Delta R = \sqrt{\Delta \eta^2 + \Delta \phi^2}$.

ATLAS consists of three primary detection systems layered radially around the beam pipe to have maximum coverage in $\eta$ and $\phi$. From inside out these sub-detectors are: a central inner tracker embedded in a 2 T solenoidal magnetic field for measurement of position and momentum of charged particles, a calorimeter system for energy measurement of both neutral and electrically charged particles, and a muon spectrometer located within a large toroidal magnetic field to measure the position and momentum of muons. The wide coverage in $\eta$ is achieved by placing part of these detectors in two caps, one at each end of the cylinder or central barrel. Also, ATLAS has a sophisticated data acquisition and trigger system that allows to store the most interesting 200 events in the 400 million events that are generated per second.

1.2.1 Inner detector

The goal of the Inner Detector (ID) is to make precision measurements of the tracks and momentum of charged particles that emerge from the interaction point in an event. It allows to reconstruct the primary interaction vertex and the secondary vertices if there are any. The inner detector is immersed in a 2 T magnetic field generated by the central solenoid (of approximately 6 m long and 1 m radius), whose magnetic field bends the trajectories of the charged particles and allows to measure direction and momentum for particles having nominally $p_T \gtrsim 500 \text{ GeV}$ within the acceptance pseudorapidity range $|\eta| < 2.5$. The impact parameter can be measured with a 10 $\mu$m resolution. These measurements are made combining three different technologies, the Pixel detector (Pixel), the SemiConductor Tracker (SCT) and the Transition Radiation Tracker (TRT). Figure 1.4 shows diagrams of the inner detector and its components.

The silicon pixel tracker is the innermost part of the tracking system built with semiconductor materials and arranged in three layers in the barrel and three additional disks at
Figure 1.4: a) Scheme of the ATLAS Inner Detector layout. b) ATLAS inner detector barrel being crossed by one high-energy particle.
each end-cap. The physical principle for the Pixel detector is as follows: ionizing particles traversing the semiconductor produce electron-hole pairs whose currents are readout by the front-end chips, providing a binary response: a “hit” is registered only if the pulse height in a channel exceeds a preset threshold. Pixels have 50 µm by 400 µm in size each and are organized in identical modules, 1744 in total, each covering an area of $19 \times 63 \text{mm}^2$. This results in a total of 80 million silicon pixels. In the barrel section pixels are oriented parallel to the beam and have resolutions of 10 µm in the $R - \phi$ plane and 115 µm in the $z$ axis. In the end-cap region, the modules are arranged radially and have a resolution of 10 µm in the $R - \phi$ plane and 115 µm in the $R$ radial direction. The central layer, located approximately at 5 cm from the beam pipe center, is called b-layer and is fundamental to the reconstruction of secondary vertices used in the identification of jets derived from heavy-flavour quarks (bottom and charm quarks).

The SemiConductor Tracker (SCT) has eight strip layers to provide eight precision measurements per track. The physical principle is the same as for the Pixel detector. There are 4088 SCT modules, each composed of two small-angle (40 mrad) stereo strips to provide two-dimensional hit localization, with one set of strips in each layer parallel to the beam direction, measuring $R - \phi$ in the barrel region and in the end-cap region one set of strips runs radially with a set of stereo strips with a 40-mrad angle. The SCT is divided into four concentric barrel layers ($|\eta| < 1.1$) and nine end-cap disks ($1.1 < |\eta| < 2.5$) at each side of the barrel. In the barrel each module has a resolution of 17 µm in the $R - \phi$ plane and 580 µm in the $z$ axis. In the end-cap region the resolution is 17 µm in $R - \phi$ and 580 µm in $R$.

The transition radiation tracker surrounds the silicon detectors providing measurements of charged particles up to $|\eta| < 2.0$. The TRT is comprised of strawtubes of 4 mm diameter, that are made of metal straws (cathode) filled with an ionizing gas mixture of xenon, oxygen and CO$_2$, with a wire (anode) running down the center of the straw. The charged particle travels through the gas producing positive ions and free electrons, which travel to the cathode and anode, respectively, under the influence of an applied voltage of 1600 V. This produces a current pulse in the wire, where electrons are collected. This method gives no information on the position along the length of the strawtube. The combination of many wires with signals creates a pattern of straw hits that allow the path of the particle to be determined (tracker). The transition radiation is generated when a charged particle passes through the interface between two materials of different dielectric constants. Polypropylene sheets are placed between the strawtubes to produce this radiation, which allows to discriminate electrons from hadrons, since the signal produced by the electrons is greater than the less ionizing particles.

The TRT only provides information in the $R - \phi$ direction and has a resolution of 130 µm. In the barrel there are approximately 52544 strawtubes oriented parallel to the $z$ axis, each of them 144 cm long. In the end-cap region there are 12288 straws, each
of them 37 cm long, arranged radially in wheels. Particles usually traverse an average of 36 tubes, providing continuous tracking to improve the pattern recognition and the momentum resolution (a schematic of the track reconstruction is shown in figure 1.5).

1.2.2 Calorimeter system

The calorimeter is the detector used for jet reconstruction. It is symmetric in $\phi$, covers a large pseudorapidity range of $|\eta| < 4.9$ ($0.5^\circ \lesssim \theta \lesssim 179.5^\circ$) and is composed of different sub-detectors. It’s divided mainly in two parts, electromagnetic calorimeter (EM) and hadronic calorimeter (figure 1.6 shows the layout of the calorimeter with each of its sub-regions). Over the $\eta$ region that coincides with the Inner Detector, the EM calorimeter granularity is ideal for precision measurements of photons and electrons, while in the rest of the detector, the coarser granularity is sufficient to satisfy the requirements for jet reconstruction and missing energy ($E^\text{miss}_T$) measurements.

The calorimeters are devices that measure the energy of particles through total absorption. The ATLAS calorimeter system is a non-compensating sampling calorimeter. The term “sampling” means that the detector utilizes intercalated layers of absorber material and active medium. When a charged particle crosses the active medium, they originate a shower of decreasingly lower-energy as consequence of the interactions with the material. These can be periodically absorbed in a high-density material and sampled using generally ionization or scintillation light proportional to the number of particles in the active medium, thus inferring the energy of the original particles from this measurement. “Non-compensation” means that part of the energy of nuclear collisions between the high-energy particles from the event and the detector material nuclei is lost in the form of nuclear recoil or fission and therefore remains “invisible” to the active readout. A consequence of this is that the energy response for hadronic particles (interacting strongly with the nuclei) is smaller than for particles of the same energy that interact predominantly by electromagnetic forces.
The innermost calorimeters cover the region closest to the beam, and use liquid argon (LAr) as the active medium. These are housed in three cryostats, one barrel and two end-caps. The barrel cryostat contains the lead/LAr electromagnetic barrel calorimeter, whereas the two end-cap cryostats each contain the lead/LAr electromagnetic end-cap calorimeter, the copper/LAr hadronic end-cap calorimeter, and the copper-tungsten/LAr forward calorimeter. The outermost detector is the steel/scintillator-tile hadronic calorimeter, divided in three parts: the central barrel and the two extended barrels.

**Electromagnetic Calorimeter**

The electromagnetic calorimeter covers two regions, the barrel in $|\eta| < 1.48$ and the end-caps covering an $\eta$ range $1.375 < |\eta| < 3.2$. The EM calorimeter uses lead as absorber and LAr as sampling and has an accordion shape to provide full coverage in $\phi$ and to minimize gaps. A photon traveling through the absorber will interact with the heavy nucleus via Compton scattering or photo-electric effect, producing low-energy electrons or producing electron/positron pairs. An electron or positron, in turn, can produce bremsstrahlung
photons or produce more charged particles via ionization. Therefore, incident photons, electrons or positrons produce a shower of photons, electrons and positrons that lose their energy through successive interactions in the absorber. The produced particles ionize the liquid argon, and the charge is collected by electrodes located in the liquid argon gap. These electrodes consist of three layers of copper sheets, the two exterior layers provide a high-voltage potential and the inner layer is used to readout the signal. Figure 1.7 shows the accordion-shaped absorber with the different layers and granularities.

The primary layer of the electromagnetic barrel calorimeter (EMB) has a fine segmentation ($\Delta \eta \times \Delta \phi = 0.003 \times 0.1$) that allows a discrimination between photons and neutral pions. A second layer with a segmentation of $\Delta \eta \times \Delta \phi = 0.025 \times 0.025$ collects the largest fraction of the energy of the electromagnetic shower (around 80%) and because of its fine segmentation it permits a detailed mapping of the electromagnetic shower. The third layer has half of the granularity of the previous layer and measures the tail of the electromagnetic shower. The EMB is housed inside the same cryostat that holds the inner detector’s solenoid magnet. In the region $|\eta| < 1.8$ a presampler calorimeter corrects for the energy lost in the cryostat material.

In the two electromagnetic end-cap calorimeters (EMEC) the accordion waves are parallel to the radial direction and the granularity is slightly larger than in the barrel.
The signal readout chain for all the calorimeter systems is divided into a fast analog readout for the trigger system and a slower digital readout used for more redefined trigger decisions and the offline reconstruction. Groups of cells called trigger towers defined by the sum of readout channels within $\Delta \phi \times \Delta \eta = 0.1 \times 0.1$ are inputs for rapid trigger evaluation (but without resolution along the direction in which the shower develops).

**Hadronic Calorimeter**

As it was mentioned before, the hadronic calorimeter is located outside the EM calorimeter. The barrel region is called Tile calorimeter and uses iron as absorber, with scintillating tiles intercalated. The hadronic enc-cap calorimeter has a copper absorber and the forward calorimeter has a tungsten absorber, with liquid argon as the active material.

The scintillating tiles in the tile calorimeter are arranged to lie parallel to the incoming particle direction. It’s divided in the barrel calorimeter covering $|\eta| < 1.0$ and two extended barrel calorimeters, covering the range $0.8 < |\eta| < 1.7$. The readout fibers of several tiles are grouped to a photomultiplier tube forming cells in $\eta \times \phi$ space. These cells are divided in three layers, the first two of size $\Delta \eta \times \Delta \phi = 0.1 \times 0.1$ and the last one of size $\Delta \eta \times \Delta \phi = 0.2 \times 0.1$.

The hadronic end-cap calorimeter (HEC), located at $1.5 < |\eta| < 3.2$, has a flat-plate design, unlike the accordion of the EMB or the EMEC, which also uses LAr as the active medium but has copper as the absorber material.

The forward calorimeters (FCal) are located at each end-cap cryostat covering the high pseudorapidity region $3.1 < |\eta| < 4.9$. Since it is the only calorimeter that covers this very forward region, it must provide both electromagnetic and hadronic measurements. The forward calorimeters at each side is split into three modules of 45 cm depth in $z$, an innermost electromagnetic layer is composed by copper/LAr modules, and two outermost hadronic layers are made of cooper-tungsten/LAr modules. Cooper plates facilitate heat interchange, while the tungsten absorbers minimize the lateral spread of hadronic showers.

### 1.2.3 Muon spectrometer

The ATLAS Muon spectrometer is based on the magnetic deflection of muon tracks in the large superconducting air-core toroid magnets and has two main functions: provide high-precision tracking measurements and trigger. A system consisting of 8 air-core toroidal magnets in the central barrel plus one smaller magnet in each end-cap, provides a toroidal magnetic field designed to bend the track of charged particles in the $\eta$ direction. Given that the spectrometer is located outside the calorimeter, the only charged particles that travel through are muons. The layout of the muon spectrometer is shown in figure 1.8. The spectrometer consists of four systems: two precision muon trackers, the monitored drift
Figure 1.8: Overview of the Atlas muon spectrometer components. The primary sub-detectors are shown: Monitored drift tubes (MDT), Cathode strip chambers (CSC), Resistive plate chambers (RPC) and Thin gap chambers (TGC).

The precision momentum measurement is performed mostly by the MDTs, that span over the range $|\eta| < 2.7$, except in the innermost end-cap layer where their coverage is limited to $|\eta| < 2.0$. Figure 1.9 shows the cross-section of the muon system in the bending plane (containing the beam axis) and shows some $\eta$ regions. In the barrel region, the monitored drift tubes are arranged in chambers located between the eight coils of the barrel toroid magnet in three cylindrical layers around the beam axis and in the transition and end-caps, the chambers are installed also in three layers perpendicular to the beam. The MDTs consists of drift tubes with a diameter of approximately 30 mm, operating with Ar/CO$_2$ gas (93/7 %) at 3 bar pressure (this mixture was selected because of its aging properties). The electrons that result from ionization due to crossing muons are collected at the central wire acting as an anode. The maximum drift tube time from the wall to the wire is about 700 ns, when operated at a nominal potential of 3080 V.
Figure 1.9: Cross-section of the muon system in a plane containing the beam axis (bending plane).

The particle fluxes and muon track density are highest in the forward direction ($2 < |\eta| < 2.7$) and the MDTs operation counting rate of about 150 Hz/cm$^2$ is exceeded in this region. Instead, Cathode Strip Chambers are used in the inner most tracking layer due to their higher rate capability and time resolution (CSC’s are safe to use in counting rates up to 1000 Hz/cm$^2$, which is sufficient up to the forward boundary of $|\eta| = 2.7$). The CSCs are multi-wire proportional chambers with cathode planes segmented into strips in orthogonal directions. This allows both coordinates to be measured from the induced-charge distribution. The resolution of the MDT chambers is 35 µm in the bending plane $R - z$ and the resolution of the CSC is 40 µm in the bending plane $R - z$ and 5 mm in the transverse coordinate.

The trigger system covers the pseudorapidity range $|\eta| < 2.4$. In the barrel region ($|\eta| < 1.05$) are used Resistive Plate Chambers and in the end-cap regions ($1.05 < |\eta| < 2.4$) are used Thin Gap Chambers (TGC). Both technologies deliver signals within 15 – 25 ns, thus providing the ability to tag the beam-crossing and allow the first level trigger to recognize muon multiplicity and approximate energy range. The trigger chambers measure both coordinates of the track, one in the bending ($\eta$) plane and one in the non-bending ($\phi$) plane. The RPC is a gaseous parallel electrode-plate detector (no wire), formed by two
resistive plates, kept parallel at a distance of 2 mm by insulating spacers. The electric field between the plates of about 4.9 kV/mm allows avalanches to form along the ionizing tracks towards the anode. TGCs are multi-wire proportional chambers with the characteristic that the wire-to-cathode distance of 1.4 mm is smaller than the wire-to-wire distance of 1.8 mm. The high electric field around the TGC wires and the small wire-to-wire distance lead to a very good time resolution of 4 ns.

The resolution of the RPC’s is 10 mm in the bending plane and in the transverse coordinate and the resolution of the TGC is approximately 2 – 6 mm in the bending plane and 3 – 7 mm in the transverse plane. TGCs provide two functions in the end-cap muon spectrometer, it has a trigger function and also the determination of the second, azimuthal coordinate to complement the measurement of the MDTs in the bending (radial) direction.

1.2.4 Forward detectors

There are also three smaller detector systems in the ATLAS forward region. The main function of the first two systems is to determine the luminosity delivered to ATLAS. The first detector is LUCID (LUMinosity measurement using Cerenkov Integration Detector), located at a distance of 17 m from the interaction point, at both ends of the detector. It detects inelastic $p - p$ scattering in the forward direction and is the main online relative luminosity monitor for ATLAS. The second detector is ALFA (Absolute Luminosity For ATLAS), located at 240 m from the interaction point at each end of the detector. It consists of scintillating fibre trackers located inside Roman pots which are designed to approach as close as 1 mm to the beam. The third system is Zero-Degree Calorimeter (ZDC), located at 140 m from the interaction point at each end of the ATLAS detector. It has a role in determining the centrality of heavy-ion collisions and consists of layers of alternating quartz rods and tungsten plates which measure neutral particles at pseudorapidities $|\eta| \geq 8.2$. 
Jets and their substructure

In the framework of Quantum Chromodynamic theory (QCD) quarks and gluons emerge from collisions as constituents of final state “colorless hadrons” due to confinement. This group of particles tend to move collinearly with the direction of the original partons, which results in a collimated shower of hadrons (and also photons and leptons) that enter the detector, instead of the partons produced in the primary interaction. This group of collimated particles are what we refer as jets.

Jets are fundamental in LHC physics analysis: both theory and experimental results are often presented in terms of jet cross sections, and therefore they provide the meeting point between the two. Jets are an input to practically all physics analysis, from Higgs searches, top physics, new physics searches and many more [5].

In section 2.1 we explain several jet reconstruction algorithms and in sections 2.3 and 2.2 we present jet grooming techniques and substructure variables, both very important tools for the analysis of boosted objects.

2.1 Jet definitions

It’s important to emphasize that there’s not a univocal correspondence between a jet and a parton and neither is there a unique definition of a jet. Figure 2.1 shows $e^+e^-$ events, it’s easy to interpret the event in the left as $e^+e^- \rightarrow q\bar{q}$, where there’s a soft and collinear showering followed by transition hadrons, forming two jets. The middle event is a slightly more complicated case, where energy flow has more than two jets. One possible interpretation of this event is that a $q\bar{q}$ pair has emitted a very energetic gluon $g$ and all three undergone a soft and collinear showering. Other possible interpretation can be seen in the right picture, where the $q\bar{q}$ pair has emitted two hard gluons, which in turn undergoes
soft and collinear showering. Choosing between these two possibilities means selecting different parameters such as separation in angle and how energetic the emissions have to be in order to form a separate jet. Because of this, the jet definition may change depending on the specific application (for example, particular signatures) or experimental conditions (high/low luminosity, high/low center-of-mass energy, etc) in order to reconstruct more accurately the hard scattering process. However, jet algorithms must comply with some fundamental conditions [7]:

- they should be simple to implement in an experimental analysis and theoretical calculations;
- yield reliable results at all orders in perturbation theory;
- relatively insensitive to hadronization;
- infrared and collinear (IRC) ‘safe’ (explained below);
- experimentally ‘safe’: not strongly affected by contamination from hadron remnants and other underlying soft events (pile-up), or detector conditions (e.e, noise).

A jet is infrared safe and collinear (IRC-safe) when it doesn’t change significantly if new soft or collinear particles are added. This is illustrated in figure 2.2. The combination of particles inside a jet is known as ‘recombination scheme’, from which the most common is the E-scheme where one adds the 4-vectors of the particles, resulting in jets that are massive.

Traditionally, jet algorithms have been classified into two categories: cone and sequential recombination algorithms.

Figure 2.1: $e^+e^- \rightarrow q\bar{q}$, resulting in two jets; middle: the event can be interpreted as $e^+e^- \rightarrow q\bar{q}g$, having a 3-jet structure; right: the same event reinterpreted as $e^+e^- \rightarrow q\bar{q}gg$ resulting in a 4-jet structure. (Taken from [6])
Cone algorithms

Cone algorithms are based on the collinear nature of gluon radiation and the parton shower. The decay products of quarks and gluons and their emissions will tend to form a cone of particles as they propagate outwards. Cone algorithms aim to maximize the amount of energy inside a stable cone of fixed radius.

The standard jet finder is iterative. First, it sorts all particles according to their momentum, and identifies the one with largest transverse momentum. This particle is referred to as seed. Then, a cone of radius $R$ in $\eta - \phi$ is drawn around the seed and all objects within a cone of $\Delta R^2_{is} = (y_i - y_s)^2 + (\phi_i - \phi_s)^2$ are combined with it, where $y_i$ and $\phi_i$ are the rapidity and azimuth of the particle $i$, and $y_s$ and $\phi_s$ are the rapidity and azimuth of the seed. The direction of the sum of the momenta of the particles is calculated and if it doesn’t coincide with the seed direction, then the sum is used as a new seed direction, and it iterates until the direction of the cone is stable, that is until the direction of the sum of the cone contents coincides with the previous seed. The resulting cone is called a jet and the particles inside it are then removed from the list of particles in the event. With the remaining particles this procedure is repeated until no particles remain, obtaining the complete set of jets.

The main problem with this approach is the use of the transverse momentum of the particles to choose which particle is the seed, because $p_T$ of particles are not collinear safe quantities. This means that another, less hard particle, pointing in a different direction now becomes the hardest in the event, leading to a different set of jets. There are many other variants of cone algorithms, but most of them aren’t collinear or infrared safe.

One very popular cone algorithm used mostly at the Tevatron doesn’t use an initial seed but instead uses all particles as possible seeds, calculating all the possible stable cones.
With the list of all stable cones found from all possible seeds a ‘split-merge’ procedure is used to assign particles that appear in multiple cones to one specific jet. This method is collinear safe because it avoids using the order in $p_T$ or particles as seed but is not infrared safe, because adding an extra soft particle creates a new seed that can lead to an extra stable cone, altering the final set of jets.

A fix to this problem is a seedless algorithm where the addition of one or more soft particles doesn’t lead to new hard stable cones. This cone algorithm is called Seedless Infrared Safe Cone (SISCone, more details about this can be found in [8]).

**Sequential recombination algorithms**

Cluster-type jet finders are generally based on successive pair-wise recombination of particles, have simple definitions and are all infrared safe.

Sequential recombination algorithms start by assigning a distance to all input objects and all pair-wise combinations of those objects. Distance between two input entities $i$ and $j$ ($d_{ij}$) and between the entity $i$ and the beam ($d_{iB}$) are defined as

$$d_{ij} = \min (p_{T i}^{2p}, p_{T j}^{2p}) \frac{\Delta^2_{ij}}{R^2}$$

$$d_{iB} = p_{T i}^{2p}$$ (2.1)

where $\Delta^2_{ij} = (y_i - y_j)^2 + (\phi_i - \phi_j)^2$, $p_{Ti}$ corresponds to the transverse momentum, $y_i$ to the rapidity and $\phi_i$ to the azimuthal angle. The parameter $p$ governs the relative power of the energy versus geometrical scales of $\Delta^2_{ij}$.

The formulation of a sequential recombination algorithms is as follows:

1. Utilize the particle distance metric defined in Eq. 2.1.

2. Compute the minimum $d_{ij}$, $d_{min} = \min (d_{ij})$ among all particles

3. If $d_{min} < d_{iB}, d_{jB}$, then recombine particles $i$ and $j$ and repeat from step 1.

4. If $d_{ij} > d_{ib}$, then identify $i$ as a jet and remove it from the list.

5. Continue until all particles are considered jets or have been clustered with other particles.

For $p = 1$ we can use the inclusive $k_T$ algorithm, and for $p = 0$ the Cambridge-Aachen (C/A) algorithm, both aim to reverse the shower history. The distance used for the $k_T$ algorithm induces soft particles to be merged first, which introduces a strong sensitivity to small fluctuations of the energy density of the parton shower. The C/A algorithm relies solely on angular ordering of emissions in order to reconstruct the shower by omitting
Figure 2.3: A sample parton-level event, generated using Herwig (see section 4.1), illustrating the jets produced with the 4 most common IRC-safe algorithms, showing the degree of regularity (or not) of the boundaries of the resulting jets and their extents in the rapidity-azimuth plane (taken from [6]).

the transverse momentum from $d_{ij}$, which again results in a strong dependence on the experimental conditions. A case of great interests is when $p = -1$, this is called the anti-$k_T$ algorithm. It first clusters hard objects together which results in more regular jets with respect to the $k_T$ and C/A algorithms.

The $k_T$ jet algorithm has several benefits, for example, it mimics a walk backwards through the QCD branching sequence, which means that the reconstructed jets collect most of the particles radiated from an original hard parton. This allows for better mass measurements and general kinematic reconstruction, to mention a few advantages. Also, the $k_T$ algorithm allows to decompose a jet into constituent subjets, which is useful for identifying decay products of fast-moving heavy particles [9].

The jet algorithms described here and many more, are implemented in the Fastjet [9] software package for jet-finding.


2.2 Jet substructure

Because of the high center of mass energy of the LHC, highly boosted electroweak bosons and/or top quarks are produced. With a high enough boost, the decay and fragmentation of a boosted object yields a collimated spray of hadrons which a standard jet algorithm would reconstruct as a single jet. Therefore, the standard methods to reconstruct electroweak bosons and top quarks become ineffective. A relatively recent approach to solve this problem is using jet substructure techniques. There are several methods that use information from the jet clustering procedure to extract the internal structure of jets, therefore successfully distinguishing between jets that originate from boosted electroweak boson and top quarks (“W jets”, “top jets”, etc) and those originating from light quarks or gluons (“QCD jets”). Jet shape methods are able to efficiently tag boosted objects using jet-based observables, taking advantage of the different energy flow in the decay pattern of signal and background jets.

2.2.1 N-subjettiness

Taking advantage of the multi-body kinematics in the decay pattern of boosted hadronic jets, N-subjettiness, called $\tau_N$, “counts” the number of subjets in a given jet. Boosted hadronic objects have a fundamentally different energy pattern than QCD jets that have comparable invariant mass. For example, $W$ decays to two quarks, therefore, a single jet corresponding to a boosted $W$ should be composed of two distinct hard subjets with a combined mass similar to $W_M$ (around 80 GeV). However, boosted QCD jets with an invariant mass of 80 GeV usually originates from a single hard parton and acquires mass through large angle soft splittings. The goal of N-subjettiness is to exploit this difference in expected energy flow by ‘counting’ the number of hard lobes of energy within a jet. The magnitude $\tau_2/\tau_1$ is an optimal discriminating variable to identify two-prong objects like boosted $W$, $Z$ and Higgs bosons and $\tau_3/\tau_2$ is effective to identify three-prong objects like boosted quarks [10].

The N-subjettiness algorithm can be summarized in the next three steps:

1. First, one reconstructs a candidate $W$ jet using a jet algorithm. In this work we used a C/A jet with $R = 1.2$ as described in section 3.5.

2. Then, one identifies $N$ candidate subjets using the exclusive $k_t$ clustering algorithm, forcing it to return exactly $N$ jets. The reclustering is done only with the constituents of the C/A jet as input and the election of the kT algorithm is because it provides a QCD history while, for example, an Anti-Kt algorithm doesn’t.
3. With these candidate subjets, one calculates $\tau_N$ as follows

$$\tau_N = \frac{1}{d_0} \sum_k p_{T,k} \min(\Delta R_{1,k}, \Delta R_{2,k}, \ldots, \Delta R_{N,k})$$ (2.3)

where $k$ runs over the constituent particles in a given jet, $p_{T,k}$ are their transverse momenta and $\Delta R_{J,k} = \sqrt{(\Delta \eta)^2 + (\Delta \phi)^2}$ is the distance in the rapidity-azimuthal plane between candidate subjet $J$ and a constituent particle $k$. The normalization factor $d_0$ is calculated as follows

$$d_0 = \sum_k p_{T,k} R_0$$

where $R_0$ is the characteristic jet radius used in the original jet clustering algorithm.

It is easy to see that $\tau_N$ quantifies to what degree a particular jet can be regarded as a jet composed of $N$ subjets. Jets with $\tau_N \approx 0$ have all their radiation aligned with the candidate subjet directions and therefore have $N$ (or fewer) subjets while jets with $\tau_N >> 0$ have a large fraction of their energy distributed away from the candidate subjet directions, and therefore have at least $N + 1$ subjets. Figure 2.4 shows $\tau_1$ and $\tau_2$ plots comparing $W$-jets and QCD jets.

Contrary to what one would think, $\tau_2$ by itself is not a useful variable to identify $W$ jets, because QCD jets can also have a small $\tau_2$ as shown in the right plot of figure 2.4. Similarly, though $W$ jets are likely to have a large $\tau_1$ value, QCD jets with a diffuse spray of large angle radiation can also have a large $\tau_1$, as shown in figure 2.4 on the left. However, those QCD jets with a large $\tau_1$ usually also have large values of $\tau_2$. Therefore, as it was mentioned at the beginning, is the ratio $\tau_2/\tau_1$ the one that has a high discriminating power. Figure 2.5 shows the ratio between $\tau_2$ and $\tau_1$ for $W$ and QCD jets.

![Figure 2.4: Distributions of $\tau_1$ (left) and $\tau_2$ (right) for boosted $W$ and background. Neither one offers a good discriminating power for boosted objects.](image)
Figure 2.5: Distribution of $\tau_2/\tau_1$ for boosted W and background. Here, the $\tau_2/\tau_1$ ratio gives a considerable separation between W jets and QCD jets.

Jet mass

The jet mass $M$ is calculated from the energies and momenta of its constituents (particles or clusters) as follows:

$$M^2 = \left( \sum_i E_i \right)^2 - \left( \sum_i \vec{p}_i \right)^2$$

(2.4)

where $E_i$ and $\vec{p}_i$ correspond to the energy and momentum of the constituent $i$. The standard ATLAS reconstruction is the following: clusters have their masses set to zero, while Monte Carlo particles are assigned their correct masses.

Jet width

The jet width is a very straightforward observable. It’s defined as:

$$Width = \frac{\sum_i \Delta R_i p_{T_i}}{\sum_i p_{T_i}}$$

(2.5)

where $\Delta R_i = \sqrt{(\Delta \phi_i)^2 + (\Delta \eta_i)^2}$ is the distance from the jet axis in the rapidity-azimuth plane of the $i$ jet constituent and $p_{T_i}$ is the constituent $p_T$ with respect to the beam axis. A low value of width for a jet means that most of its constituents are close to the jet axis. Therefore a W jet, for example, would have a higher value of jet width.

$K_T$ splitting scales

The $k_t$ splitting scales observables are very straightforward to calculate. Given a certain jet, in this work a C/A jet with $R = 1.2$, the constituents are reclustered using a $k_t$ algorithm,
which as it was mentioned before tends to combine the harder constituents last. At the final step of the jet recombination, the $k_t$ distance definition $d_{ij}$ between the two remaining subjets is used to define a “splitting scale” variable as:

$$\sqrt{d_{ij}} = \min(p_T, p_{T_j}) \Delta R_{ij}$$

(2.6)

where $\Delta R_{I,k}$ is the distance in the rapidity-azimuth plane. Using this definition, the subjets of the last step of the reclustering provide the $\sqrt{d_{12}}$ observable, while the splitting scale of the second-to-last step of the reclustering provides the $\sqrt{d_{23}}$ observable. The parameters $\sqrt{d_{12}}$ and $\sqrt{d_{23}}$ can then be used to distinguish heavy-particle decays, which tend to be reasonably symmetric, from the largely asymmetric splittings that originate from QCD radiation in light-quark or gluon jets. The expected value for a two-body heavy-particle decay is approximately $\sqrt{d_{12}} \approx m_{\text{particle}} / 2$, whereas jets from the parton shower of gluons and light quarks tend to have smaller values of the splitting scales and to exhibit a steeply falling spectrum for both $\sqrt{d_{12}}$ and $\sqrt{d_{23}}$.

$z_{\text{cut}}$

The substructure variable $z_{\text{cut}}$ is closely related to $k_t$ splitting scales, and it takes into account the energy sharing between the two subjets [11]. $z_{\text{cut}}$ is defined as follows:

$$z_{\text{cut}} \equiv \frac{d_{\text{cut}}}{d_{\text{cut}} + Q^2_M} \rightarrow \min(E_A, E_B) \frac{E_A + E_B}{E_A + E_B}$$

(2.7)

where $d_{\text{cut}}$ is the variable defined in eq. 2.6, $Q_M$ corresponds to the mass of the fat jet and $E_A$ and $E_B$ are the energies of the two subjets. Larger values of $z_{\text{cut}}$ are expected for boosted objects, and smaller values for softer objects.

**Dipolarity**

Jet dipolarity, $D$, is based on color flow. The distribution of radiation in an event doesn’t only depend on the hard parton collision, it also depends on the color flow of an event. One expects a color-connection of particle strings to form between two partons given place to a multitude of particles in the space between them. In the case of the $W$, where the color connection exists between the two quarks, one assumes a preference to radiate color particles in between the two decay products [12].

In order to calculate $D$, first consider a jet $J$, with two subjets, $j_1$ and $j_2$, whose centers are located at $(\eta_1, \phi_1)$ and $(\eta_2, \phi_2)$, respectively. For each calorimeter cell $(\eta_i, \phi_i)$ with transverse momentum $p_{T_i}$ let $R_i$ be the distance in the rapidity-azimuth plane between $(\eta_i, \phi_i)$ and the segment that runs from $(\eta_1, \phi_1)$ to $(\eta_2, \phi_2)$. Dipolarity is calculated as follows

$$D = \frac{1}{R_{12}^2} \sum_{i \in J} \frac{p_T}{p_T} R_i^2$$

(2.8)
where $R_{12}$ is the distance of the two subjets in the rapidity-azimuth plane. If the algorithms used to reconstruct $J$, $j_1$ and $j_2$ are infrared and collinear safe (IRC), then dipolarity is also IRC. Dipolarity is small when most of the radiation within $J$ occurs in the region between $j_1$ and $j_2$ and will be large whenever a considerable amount of radiation is elsewhere. Dipolarity receives large contributions from semi-soft radiation away from the cores of $j_1$ and $j_2$, this semi-soft radiation is expected to reflect the color configuration of $J$. One expects that color singlets that decay into two jets will have small $D$, while radiation emitted by colored objects will tend to yield larger values of $D$.

**Angularity**

Angularity, called $\varsigma$, is an observable sensitive to the symmetry in the energy flow inside a jet [13]. The general formulation for angularity is

$$\varsigma_a = \frac{1}{M} \sum_i E_i \sin \theta_i^2 [1 - \cos \theta_i]^{1-a}$$

(2.9)

where $a$ is a free parameter that depending of its value can emphasize radiation near the edges ($a < 0$) or core ($a > 0$) of the jet, $M$ is the jet mass, $E_i$ and $\theta_i$ are the energy and angle with respect to the jet axis of the $i$ particle, respectively.

In the limit of small-angle radiation ($\theta_i << 1$), $\varsigma$ can be approximated by

$$\varsigma_a \simeq \frac{2^{a-1}}{M} \sum_i E_i \theta_i^{2-a}$$

(2.10)

Angularities are infrared-safe for $a \leq 2$. In this work, the value chosen is $a = -2$. The resulting $\varsigma_{-2}$ observable can serve as a discriminator between QCD jets and boosted particle decays by virtue of the broader tail expected in the QCD distribution.

**Planar Flow**

This observable characterizes the geometric distribution of energy deposition from a jet. Essentially planar flow measures how the jet’s energy is evenly spread over the plane across the face of the jet (high planar flow) versus spread linearly across the face of the jet (small planar flow) [13]. To calculate planar flow, for a given jet one first constructs a two dimensional matrix

$$I^{kl}_E = \frac{1}{M} \sum_i \frac{1}{E_i} p_{i,k} p_{i,l}$$

where $M$ is the jet mass, $E_i$ is the energy of the $i$ constituent of the jet and $p_{i,k}$ and $p_{i,l}$ are the $k$ and $l$ components of its transverse momentum calculated with respect to the jet axis. Planar flow is the defines as follows:

$$P_f = \frac{\det(I_E)}{Tr(I_E)^2} = \frac{\lambda_1 \lambda_2}{(\lambda_1 + \lambda_2)^2}$$

(2.11)
where $\lambda_{1,2}$ are eigenvalues of $I_E$. Small or close to zero values of planar flow ($P_f \rightarrow 0$) corresponds to a light energy deposition, as in the case of a two-pronged decay (like the $W$ jet for example), while values close to one ($P_f \rightarrow 1$) corresponds to completely isotropic energy distributions. This is the general case of QCD jets.

### 2.3 Jet grooming

Grooming algorithms include filtering (usually used together with the mass drop method), pruning and trimming. Each of these algorithms has two adjustable parameters that can vary to maximize the significance (signal over background ration). There are many previous works where these parameters have been studied in detail (for example, [14], [15], [16]). After jet grooming, the jet mass is always shifted lower, with the mass of signal jets around the $W$ mass and background jets concentrated around much lower values. This is shown in figure 2.6. Because of this, one can apply a mass window cut to efficiently reduce the number of background. This is explained in the event selection in section 3.5. Essentially all these grooming techniques are qualitatively similar but differ in details, as explained next.
Figure 2.7: Mass-drop filtering method. The top diagram shows the mass-drop and symmetric splitting. The bottom diagram shows the filtering procedure where a filtered jet is reconstructed using the three hardest subjets reclustered with the $R_{filt}$ parameter (taken from [16]).

2.3.1 Filtering

Filtering with mass drop [14]: Given a ‘fat’ jet jet with a recombination parameter $R$, an iterative decomposition algorithm is applied on the jet as follows:

1. For jet $j$ undo the last step of jet clustering. The two resulting jets $j_1$ and $j_2$ are ordered in mass such that $m_{j_1} > m_{j_2}$;

2. If a significant mass drop is found and the splitting is not too asymmetric then exit the loop. The significant mass drop and the asymmetry conditions are as follows

$$m_{j_1} < \mu m_j$$

$$y \equiv \frac{\min(p_{Tj_1}^2, p_{Tj_2}^2)}{m_j^2} \Delta R_{j_1,j_2}^2 > y_{cut}$$

where $\mu$ and $y_{cut}$ are free parameters.

3. Otherwise redefine subjet $j_1$ as $j$ and repeat.
Figure 2.8: Pruning method. The initial jet is reclustered using a C/A algorithm and each time the subjets have to be merged, conditions about their $p_T$ and $\Delta R$ have to be checked. If the conditions are met, then the two subjets are not merged and the subjet with the lower $p_T$ is discarded (taken from [16]).

The mass drop algorithm is illustrated in the top part of figure 2.7. When a mass drop is found, i.e., when the condition 2 is fulfilled, particles contained in $j_1$ and $j_2$ are reclustered using $R_{\text{filt}} = \min(0.3, R_{j_1,j_2}/2)$. The three hardest subjets are then combined as the new “filtered” jet and all constituents outside these three jets are discarded. This choice of keeping the three hardest jets allows one additional radiation from a two-body decay to be captured. By using the C/A algorithm to isolate $j_1$ and $j_2$, the angular scale of any potential massive particle decay is known. Reclustering the jet dynamically at an appropriate angular scale that allows to resolve that structure, maximizes the sensitivity to highly collimated decays, as can be seen in the bottom part of figure 2.7.

It’s possible to do the filtering (reclustering) without doing the mass drop requirement. However, in this analysis mass drop is always included, as will be mentioned in section 3.5 along with the parameters used.

2.3.2 Pruning

Given a jet, one reclusters it with a C/A algorithm and each time that subjets $i, j \rightarrow p$ have to be merged, first the following condition has to be checked:

$$z \equiv \frac{\min(p_{T,i}, p_{T,j})}{p_{T_p}} < z_{\text{cut}} \quad \text{and} \quad \Delta R_{i,j} > D_{\text{cut}} \quad (2.11)$$

where $z_{\text{cut}}$ and $D_{\text{cut}}$ are free parameters. If condition 2.11 is met, then the two subjets don’t have to be merged and the jet with the lower $p_T$ is discarded. This continues until all particles are either clustered or discarded. This technique is illustrated in figure 2.8.
2.3.3 Trimming

Given a jet, one reclusters it with a $k_t$ algorithm with radius $R_{sub}$ to identify the subjets. Then, the following condition is checked:

$$p_{T_i} < f_{cut} P_{T_{jet}}$$  \hspace{1cm} (2.12)

where $P_{T_{jet}}$ is the $p_T$ of the original jet. When condition 2.12 is met, the corresponding subjet $i$ is discarded. The difference between trimming and filtering is that in the filtering the number of subjets is fixed while in trimming whether a subjet is kept or not is determined by it’s $p_T$. A representation of this technique is shown in figure 2.9.
In ATLAS the event reconstruction is done by packages implemented in the software framework Athena [17], which process the events, starting from raw data obtained from the different sub-detectors, through different stages to finally interpreting them as a set of objects such as jets, muons, electrons, and many more. In sections 3.2 and 3.3 we give an overview of the reconstruction of these objects. In section 3.1 we describe the trigger requirements for the ‘on-line’ event selection and in section 3.5 we describe the complete object and event selection used in the analysis to discriminate WW processes decaying semileptonically.

3.1 Trigger selection

Before undergoing an ‘off-line’ event selection, where any processing/cut/identification is done to data saved on disk, there’s an ‘on-line’ selection that corresponds to data that is being acquired, this job is done by triggers.

In this analysis, the events used are triggered using an electron candidate of transverse energy $E_T > 24$ GeV or $p_T > 60$ GeV or a muon candidate with transverse momentum $p_T > 24$ GeV or $p_T > 36$ GeV. The final levels of the trigger selection (EF for Event Filter) for a single electron used are $EF_{e24vhi\_medium1}$ and $EF_{e60\_medium1}$. Electrons are required to pass the medium1 selection and to be isolated. Medium1 selection means the tightness of the identification. There are three reference sets of cuts defined with increasing background rejection power, while keeping high efficiency: loose, medium and tight [4]. The loose selection uses simple shower-shape cuts and very loose matching cuts between reconstructed track and calorimeter cluster. The medium selection adds shower-shape cuts using information provided by the first layer of the EMC and track-quality cuts. Finally, the
tight selection incorporates cluster energy over track momentum information \((E/p)\), particle identification using the TRT, and discrimination against photon conversions. Electrons are considered isolated if the total transverse momentum around the electron within a cone of \(R = 0.2\) is less than 10% of the transverse energy of the cluster \(p_{T\text{cone}} < 0.1E_T^{\text{electron}}\) [18].

The single muon triggers used are \(\text{EF} \_\text{mu24i\_tight}\) and \(\text{EF} \_\text{mu36\_tight}\), which require at least one isolated muon, where the isolation criterion is made with inner detector tracks. Similar to the approach of the electrons regarding loose, medium and tight, these selections essentially refine the \(p_T\) thresholds, enhance the number of hits requirements and impose conditions on their layer location, and tune the transverse and longitudinal impact parameter with respect to the primary vertex to reject possible overlapping cosmic rays, among others.

### 3.2 Jet reconstruction

**Jet calibration**

Jets are reconstructed using the anti-\(K_t\) algorithm with \(R = 0.4\) using as input topological clusters (topo-clusters) built from energy deposited in the calorimeter cells. The topological clusters are calibrated using the Local Cluster Weighting (LCW) method, which partially corrects the response and reduces fluctuations due to the non-compensating nature of the ATLAS calorimeters. Next is a short explanation of the topological cluster algorithm that forms topo-clusters and how local hadron calibration uses this topo-clusters as input to jet calibration (more information can be found in [19]).

The calorimeter cells don’t reproduce the actual shape of the particle cascade and are susceptible to stochastic fluctuations of the detector state. There are basically two sources of noise in the calorimeter: one source of noise comes from the readout electronics, and the second is due to pile up, either in time or out-of-time. The main idea of topological clustering is to group neighboring cells that have significant energies compared to the expected noise into clusters. The topo-cluster algorithm results in a reduction of the calorimeter noise. Very noisy cells are excluded entirely from the topo-cluster formation, resulting in a minor loss of cells (\(\approx 0.1\%\) of all calorimeter cells). Topo-clusters that have a negative energy are rejected from the jet reconstruction.

The local cluster calibration (LCW) improves the jet energy resolution by weighting in a different way energy deposits that come from the electromagnetic and hadronic showers [20][21]. The LCW method classifies topo-clusters as ‘electromagnetic-like’ or ‘hadronic-like’ based on cluster shape variables. These variables are the energy density in cells in topoclusters, the cluster energy fraction deposited in different calorimeter layers, the isolation (that characterizes the energy around the cluster) and the depth of the cluster
barycentre in the calorimeter. These variables are used to calculate a classification weight that denotes the probability for a cluster to stem from a hadronic interaction. Jets are then reconstructed using hadronically calibrated topo-clusters. This calibration scheme reduces some of the sources of fluctuations in the jet energy response, thus improving the jet energy resolution \[22\] \[23\].

**Large-\(R\) jets**

Large-\(R\) jets, also called ‘fat jets’, refers to jets with \(R \geq 1.0\). In this work, we used Cambridge-Aachen (C/A) jets with \(R = 1.2\). These jets use as input locally calibrated topo-clusters. The jet collection used is called CamKt12LCTopo (ungroomed).

### 3.3 Lepton Reconstruction

#### 3.3.1 Electrons

Electron candidates are basically clusters of energy deposited in the electromagnetic calorimeter that are associated to a track reconstructed in the inner detector. Selected events must have at least one electron with the following parameters:

- \(p_T > 25\) GeV;
- \(|\eta| < 2.4\);
- impact parameter along the beam direction: \(z_0 < 0.5\) mm;
- transverse impact parameter significance: \(|d_0/\sigma(d_0)| < 3\);
- track isolation: \(\frac{p_{T_{cone30}}}{p_T} = \frac{\sum_{\Delta R < 0.3} p_{T_{track}}}{p_T} < 0.1\), where \(p_{T_{cone20}}\) is the sum of the transverse momentum of the tracks in a cone of size \(\Delta R = 0.2\) around the electron direction, excluding the electron.

#### 3.3.2 Muons

Muons selected for the present analysis were reconstructed using the combined muon reconstruction algorithm STACO. This algorithm reconstructs independently a track in the inner detector and a track in the muon spectrometer backtracked down to the \(z\)-axis and then it combines the track parameters obtained by a statistical averaging of the inner detector and muon spectrometer track parameters. Muons are also required to be originated in the primary vertex. To make sure that the muon is originating from a \(W\), selected events with at least one muon have to comply with:
• \( p_T > 25 \text{ GeV}; \)

• \(|\eta| < 2.4;\)

• impact parameter along the beam direction: \( z_0 < 1 \text{ mm}; \)

• transverse impact parameter significance: \(|d_0/\sigma(d_0)| < 3;\)

• \( \frac{p_{T\text{cone}20}}{p_T} = \frac{\sum_{\Delta R < 0.2} p_T^{\text{track}}}{p_T} < 0.15, \) where \( p_{T\text{cone}20} \) is the sum of the transverse momentum of the tracks in a cone of size \( \Delta R = 0.2 \) around the muon direction, excluding the muon.

3.4 Missing transverse momentum

The observables described so far, electrons, muons and jets, are defined to measure the signatures of electrically charged and/or strongly interacting particles detected by the different ATLAS sub-systems. The exceptions to this are electrically neutral color singlets like neutrinos. Missing transverse momentum, denoted \( E_T^{\text{miss}} \), accounts for the existence of these ‘invisible’ particles. It’s defined by the momentum imbalance in the plane transverse to the beam axis, where momentum conservation is expected. The vector momentum imbalance is obtained from the negative vector sum of the momenta of all particles detected in a \( pp \) collision:

\[
E_T^{\text{miss}} = -\sum_i \vec{p}_T^i
\]

where \( i \) corresponds to calorimeter cells, tracks, muons, etc. The estimation of the \( E_T^{\text{miss}} \) is done from reconstructed electrons and muons with \( \eta < 2.7 \), jets with \( \eta < 4.9 \) and clusters of energy in the calorimeters with \( \eta < 4.5 \) that were not included in the reconstructed objects (the expected energy deposit of identified muons in the calorimeter is subtracted). The total transverse missing momentum is named \( \text{METRefFinal} \).

3.5 Event selection

The event selection was performed in two steps, a first step called ‘pre-selection’ where we require the events to pass certain quality conditions, with also some broad filters for jets and leptons, and then a second selection, with more strict and detailed filters. Next we explain both steps and we show how many events remain after the event selection.
3.5.1 Event pre-selection

A preselection was done to decide whether the event and the objects were suitable for further analysis. This was done using the TopROOTCore software [18], which is composed of several analysis packages that have different functionalities like slimming (removal of parts of an object, the full object is not kept after this) or skimming (removal of unwanted events). This software is mostly used in top analysis, and the version used in this thesis (recommended for 2012 data) is AnalysisTop-1.9.0. Next is detailed the event preselection. This event selection is done once specifying the electron channel and one selecting the muon channel.

- **GRL** The data used for the analysis has to satisfy the Good Run List which is used to identify the quality of the run. A run is a data taking period in ATLAS, which is subdivided into luminosity blocks (typically 1 or 2 minutes long) where the instantaneous luminosity is approximately constant. The status of each sub-detector is checked during the run and data-quality flags are used to identify luminosity blocks suitable for each physics analysis. The good run lists contain these data-quality flags.

- **Trigger** The event has to satisfy the trigger requirements (EF_e24vhi_medium1 or EF_e60_medium1 in the electron case and EF_mu24i_tight or EF_mu36_tight in the muon case).

- **NVP** When calculating the number of primary vertices, at least 3 tracks per vertex are required.

- **≥ 1 lepton** The event has to have at least one lepton of the specified channel (if we want to pre-select events in the electron channel, then it has to find at least one electron that has some quality conditions for reconstruction, as mentioned in 3.3).

- **1 lepton** There mustn’t be any second of this lepton that passes the same quality requirements.

- **0 other lepton** Depending on the channel selected, there mustn’t be any lepton of other type passing those quality requirements. For example, if we selected the electron channel, then if the event also has a muon which fulfills the quality requirements, then the event is discarded.

- **Trigger match** The lepton in the event has to be the one that fired the trigger.

- **Jet cleaning** This cut is to identify so-called ‘bad’ jets. These jets are not originated from the event, instead they are caused by various sources such as hardware problems in the calorimeter, LHC beam-gas interactions and cosmic-ray showers. The
recommended `isBadLooseMinus` criteria is used. Also, the jet is pile up corrected using a cut on the Jet Vertex Fraction $|JVF| < 0.5$. This is applied to all ‘small’ (R = 0.4) jets with $p_T < 50\text{GeV}$ and $|\eta| < 2.4$. Also, there’s an overlap removal between the small jets and the leptons.

- **≥ 1 jet** The event has to have at least one jet with $p_T > 25 \text{GeV}$.
- **≥ 2 jets** The event has to have at least two jets with $p_T > 25 \text{GeV}$.
- **1 C/A jet** The event has to have at least one C/A jet with $p_T > 40 \text{GeV}$ and $|\eta| < 3.5$.

### 3.5.2 Full event selection

Events that pass the event preselection are slimmed, *i.e.*, only retain the necessary variables for the full event selection and all the extra variables that were in the original Common Ntuple are discarded (for example, several containers with different jets reconstructions not necessary for the analysis are discarded). The advantage of the slimming is that the resulting Ntuple is much smaller and therefore is significantly less time consuming for the analysis.

WW events decaying in the semileptonic channel ($WW \rightarrow \ell \nu jj$) are identified in principle requiring two jets, one ‘fat jet’, one high-pt lepton (either electron or muon) and missing transverse energy. Additional cuts are included to improve pileup rejection and background rejection. During the present work, some variations of the event selection were implemented, and the final configuration used for most of the analysis is the following:

- **Trigger:** Events are required to pass the `EF_e24vhi_medium1` trigger in the electron case or the `EF_mu24i_tight` in the muon case.

- **SL Event and Channel:** Only events that decay semileptonically are accepted (SemiLeptonic Event). Channel checks that the lepton that passed the single lepton requirement comes from the decay of the $W$. If that’s not the case, it rejects the event.

- **Lepton $p_T/|\eta|$** Leptons are required to have $p_T > 25 \text{GeV}$ and $|\eta| < 2.4$.

- **Masked modules and 2 HadJets:** These conditions mean that the event is required to have at least two jets and the jets can’t be masked. For a variety of reasons, there are modules which are either temporarily or permanently masked throughout all data taking periods. Studies have shown that medium to high $p_T$ jets which fall within a masked tile calorimeter module (this region is called ‘core’) are usually undercorrected, while jets in modules adjacent to a masked tile calorimeter module (this region is called ‘edge’) are overcorrected. It becomes important to kill events where such
jets fall into masked regions, as otherwise the jet is poorly reconstructed. Therefore, events that are masked are rejected. For this purpose, we used the BCHCleaningTool that has a function called TileTripReader that handles dead regions in data.

To kill these events, it is necessary to have a handle on when modules are masked, for a given run and lumiblock. For MC, it’s more complicated, as one has to ensure that the same thing is done in both data and MC. The BCHCleaningTool methods for MC need to know about the right time-dependence of the masking to equalize the data and MC and emulate the inefficiency in data in terms of masked modules. This is achieved by passing random run and lumiblock numbers generated from those present in the Good Run List in the MC and then the TileTripReader is used in the same way as in data.

- **> 0 FatJets and FatJetEta:** Events are required to have at least one C/A jet with \( p_T > 20 \text{ GeV} \) and \( |\eta| < 2.8 \). The C/A jets used were reclustered using the constituent clusters of the original CamKt12LCTopo and doing an overlap removal between the clusters and the electron, in order to reject a fake jet from an electron. Essentially, if \( \Delta R_{cl,el} < 0.2 \) (where \( \Delta R_{cl,el} \) means the distance in the rapidity-azimuth plane between the cluster and electron), the cluster was rejected and, therefore, it wasn’t used to generate the new C/A jet with \( R = 1.2 \).

- **dPhiJ1MET and dPhiLeMET:** Events are required that the azimuthal separation between the leading jet and the missing transverse energy must be \( \Delta \phi(E_T^{\text{miss}}, j_1) > 1.5 \). Also, the azimuthal distance between the lepton and the missing transverse energy has to be \( \Delta \phi(E_T^{\text{miss}}, \ell) < 1.7 \). These cuts allowed for QCD multijet background rejection.

- **TwoGoodJets:** We required at least two antiKt4 jets, with the leading jet above 30 GeV and the second leading jet above 25 GeV. Any event that has a third jet with \( p_T > 25 \text{ GeV} \) is vetoed.

- **FatJetPt and FatJetMass:** Events need to have a C/A jet with \( p_T > 150 \text{ GeV} \) and \( m > 60 \text{ GeV} \).

- **dRJ12<1.6:** The angular distance between the two leading jets has to be \( \Delta R_{j1,j2} < 1.6 \).

- **mT and EtMiss:** Events are required to have \( E_T^{\text{miss}} > 30 \text{ GeV} \) in order to take into account the neutrino from the \( W \to \ell \nu \) decay, and the transverse mass \( m_T \)\(^1\)

\[^1\]The transverse mass, \( m_T(W) \), is calculated from the lepton transverse momentum \( p_T^{\ell \nu} \) and the difference of the azimuthal angle \( \Delta \phi \) between the \( E_T^{\text{miss}} \) and \( p_T^{\ell \nu} \) vector as \( m_T(W) = \sqrt{2 E_T^{\text{miss}} p_T^{\ell \nu} (1 - \cos (\Delta \phi(E_T^{\text{miss}}, p_T^{\ell \nu})))} \).
is required to be greater than 40 GeV. These two cuts greatly suppress the QCD multijet background.

In the analysis, the muon and electron channels are analyzed separately but the event selection for both is the same. Table 3.1 summarizes the event selection and figure 3.1 shows the number of events that survive each of the requirements for a standard Monte Carlo sample and data.

The very small amount of events that survived after the full event selection in the official Monte Carlo sample made clear the need to produce more Monte Carlo samples. The low efficiency of the event selection (approximately 0.5%) motivated the inclusion of filters at the generator level, producing already boosted events. When later analyzed using the full event selection, these already preselected boosted events would have a better efficiency, resulting in having to produce fewer events (a lot less computational time required to generate the Monte Carlo samples) and also reducing the offline analysis time.
Table 3.1: The cut-flow for the series of vetoes described in subsection 3.5.2 applied to the official Monte Carlo $WW$ sample. A pre-selection was applied first, as described in subsection 3.5.1 where the initial number of events was 2.5 million. The values here are only for the electron channel, but the muon channel yields similar values. Approximately 0.4% of the events remain after the full event selection considering both muon and electron channels.

<table>
<thead>
<tr>
<th>Cut</th>
<th>Name</th>
<th>Description</th>
<th># events</th>
</tr>
</thead>
<tbody>
<tr>
<td>C0</td>
<td>Initial</td>
<td></td>
<td>46947</td>
</tr>
<tr>
<td>C1</td>
<td>Trigger</td>
<td>$\text{EF}_{\text{e24vhi_medium1 or EF_mu24i_tight}}$</td>
<td>45551</td>
</tr>
<tr>
<td>C2</td>
<td>SL event</td>
<td>Semileptonic event</td>
<td>37044</td>
</tr>
<tr>
<td>C3</td>
<td>Channel</td>
<td></td>
<td>37044</td>
</tr>
<tr>
<td>C4</td>
<td>lepton $p_T/\eta$</td>
<td>$p_T &gt; 25\text{ GeV, }</td>
<td>\eta</td>
</tr>
<tr>
<td>C5</td>
<td>Masked</td>
<td>Masked Modules</td>
<td>35272</td>
</tr>
<tr>
<td>C6</td>
<td>2 HadJets</td>
<td>At least two jets</td>
<td>35272</td>
</tr>
<tr>
<td>C7</td>
<td>&gt;0 FatJets</td>
<td>At least one C/A jet with $p_T &gt; 20\text{ GeV}$</td>
<td>35272</td>
</tr>
<tr>
<td>C8</td>
<td>FatJet Eta</td>
<td>C/A jet with $</td>
<td>\eta</td>
</tr>
<tr>
<td>C9</td>
<td>dPhiJ1MET</td>
<td>$\Delta\phi(E_T^{\text{miss}, j_1}) &gt; 1.5$</td>
<td>28354</td>
</tr>
<tr>
<td>C10</td>
<td>dPhiLeMET</td>
<td>$\Delta\phi(E_T^{\text{miss}, \ell}) &lt; 1.7$</td>
<td>24451</td>
</tr>
<tr>
<td>C11</td>
<td>twogoodjets</td>
<td>$p_{T_{j_1}} &gt; 30\text{ GeV, } p_{T_{j_2}} &gt; 25\text{ GeV, } p_{T_{j_3}} &lt; 25\text{ GeV}$</td>
<td>14357</td>
</tr>
<tr>
<td>C12</td>
<td>FatJetPt</td>
<td>C/A jet with $p_T &gt; 150\text{ GeV}$</td>
<td>12109</td>
</tr>
<tr>
<td>C13</td>
<td>FatJetMass</td>
<td>C/A jet with $m &gt; 60\text{ GeV}$</td>
<td>11297</td>
</tr>
<tr>
<td>C14</td>
<td>dRJ12&lt;1.6</td>
<td>$\Delta R(j_1, j_2) &lt; 1.6$</td>
<td>7931</td>
</tr>
<tr>
<td>C15</td>
<td>$m_T$</td>
<td>$m_T &gt; 40\text{ GeV}$</td>
<td>4931</td>
</tr>
<tr>
<td>C16</td>
<td>EtMiss</td>
<td>$E_T^{\text{miss}} &gt; 30\text{ GeV}$</td>
<td>4664</td>
</tr>
<tr>
<td>C17</td>
<td>Tag</td>
<td>FatJet mass drop tag</td>
<td>4656</td>
</tr>
</tbody>
</table>
MC simulation and data

In this chapter we focus on the generation of Monte Carlo samples of $WW$ processes, introducing an event selection at generator level in order to improve the efficiency in the off-line analysis. As it was showed at the end of chapter 3, the existing Monte Carlo samples didn’t have enough events after doing the final event selection so it became necessary to produce more samples. In order to have a better efficiency in the off-line analysis, instead of producing the Monte Carlo samples as they were produced initially, we added a primitive event selection to the generator. This work was done with two different MC generators, Pythia and Herwig. In section 4.1 we describe the stages involved in a full simulation and give a short overview of the generators used. In section 4.3 we describe the event selection used in both generators and discuss the main results. Finally, in section 4.4 we show the validation of the produced MC samples compared to the official sample after the complete event selection is applied.

4.1 Stages of the full simulation

Data simulation in Atlas provides access to information of the event in many different levels, some of these inaccessible in the case of real data. For example, in a simulation we can know exactly what was the interaction that led to the event and how the particles interacted afterwards. With the goal of modeling real events, event generators are used, together with sophisticated detector simulation. This simulation is finally processed with the same reconstruction algorithm used in real data to obtain a set of events that is equivalent to the experimental results. As it was mentioned before, in ATLAS simulation and reconstruction are done in the Athena framework.

The process to generate simulated events in ATLAS is as follows:
• **Generation:** an event generator that simulates the interaction between two protons is used. Generators like **Pythia** or **Herwig**, explained below, simulate the interaction between the quarks and the subsequent parton shower. All possible channels of interaction between protons are simulated, weighted by their cross section.

• **Simulation:** in order to model events that might be observed with the detector, the events produced in the generator phase are used as input to a detector simulation model. **ATLAS** uses Geant4 toolkit for this[24] [25]. Geant4 is an extensive particle simulation toolkit that governs all aspects of the propagation of particles through detectors, based on a description of the geometry of the detector components and the magnetic field. The output of this algorithm indicates the energy deposited by the generated particles when they interact with the detector.

Due to the detailed and complex geometry of the ATLAS detector, and the many physics processes involved, the consumed computing time per event is large ($O(1 \text{ hour})$). This is why the simulation of events are normally done in parallel, taking advantage of grid computing.

• **Digitisation:** This step is a second level of detector simulation where the response of the detector is simulated (obtaining voltages and time values). The output from this stage is similar to that which might be expected from the readout electronics in the actual experiment. The digitisation stage is very fast, except when pile-up at high luminosity is added.

• **Reconstruction:** The reconstruction is done in two stages, first, the signal is reconstructed in a stand-alone mode. Second, the information from all detectors is combined to get a more accurate measurement and identification of the final objects used in the analysis (jets, leptons, photons, missing transverse energy, etc). In this stage signals from the digitisation are introduced in the same reconstruction algorithm as signals originating from a real event.

As it was mentioned in the end of the previous chapter, due to the strict cuts needed to analyze $WW$ processes in the boosted regime, where only 0.05% of the events survive the selection, it became mandatory to increase the number of produced Monte Carlo samples. We did several studies introducing a preliminary event selection at generator level using **Pythia** and **Herwig** generators. Next is a brief overview of the physics process and how simulation is done in these generators. The main difference between these two generators relies on how each produces the showering process and their treatment of the hadronization, but both generators are usually used to generate the same process and calculate systematic uncertainties.
PYTHIA

The original PYTHIA generator started in 1978 as part of the development of another event generator, JETSET. Originally it was written in Fortran 77, a language that continued being used until PYTHIA 6. Since Fortran support started to decline at this point, in PYTHIA 7 part of the code was migrated to C++, with mixed results. PYTHIA 8 became the successor to PYTHIA 6 and it was rewritten entirely in C++ [26].

The possible hard scatter processes allowed in PYTHIA are $pp$, $\bar{p}p$, $e^+e^-$ and $\mu^+\mu^-$, there is no option for $ep$ collisions or for incoming photon beams. It has over 200 possible subprocesses, which are mainly $2 \rightarrow 1$ and $2 \rightarrow 2$. In PYTHIA showers are ordered in transverse momentum both for Initial State Radiation and for Final State Radiation.

The program structure can be summarized in three stages, as follows [27]:

1. The generation of a certain process, usually a “hard process”, such as $gg \rightarrow h^0 \rightarrow Z^0Z^0 \rightarrow \mu^+\mu^-q\bar{q}$, that is calculated in leading order perturbation theory. Only very few partons/particles are defined at this level, so only the main aspects of the event structure are covered.

2. In the second stage, all subsequent activity on the parton level is generated, involving initial and final-state radiation, multiple parton-parton interactions and the structure of beam remnants. Much of the phenomena are under an (approximate) perturbative control, but nonperturbative physics aspects are also present. At the end of this stage there’s a realistic partonic structure (e.g. with broadened jets and an underlying-event activity).

3. In the last stage the hadronization of the parton configuration occurs, by string fragmentation, followed by the decays of unstable particles. This part is almost completely nonperturbative, and so requires extensive modeling and tuning or parametrization of existing data. Only at the end of this step realistic events are available, as they could be observed by the detector.

For the present work, simulated samples of WW events from $pp$ collisions were generated using PYTHIA version 8.186 released in 2014.

HERWIG

HERWIG (Hadron Emission Reactions With Interfering Gluons) is a general-purpose Monte Carlo event generator, which includes the simulation of hard lepton-lepton, lepton-hadron and hadron-hadron scattering and soft hadron-hadron collisions. HERWIG was first published in 1986 and was written in Fortran. As in the case of PYTHIA, at the beginning of the LHC era, HERWIG was rewritten entirely in C++ and has been renamed HERWIG++ [28].
The event simulation in Herwig can be divided in 4 stages, that correspond roughly to increasing scales of distance and time [29]:

1. **Elementary hard subprocess.** A pair of incoming beam particles or their constituents interact, producing one or more primary outgoing objects. In general this is computed at leading order in perturbation theory.

2. **Initial- and final-state parton showers.** The colored particles in the event evolve perturbatively. This occurs for both the particles produced in the collision, the final-state shower, and the initial partons involved in the collision for processes with incoming hadrons, the initial-state shower.

3. **Heavy object decays.** Massive fundamental particles such as the top quark, electroweak gauge bosons, Higgs bosons and more, decay on time-scales that may be shorter than or comparable to that of the QCD parton showers. Depending on their nature and the decay mode, they may also initiate parton showers before and/or after decaying.

4. **Hadronization process.** In order to construct a realistic simulation one needs to combine the partons into hadrons. This applies to the constituent partons of incoming hadronic beams as well as to outgoing products of parton showering, which give rise to hadronic jets. Since perturbation theory is not applicable because the hadronization process takes place at a low momentum transfer scale and the strong coupling is large, a phenomenological model is used.

The version of Herwig used for the event generation on this thesis was 6.520.2 released in 2010.

### 4.2 Monte Carlo samples and Data-sets

#### 4.2.1 Standard model MC

The $WW/WZ$ events are generated using Herwig 6.520.2 with AUET2 tune and PDF set CTEQ6L1. The background processes used in this work are $W/Z$+jets, $W/Z$ plus heavy flavor jets, single top and $t\bar{t}$. The vector boson processes $W/Z$+jets and $W/Z$ plus heavy flavor jets are produced using Sherpa 1.4.1 with PDF set CT10. Samples of single top and $t\bar{t}$ events are produced using Powheg+Pythia, with Powheg PDF set CT10 and Pythia 6.426 PDF set CTEQ6L1.

All MC samples are simulated with pileup and reweighted using TPileupReweighting to make the average number of interactions per bunch-crossing consistent with the value
in data. Also, additional weights are applied depending on filter efficiencies, MC weights, k-factors and cross-sections for the different samples.

4.2.2 Recorded Data

The data sample used in this analysis corresponds to the \( pp \) collision data recorded by the ATLAS detector during 2012 with a center-of-mass energy of \( \sqrt{s} = 8 \) TeV. The integrated luminosity for this period was \( 20.3 \text{ fb}^{-1} \) and the mean number of \( pp \) interactions per bunch crossing (\( \mu \)) ranged between 5 and 40, resulting in big variations of pile up conditions during the data taking period.

4.3 Event selection at generator level

In order to increase the number of Monte Carlo samples, a straightforward approach would be to replicate the official Monte Carlo process, with the addition of a one lepton filter. A slightly more complicated approach, but more beneficial in terms of efficiency, is to include more filters at generator level. This means that we would need to produce a lot less events in order to have a significant amount of events after the offline selection. Also, in case of producing fully simulated samples, having to generate a huge amount of events would be extremely time consuming, especially in the simulation phase. If one can already have a preselection at generator level, this means that less events would have to be simulated. Including filters in the generator of course makes it more time consuming, because the generator itself has to create a lot more events and filter only the one that pass the requirements. But the extra time required to generate the events is orders of magnitude shorter than the time it would take to simulate a greater amount of events without the initial selection.

4.3.1 PYTHIA

The first samples generated were done using the PYTHIA generator. The goal was to be able to study \( WW \) phase space with a large statistical sample that was easy to produce. The events were generated in two parts, first we produced events with the \( W^+ \) decaying leptonically and the \( W^- \) decaying hadronically, and then this was interchanged. The focus of the studies presented in this part using PYTHIA is to have a good simulation of the event at generator level. Generator level means to have only the generated events produced by PYTHIA without doing the further steps in the full simulation (i.e. without going through simulation, digitisation and reconstruction as explained in section 4.1).

As can be seen in figure 3.1, many events are lost in the Fat Jet high \( p_T \) requirement as well as in the two ‘small’ jets (\( \text{AntiKt4} \)) requirements. Because of this, we required at
least two small jets above 25 GeV with $\eta < 2.8$, one C/A jet above 120 GeV and $\eta < 2.8$, and one lepton with $p_T > 25$ GeV and $\eta < 2.4$. Any event that had a third small jet with transverse momentum above 25 GeV and had a second fat jet was vetoed. Because of the filters, the efficiency of the generator was around 10%. Below we present the summary of the generator requirements:

Signal: $WW \rightarrow l\nu qq$

Filters:
- Lepton Filter: $p_T \geq 25$ GeV and $|\eta| \leq 2.4$.
- Small Jet Filter: Min 2 jets, $p_T \geq 25$ GeV, $|\eta| \leq 2.8$.
- Fat Jet Filter: $p_T \geq 120$ GeV, $|\eta| \leq 2.8$.

Vetoes:
- Third small jet veto: we veto events that have a third jet above 25 GeV.
- One Fat Jet only: we veto events that have a second FatJet.

Center of mass energy of 8 TeV.

Figure 4.1 shows the $p_T$ and spatial distributions of the $W^+$ right before decaying but after undergoing the successive energy corrections after the hard subprocess. The right bottom panel is the $p_T$ histogram with a mean of 178 GeV. Obtaining a $W$ with such high transverse momentum is evidence that the high $p_T$ jet filters worked. Figure 4.2 shows the mass of the leading two ‘small’ jets in the top, and in the bottom the sum of the first three leading small jets (left) and the ‘fat jet’ (right), these last two have masses around 90 GeV which is close to the mass of the $W$.

Figure A.1, shows the comparison between the signal events obtained after the event generation applying the filters as described above and the official Monte Carlo, the results observe don’t show a good agreement. In order to have a good description of $WW$ events and to be able to compare this simulation with the official MC samples, it’s necessary to add pile up to the signal generated. For this we used PYTHIA to produce soft QCD events. This corresponds to the inelastic nondiffractive part of the total cross section, what is usually called the “minimum-bias” component. No hard process at all is defined at the process-level part of the event generation. The resulting histograms of the analysis done with pile are shown in figures A.2 and A.3. Below is the summary of the pile up generation.
Figure 4.1: $\eta$ (top left), $\phi$ (top right), $p_T$ (bottom left) and mass (bottom right) distributions for the truth $W$ boson generated with PYTHIA, with status 62, i.e., after the energy corrections are applied before it decays.

Figure 4.2: Mass of the two leading 'small' jets (top) and sum of the masses of the two leading small jets (bottom left) and mass of the fat jet (bottom right).
Pile up: Soft QCD (minBias) 
No filters or vetoes.
Center of mass energy of 8 TeV.

The second goal of these generator level studies was to also produce background samples. As it was mentioned in subsection 4.2.1, the dominant background of $WW$ signal is $W$+jets production. The samples produced in this part have been generated without detector simulation, as in the case of $WW$. Additionally, as in the case of $WW$ production, the same filters were added to the generator. Next is the summary of the event generation for $W$+jets.

$W + $ Jets: $Wg \rightarrow l\nu g$

Filters:
- Lepton Filter: $p_T \geq 25$ GeV and $|\eta| \leq 2.4$.
- Small Jet Filter: Min 2 jets, $p_T \geq 25$ GeV, $|\eta| \leq 2.8$.
- Fat Jet Filter: $p_T \geq 120$ GeV, $|\eta| \leq 2.8$.

Vetoes:
- Third small jet veto: we veto events that have a third jet above 25 GeV.
- One Fat Jet only: we veto events that have a second FatJet.

Center of mass energy of 8 TeV.

In order to have a better description of the background, pile up was added to $W$+jets events. As mentioned in section 1.1.1, the average interaction per bunch crossing during data taking in 2008 was 20.7. The first attempt to add pile up was following the distribution shown in 1.2. This proved to be excessive and resulted in jets that were too massive when compared to official MC samples.

This is because the pile up that we added is only in-time pile up. In reality, the structure of pile up is more complicated and in-time and out-of-time pile up are cancelling partially. Figure A.5 corresponds to the pulse shape in LAr. It shows that inside a bunch train the average response in LAr is 0. Thus, in-time and out-of-time pile up cancel on average. But if we have only moderate pileup the occupancy is such that the cancellation does not work locally - only on a global or time averaged scale. Individual events and jets have more or less contribution from in-time and out-of-time PileUp and thus do not come with “0” but some residual energy. In an attempt to mimic this residual amount of the 20 pile up events
every bunch crossing, we tried adding pile up in lower numbers, for example 3, 5 and 7 soft QCD events per $W+jet$ signal event. In the end, the number that better adjusted to official samples was a Poisson distribution with $\mu = 5$. In figure 4.3 we show the mass and $p_T$ of the C/A jet and the two leading small jets compared to the official sample. As it can be seen from the figure, the description of the fat jet is very good but not so much the description of the two small jets.

These discrepancies motivated to follow a different approach in the Monte Carlo production. Also, it was decisive in this matter that the substructure variables calculated from the C/A jet greatly differed from the original sample, especially variables like $\tau_2/\tau_1$, which is an important variable because of it’s great discriminating power between signal and background. Figure A.4 shows the comparison of some of the substructure variables between the private sample and the original one.

The first alternative considered to produce new Monte Carlo samples was to use the samples generated with PYTHIA as input to the simulation and digitisation, in order to have a fully simulated sample an fix the problems arising from adding pile up artificially. Due to incompatibilities between the releases of the full simulation and the necessary input to be able to use TopRootCore [30] to do the event preselection, we decided to switch from the PYTHIA generator to HERWIG. The main incompatibility issue was that the generation with PYTHIA was done using Athena release 20, whose output after digitisation are xAODs, which have become the new standard for ATLAS, especially for Run II, while the generation of the existing official MC samples for Run 1 were done in Athena with release 17, where the output of the digitisation are AODs (now deprecated for Run II), easily transformed in Common NTUP, which is the standard input to the TopRootCore software. The other reason to switch to the Herwig generator was that the official Monte Carlo samples were produced using this generator, in release 17 of Athena.
Figure 4.3: $p_T$ (left) and mass (right) distributions comparing $W$+jets produced with Pythia with pile up added at generator level with official Monte Carlo samples of $W$+jets fully simulated. The fat jets show a very good agreement but the small jets present big discrepancies, especially in the masses, where there’s a shift to higher values.
4.3.2 HERWIG

The first step with the HERWIG generator was to produce a fully simulated sample using the same parameters as in the official Monte Carlo production. This was done in order to check that all the stages of the simulation were working correctly, especially the digitisation stage where the pile up is added.

The test samples were produced using only a lepton filter that required $p_T > 23\text{ GeV}$ and $|\eta| < 2.4$, as in the original production. These samples presented a very good agreement with the original Monte Carlo which can be observed in figures B.1 and B.2 (in appendix B), where the comparison of relevant kinematic magnitudes are shown.

The next step was to produce events adding filters in the generator. We required a lepton with $p_T > 23\text{ GeV}$ and $|\eta| < 2.4$, two AntiKt4 jets in the range $|\eta| < 2.8$ with the leading jet $p_T$ above 28 GeV and the second-leading jet $p_T$ above 23 GeV. We also required a fat jet ($C/A$ jet with $R = 1.2$) with $p_T > 130\text{ GeV}$ and $|\eta| < 2.9$. In addition, any event with a third small jet with $p_T > 27\text{ GeV}$ or any event with a second fat jet with $p_T > 40\text{ GeV}$ was vetoed. We also added a missing transverse energy filter, with $MET > 30\text{ GeV}$. In order to improve the production efficiency, we included a parameter called $p_T\, Min$ that corresponds to the $p_T$ of the $W$ when is produced (before subsequent energy corrections). Having a $p_T\, Min$ allows to produce more energetic events, therefore improving the efficiency of the generator, but it’s important to have a reasonable value, in order to have a full phase space of events and not bias the sample. Next is the summary of the generator configuration.

Signal: $WW \rightarrow l\nu qq$

$p_T\, Min = 70\text{ GeV}$ ($p_T$ of the $W$ when is produced, before energy corrections).

Filters:

- Lepton Filter: $p_T > 23\text{ GeV}$, $|\eta| < 2.4$
- MET Filter: $MET > 30\text{ GeV}$
- 2 jets: $p_{T1} > 28\text{ GeV}$, $p_{T2} > 23\text{ GeV}$, $|\eta| < 2.8$
- FatJet: $p_T > 130\text{ GeV}$, $|\eta| < 2.9$

Vetoes:

- 3rd jet veto: Veto on any third jet with $p_T > 27\text{ GeV}$
- 2nd FatJet veto: Veto on any second FatJet with $p_T > 40\text{ GeV}$

Center of mass energy of 8 TeV.

Generator Eff: 0.94%
Figure 4.4: Plots show the number of times the trigger fired in the muon channel (left) and electron channel (right). Plots on top correspond to the official Monte Carlo sample where one observes that both channels have similar values. Plots on bottom correspond to our MC samples where there’s a big difference between the muon and electron channel. See subsection 4.3.2 for details.

There were several issues with this configuration of the generator. The filters were too tight, and therefore we didn’t have a good event simulation compared with the official production. In addition, one of the main features was that when looking at the final events, there was a huge discrepancy between the amount of times that the muon trigger was fired versus the amount of times where the electron trigger fired. In the official sample, the number of times that the triggered electron was fired was roughly 50%, as in the muon case. This wasn’t the case in our private sample, as evidenced in the plots of figure 4.4. As it was mentioned in the previous chapter, electrons can simulate jets, given that they leave energy deposits in the calorimeters as in the case of the hadronic showers. In our case, very energetic electrons (because of the lepton filter) were faking jets, and because of the jet vetoes, these events were being discarded. Muons don’t have this problem because they can’t fake jets, so events were the $W$ decayed to a muon weren’t rejected because of the jets vetoes. The solution to this was to modify the way jets were reclustered in the generator. Jets at generator level (truth level) use as input only stable particles, but before they are stable, particles undergo some energy corrections. A $W$, for example, doesn’t decay to a stable electron, rather it decays to an electron that will loose some energy until it reaches stable status. Therefore, in order to exclude the correct electrons from the jet reconstruction algorithm (this means, electrons that are generated from the
W decay), one has to look for all stable electrons and go up the family tree to see if they come from a W. Softer electrons produced from quarks are not excluded. Also, electrons that were produced from a tau particle (always unstable) that was the result of a W decay were also rejected.

With the changes implemented in the jet reconstruction algorithm at generator level, and modifying the filters to make them less strict, we arrived at the final configuration of the generator. In this configuration we require a lepton with $p_T > 20\text{ GeV}$ and $|\eta| < 2.4$, two small jets in $|\eta| < 2.8$ with $p_T$ above 28 GeV and 23 GeV for the leading and second-leading respectively, and a fat jet with $p_T > 100\text{ GeV}$ and $|\eta| < 2.8$. Also, any event with a third jet with $p_T > 27\text{ GeV}$ or with a second fat jet with $p_T > 50\text{ GeV}$ is vetoed. In the next section we present the validation of the private Monte Carlo sample. The summary with the configuration of the generator is shown below.

**Signal:** $WW \rightarrow l\nu qq$

**Filters:**
- Lepton Filter: $p_T > 20\text{ GeV}$, $|\eta| < 2.4$
- 2 jets: $p_{T1} > 28\text{ GeV}$, $p_{T2} > 23\text{ GeV}$, $|\eta| < 2.8$
- FatJet: $p_T > 100\text{ GeV}$, $|\eta| < 2.9$

**Vetoes:**
- 3rd jet veto: Veto on any third jet with $p_T > 27\text{ GeV}$
- 2nd FatJet veto: Veto on any second FatJet with $p_T > 50\text{ GeV}$

Center of mass energy of 8 TeV.

Generator Eff: 1.0046%

### 4.4 Validation of private MC sample

The private Monte Carlo sample showed a good agreement with the official sample, but there were still some small discrepancies especially in the $\tau_2/\tau_1$ substructure variable, as shown in figure 4.5, where the most relevant quantities for the electron and muon channel are presented.

Since $\tau_2/\tau_1$ is one of the most useful variables to discriminate between signal and background, we decided to reweight the sample calculating the ratio between the $\tau_2/\tau_1$ histograms of the official sample and the private one. This is shown in the four top plots of figure 4.6, where the figures on the left correspond to the electron channel and the ones on
Figure 4.5: Plots on the left correspond to the electron channel and plots on the right correspond to the muon channel. The $p_T$ of the leading small jet (top) and the leading fat jet (middle) from our private sample show a good agreement with the official sample. Plots on bottom correspond to $\tau_2/\tau_1$ variable, which shows some discrepancies with the official sample.
the right to the muon channel. Then we fitted this ratio using polynomial functions, taking into account the statistical uncertainties in the fit. This is shown in the four bottom plots of figure 4.6. Already with a polynomial of order one, the $\chi^2$ over number of degrees is very small and doesn’t change much with regards to the second order polynomial. Nevertheless, looking at the complete set of kinematic and substructure variables, we decided to use the quadratic fit polynomial as a weight for the private MC sample.

The $p_T$ of the leading small jets and fat jets along with $\tau_2/\tau_1$ are shown in figure 4.7 (for the electron channel) and 4.8 (for the muon channel). On the left we show the original un-weighted histograms, and on the right the reweighted histograms using the quadratic fit. It can be seen that the agreement improves with the fit. Also, other magnitudes like number of jets in the event, second and third leading jets improve, but the comparisons are not presented here. Finally, in section B.2 we show the most relevant kinematic and substructure variables of our private production with the reweighting applied compared with the official MC sample. The comparison of the results using the un-weighted and weighted events could be treated as systematic error for following analysis.
Figure 4.6: Plots on top show the ratio between $\tau_2/\tau_1$ histograms of the official sample and the private one, including the statistical error. The second plot from top shows a zoom of this ratio, excluding the most extreme values. The four plots on the lower part correspond to a linear fit and a quadratic fit of the ratio between $\tau_2/\tau_1$. Plots on the left correspond to the electron channel and plots on the right correspond to the muon channel.
Figure 4.7: On the left are the distributions of the $p_T$ of the leading small jet, fat jet and $\tau_2/\tau_1$ of the original sample and private one. On the right are the same quantities weighted by the quadratic fit of the ratio between $\tau_2/\tau_1$ of the official sample and the private one. One can observe a better agreement in the plot of the right. All plots correspond to the electron channel.
Figure 4.8: On the left are the distributions of the $p_T$ of the leading small jet, fat jet and $\tau_2/\tau_1$ of the original sample and private one. On the right are the same quantities weighted by the quadratic fit of the ratio between $\tau_2/\tau_1$ of the official sample and the private one. One can observe a better agreement in the plot of the right. All plots correspond to the muon channel.
Multivariate Analysis techniques (MVA) have been used in high energy physics for many years with considerable success, and as analysis are becoming increasingly challenging, MVAs are becoming a new standard for physics analysis. The most typical areas of application are background suppression (classification) and parameter estimation (regression), where a physical quantity is extracted from a set of directly measured observables. In this work we only concentrate in classification using as input event and substructure variables. In section 5.1 we evaluate the results of using substructure variables for the discrimination of background and in section 5.2 we present the analysis done using Artificial Neural Networks (ANN) and different types of Boosted Decision Trees (BDT).

5.1 Jet substructure results

In sections 2.2 and 2.3 we described several jet substructure variables and grooming techniques that have a good background rejection for boosted objects like $W$ bosons. In this section we briefly present the results obtained comparing the validated $WW$ sample that was produced including filters at generator level with a $W + \text{jets}$ sample and evaluate which variables present the highest discrimination power. All the plots presented here correspond to the electron channel. The muon channel yields very similar results. Essentially, all variables behave in the same way for both muon and electron channel.

Plots shown in figure C.2 correspond to the N-subjettiness variables and Width, as explained in 2.2.1 and 2.2.1. The top plots correspond to $\tau_1$ and $\tau_2$, which as previously explained don’t have a good discriminating power by themselves. Only when combined
Figure 5.1: Plots on top show the variables $\tau_1$ (left) and $\tau_2$ (right). As it was explained, these two variables by themselves are not useful for discriminating between signal and background. Plots on bottom correspond to $\tau_2/\tau_1$ (left) and Width (right), here one observes that $\tau_2/\tau_1$ shows a high discriminating power while Width doesn’t.

($\tau_2/\tau_1$), as shown in the bottom left plot, they present a high discriminating power. In the bottom right the width is plotted. One would expect smaller values for the signal, but this doesn’t occur because of the mass-drop/filtering, which requires substructure, resulting in large jet widths.

Figure 5.2 shows several observables related to the mass drop/filter technique (explained in 2.2). The top plots show the mass corresponding to the two and three hardest subjets (from left to right). Both are good discriminants, but using only the two hardest subjets is not a valid option, since there’s a significant loss in the mass and therefore it doesn’t reconstruct the $W$ correctly. One could also compare the $p_T$ of the three hardest subjets, but this observable is considerably worse than the mass. The middle plots show, from left to right, $R_{j_1,j_2}$ (the distance between the two subjets when condition 2 is achieved) and $R_{filt}$ (defined as $\min(0.3, R_{j_1,j_2}/2)$). Although it doesn’t perform as well as the mass of the three subjets filtered, $R_{filt}$ is still a good discriminator between signal and background.
The two bottom plots correspond to the variables $\mu$ and $y$, they both present a poor discrimination power. Additional plots with more substructure and grooming variables can be found in appendix C.

**Figure 5.2:** All distributions correspond to different variables calculated with the mass-drop/filtering technique. Here, as expected, the mass of the filtered jet is a variable with high discriminating power, and also $R_{\text{filt}}$ is useful, while the rest of the variables aren’t significantly different.
Plots in figure 5.3 correspond to angularity and dipolarity (on top), planar flow and the fat jet mass over the $p_T$ ratio ($M_J/p_T$) (middle) and splitting scales and $Z_{cut}$ (bottom). We can observe that only the splitting scale variable ($\sqrt{d_{12}}$) presents a good separation between signal and background.

**Figure 5.3:** Distributions correspond to substructure variables, angularity and dipolarity (top), planar flow and fat jet mass over $p_T$ (middle) and splitting scales and $Z_{cut}$ (bottom). Only splitting scales shows a significant difference between signal and background.
5.2 Multivariate Analysis

In the previous section we showed several discriminating variables based on substructure and grooming techniques. Next we present a simple overview of the two methods used for the MVA, artificial neural networks and boosted decision trees. The analysis was done using the multivariate methods implemented in TMVA (Toolkit for Multivariate Data Analysis) [31].

Neural Networks

Originally, artificial neural networks were inspired by biological neural networks models, emulating their structure to process information. Nowadays, ANNs are mostly mathematical models based on statistics and optimization theory, leaving behind the comparison with the central nervous system [32][33].

![Figure 5.4: A schematic representation of feed-forward neural network with two hidden layers.](image_url)

A neural network is a nonlinear function \( f : x \rightarrow y \) which can model relationships between input and outputs using adjustable parameters called weights. This function \( f \) is a composition of other functions \( g_i \), hence the term "network", which can in turn be a composition of other functions \( h_i \), etc. Figure 5.4 displays a representation of this network in which each function is represented by a node and the arrows show the dependences between functions. The most common used functions are the nonlinear weighted sums:

\[
f(x) = K \left( \sum_i (w_i g_i(x')) \right)
\]

where \( x' \) might be again a composition of functions acting on the input vector \( x \) and \( K \) is called the activation function. The most common activation functions (all available in TMVA) are:
The simplest and most commonly used neural network is the Feedforward, since the information is fed from input to output without going through any loop (networks with loops are called recurrent). The most important feature of neural networks is that they can be trained, which means that using a set of observations as input, we obtain an optimal function \( f \). In order to obtain this optimal function, we have to first define the cost function \( C : F \to \mathbb{R} \), where \( F \) is the space of all functions \( f \). Then, the optimal solution \( f^* \) has to fulfill the relation

\[
C(f^*) \leq C(f) \quad \forall f \in F
\]  

(5.2)

Training algorithms search through the solution space in order to find a function that minimizes \( C \). Usually the cost function is based on the \( \chi^2 \) minimization

\[
C(f) = \frac{\sum_{i=0}^{N} (f(x_i) - y_i)^2}{N}
\]  

(5.3)

where \( y_i \) is the desired output.

In supervised training one has a set of observations \( (x, y) \), where \( x \) is a known input vector and \( y \) is the expected output (this can be taken from the Monte Carlo simulation where one knows the true output). In the present study, we provided as input to the neural network two samples simulated with Monte Carlo, one for signal and one for background and therefore ‘trained’ the neural network to recognize between signal and background.

The training process can be summarized as follows: we present patterns to the neural network which generates an output. Then, this output is compared with the desired output from the training sample and the cost function is calculated. Next, the weights in nodes should be adjusted to decrease the value of the cost functions. The algorithm that takes care of this is called the backpropagation algorithm, which propagates the errors back in the neural network from the output units to the hidden units. Therefore, at the output layers, the resulting vector is compared to the expected output. If the difference is not zero, then the error is calculated using the delta rule and is propagated back through the network. Using this delta rule, the change in weight of the link between nodes \( i \) and \( j \) is given as

\[
dw_{ij} = rx_i(t_j - y_j)
\]  

(5.4)
where \( r \) is the learning rate (an arbitrary parameter), \( t_j \) is the target output, \( y_j \) is the actual output at unit \( j \), and \( x_i \) is the actual output of the node \( i \) of the preceding layer. The delta rule changes the weight vector in a way that minimizes the error (the difference between the target output and the actual output) \([34][35]\).

**Boosted Decision Trees**

![Figure 5.5: A schematic of a binary decision tree](image)

Decision trees are a machine learning technique, that by combining several weak classifiers, creates a more powerful multivariate discriminant. Essentially, by comparing samples attributes to threshold values classifies unknown events into predetermined groups. At each step of the sequence, the best cut is searched for and used to split the data and this process is continued recursively on the resulting partitions until a given terminal condition is satisfied. The training starts with the root node (figure 5.5 shows a diagram of a simple decision tree), with the entire training data set containing signal and background events \([36]\). At each iteration of the algorithm, and for each node, one finds the best cut for each variable and then the best cut overall. Therefore, the data are split using the best cut, forming two branch nodes. One stops splitting when no further reduction in impurity is possible (or when the number of events is too small). Usually the common measure used to quantify impurity is the Gini index, defined by \([37]\),

\[
Gini = 2P(1 - P)
\]  \((5.5)\)
where the purity $P$ is

$$P = \frac{S}{S + B} = \frac{\sum_s w_s}{\sum_s w_s + \sum_b w_b}$$

and $S$ ($B$) is the weighted total number of signal (background) events which ended on the node during training. The splitting at a branch node is terminated if the impurity after the split is not reduced. The node then becomes a terminal node (leaf) and an output response.

Decision trees have been available for several years and are a very powerful technique, but they also have some limitations, for example, instability with respect to the training sample (a slightly different training sample can produce a dramatically different tree), poor global generalization because the recursive splitting results in the use of fewer and fewer training data per bin and only a small fraction of the feature variables may be used to model the predictions for individual bins. Fortunately, there are several techniques that overcome these problems. One of these techniques is boosting.

The idea of boosting is to make a sequence of classifiers that work progressively harder on increasingly “difficult” events. Instead of seeking one high performance classifier, one creates an ensemble of weak classifiers that collectively have a “boosted” performance. The boosted decision trees (BDTs) have been found to be very robust.

### 5.3 Training results

The neural network and boosted decision trees used in this analysis are the ones implemented in the TMVA. The three BDTs used are the classical one (BDT), with gradient boosting (BDTG) and with decorrelation + adaptive boost (BDTD). As input, we used different subsets of the substructure and grooming variables presented in section 5.1 and in appendix C. To see which set of variables had a better discrimination we analyzed the ROC curves. A ROC curve (receiver operating characteristic) is a graphical plot that illustrates the performance of a classifier system as its discrimination threshold is varied. In the end, after a certain number of variables, the significance didn’t improve much.

The neural network used was a Multi-layer perceptron (MLP). We tested different configurations of the neural network, changing the number of hidden layers, the activation function (we tested sigmoid and tanh) and the number of epochs (number of training cycles) between 600 and 1600. The final configuration was as follows

- Neuron type: tanh
- Hidden layers: $N, N - 1$
- Number of cycles: 750
Figure 5.6: Rejection curves for the four techniques used, Neural network, BDT, BDTG and BDTD. Both the MPL and BDTG have a better performance than BDT and BDTD.

Regarding the BDTs (BDT, BDTG and BDTD), as can be seen in figure 5.6, the BDTG method shows better results. The final configuration for the BDTG is as follows:

- NTrees: 1000
- MaxDepth of the decision tree allowed: 2
- Minimum percentage of training events required in a leaf node: 1.5%

Figure 5.6 shows the ROC curves of the four methods used. Although there’s not a significant difference, MLP and BDTG have a better performance than BDT and BDTD. As it was mentioned before, using different sets of variables doesn’t change much the results in the ROC curves, but in almost all cases MPL and BDTG have a better rejection.

Both Monte Carlo samples (signal and background) were divided in two parts. In the first place, training samples of signal and training samples of background were used to build the ANN and the BDTs. Once they were built, they were applied to the other half of samples of signal and background, independently of the training samples, and were used for ANN or BDT evaluation. Each testing object is evaluated and assigned an output score. Figure 5.7 shows the output assigned to an independent sample (test sample) of signal and background for the ANN and BDTG. A cut on this score can be used to distinguish between signal and background.
Figure 5.7: Distribution of the MVA discriminant outputs for the boosted decision tree (top left), neural network (top right), BDT with gradient boosting (bottom left) and BDT with decorrelation and adaptive boost (bottom right). Background presents a lower score than Signal, therefore, a cut on this score can be used to distinguish between signal and background.

Figure 5.8: Classifier cut efficiencies showing the significance \( \frac{S}{\sqrt{S + B}} \), signal and background efficiencies and signal purity as a function of the ANN (left) and BDT (right) cut value.
Figure 5.8 corresponds to the classifier cut efficiencies showing the signal and background efficiencies, the significance \((S/\sqrt{S+B})\) and signal purity. The significance is the parameter that indicates the performance of the training method more accurately. However, when choosing a cut value, the statistics must be taken into account, this means that we should be aware of the signal efficiency that corresponds to the optimal cut value. If this value is very low, a lot of signal will be cut away, and therefore will have low statistics in the real data. This is problematic because having low statistics in data means that we can’t be sure of the quality of the cut, since the output distribution of data is different from the one of the training sample. As a next step in the analysis, we could pick one working point from the ROC curve and compare the methods at that point (either choosing fixed efficiency or fixed rejection).
Conclusions

In this thesis we present the improvement of the event selection to identify $W$–pairs, with one boosted $W$ decaying hadronically and the other $W$ decaying leptonically. The event selection was done at two different levels.

The main goal of the thesis, the implementation of an event selection at generator level, was accomplished. Through several studies, first using the Pythia generator and then using the Herwig generator, we succeeded in obtaining fully simulated samples that included a preselection of events at generator level. This has several advantages, first, the percentage of events that remain after the full offline event selection is around 10%, while in the case of the official Monte Carlo sample this was around 0.4%. The impact of this is that one has to produce fewer events (which is very time-consuming), and also, it makes the event selection offline more efficient. We validated the fully simulated samples by comparing them with Monte Carlo samples that were already validated using data. This was done after applying the full event selection to both samples and comparing the most relevant kinematic and substructure variables between them. The results of this analysis and the private Monte Carlo validation can be seen in chapter 4.

In the second place, we improved the event selection for the offline analysis optimizing jet variables and cuts, event vetoes and isolation criteria. The final configuration of the event selection is shown at the end of chapter 3. As it was explained before, since boosted objects cannot be analysed with the usual jet algorithms, we implemented several substructure variables and grooming techniques, in order to have a good discrimination between signal and background. Several of these variables were very successful and one can observe a high discriminating power. The resulting distributions of substructure variables between signal and background are shown in the beginning of chapter 5. There, variables like $\tau_2/\tau_1$, filtered mass, $Kt$ splitting and several more are very promising to discriminate signal over background using for example a multivariate analysis technique.
Finally, we applied a multivariate analysis technique using different sets of substructure and grooming variables as input. For this, we used a neural network and boosted decision trees, changing different parameters. In order to see which set of variables and parameters were better discriminants, we used ROC curves (rejection curves that compare signal vs background efficiencies). Some preliminary results of this analysis are shown in the end of chapter 5. As a further analysis, one could pick one working point from the ROC curve and compare the methods at that point (either choosing fixed efficiency or fixed rejection).
Appendices
PYTHIA PLOTS

A.1 Signal truth (no pile up) vs signal MC

Figure A.1: Plots show the $p_T$ of the two leading small jets and the fat jet of the signal sample generated with PYTHIA including filters in the generator compared to the official signal Monte Carlo samples. These plots don’t show a good agreement, especially the $p_T$ of the fat jet.
A.2 Pile up generated with pythia

![Graphs showing pT and mass distributions for Min Bias events with small and fat jets.](image)

**Figure A.2:** Plots on the left show the \( p_T \) of the leading small jet (top) and fat jet (bottom) and plots on the right show their respective masses. The very low masses and transverse momenta observed are characteristic of very soft QCD pile up events.
Figure A.3: Mass and $p_T$ distributions for pile up generated with PYTHIA vs official Min Bias samples, divided in $p_T$ windows. As expected, the two samples become more similar with higher $p_T$, but for lower values the distributions show significant disagreements.
A.3 W+jets with pile up added:

![Histograms]

**Figure A.4:** Distributions of relevant substructure variables comparing W+jets produced with 
*Pythia* with pile up added at generator level with official Monte Carlo samples of W+jets fully simulated. These variables have a great discriminating power and therefore we want to have a good agreement in the results. Here we see that it’s not the case. This motivated us to find a different approach for the analysis, as described in section 4.3.2.
Figure A.5: The pulse shape in the ATLAS LAr calorimeters. The unipolar triangular pulse is the current pulse in the liquid argon generated by fast ionizing particles. Its characteristic time is the drift time (charge collection time) of approximately 450 ns in the example for the central EMB calorimeter shown here. The shaped pulse is superimposed, with a characteristic duration of about 600 ns. The full circles on the shaped pulse indicate the nominal bunch crossings at 25 ns intervals (Taken from [38]).
**HERWIG PLOTS**

**B.1 Validation of MC without filters**

![Eta and $p_T$ distributions of the leptons (left) and neutrinos (right) from an official MC sample validated and a our private sample generated with the same conditions (only a lepton filter).](image)

**Figure B.1:** Eta and $p_T$ distributions of the leptons (left) and neutrinos (right) from an official MC sample validated and a private sample generated with the same conditions (only a lepton filter).
Figure B.2: Eta, phi, $p_T$ and mass distributions of the leading small jets and fat jet comparing an official MC sample already validated and our privately generated sample using the same conditions.
B.2 Validation of MC with filters

Figure B.3: Phi, eta, mass and $p_T$ distributions of the leading small jet in the electron channel (left) and muon channel (right), comparing our private MC sample with the event selection at generator level and a MC sample already validated with data.
Figure B.4: Phi, eta, mass and $p_T$ distributions of the leading fat jet in the electron channel (left) and muon channel (right), comparing our private MC sample with the event selection at generator level and a MC sample already validated with data.
Figure B.5: Phi, eta and $p_T$ distributions of the leading electron (left) and muon (right), comparing our private MC sample with the event selection at generator level and a MC sample already validated with data.
Figure B.6: $E_{T}^{\text{miss}}$ and transverse mass ($m_T$) distributions in the electron channel (left) and muon channel (right), comparing our private MC sample with the event selection at generator level and a MC sample already validated with data.
Figure B.7: $p_T$, mass and $\Delta R_{12}$ dijet distributions in the electron channel (left) and muon channel (right), comparing our private MC sample with the event selection at generator level and a MC sample already validated with data.
Figure B.8: Substructure variables. Mass and $p_T$ of the filtered 2 sbjets (top) and 3 sbjets (bottom) in the electron channel (left) and muon channel (right), comparing our private MC sample with the event selection at generator level and a MC sample already validated with data.
Figure B.9: Substructure variables. Mass-drop/filtering variables $\mu, y, R_{j1,j2}$ and $R_{\text{filt}}$ distributions in the electron channel (left) and muon channel (right), comparing our private MC sample with the event selection at generator level and a MC sample already validated with data.
Figure B.10: Substructure variables. $N$—subjets variables ($\tau_1$, $\tau_2$ and $\tau_2/\tau_1$) mass over $p_T$ ratio distributions in the electron channel (left) and muon channel (right), comparing our private MC sample with the event selection at generator level and a MC sample already validated with data.
Figure B.11: Substructure variables. Angularity, dipolarity, planar flow and Width distributions in the electron channel (left) and muon channel (right), comparing our private MC sample with the event selection at generator level and a MC sample already validated with data.
Figure B.12: Substructure variables. Splitting scales, $Z_{\text{cut}}$ and energy correlation distributions in the electron channel (left) and muon channel (right), comparing our private MC sample with the event selection at generator level and a MC sample already validated with data.
Figure C.1: Plots on top show the ratio between the mass of the mass-dropped/filtered jet and the fat jet (left) and the ratio between the $p_T$ of the filtered jet and the $p_T$ of the fat jet (right). Plots on the bottom show the ratio between the mass of the trimmed jet and the mass of the fat jet (left) and the ratio between the $p_T$ of the trimmed jet and the $p_T$ of the fat jet (right).
Figure C.2: Plots on top show the ratio between the mass of the pruned jet and the mass of the fat jet (left) and the ratio between the $p_T$ of the pruned jet and the $p_T$ of the fat jet (right). Plot on bottom corresponds to the number of filtered subjets with $p_T > 10$ GeV.
BIBLIOGRAPHY


[38] Atlas Collaboration. Topological cell clustering in the ATLAS calorimeters and its 