Karlsruhe Institute of Technology

Statistical Combination of Higgs Decay Channels and Determination of the Jet-Energy Scale of the CMS Experiment at the LHC

Zur Erlangung des akademisches Grades eines
DOKTORS DER NATURWISSENSCHAFTEN
von der Fakultät für Physik des
Karlsruher Institut für Technologie eingereichte

DISSERTATION

von

Dott. Danilo Piparo
aus Mailand

Mündliche Prüfung: 12. November 2010

Referent: Prof. Dr. G. Quast
Institut für Experimentelle Kernphysik

Korreferent: Prof. Dr. T. Müller
Institut für Experimentelle Kernphysik

Todo cuanto necesitamos saber está escrito en el gran libro de la naturaleza.
Basta con tener la valentía y la claridad de mente y espiritú para leerlo.
Andreas Corelli

Statistical Combination of Higgs Decay Channels and Determination of the Jet-Energy Scale of the CMS Experiment at the LHC.

At the Large Hadron Collider (LHC), numerous events containing a Z boson decaying into two muons are produced. In many of these events, the Z boson is boosted and balanced in transverse momentum by exactly one jet. Since the kinematical properties of the muons can be measured very precisely with the Compact Muon Solenoid (CMS) detector, such events are ideal candidates for a data driven technique for the absolute jet energy scale determination and calibration.
The accurate knowledge of the jet energy scale is crucial for many LHC analyses, as it represents very often the dominating systematic uncertainty. The prospects for calibration with $Z(\rightarrow \mu^+\mu^-)$+jet events show that with $200\,\mathrm{pb}^{-1}$ of acquired data, which will be collected by CMS at the beginning of 2011, an absolute jet energy calibration can be performed up to transverse momenta of 160 GeV. The $Z(\rightarrow \mu^+\mu^-)$+jet technique is used for the first time with $2.96\,\mathrm{pb}^{-1}$ of data acquired with the CMS detector at a centre of mass energy of 7 TeV. The jet energy scale determined with respect to the Monte Carlo prediction proves the need for a data driven calibration of jet energies at CMS.
Underlying event and pileup contributions alter the jet energy scale, generating additional activity in the events together with the principal scattering process. A strategy involving jet areas for the subtraction of this unwanted contamination on an event-by-event and jet-by-jet basis is tested on CMS data. Even if up to now a proof of concept of the technique was given only considering generator particles, the technique is proved to be applicable also to detector data.

One of the main goals of the LHC physics program is the discovery or exclusion of the Standard Model Higgs boson. This can only happen if searches in all possible Higgs decay channels are combined in order to obtain the maximum statistical power. Two techniques for discoveries and exclusions are discussed, namely the profile likelihood and the hypothesis testing methods, together with two strategies for systematic uncertainties treatment, marginalisation and profiling. With the aid of the RooStatsCms framework for analysis modelling, combination and statistical studies, the first Higgs analyses combination exercises are performed.

Acknowledgements

For having accepted me in his group three years ago, for being always there to give advice about the research line to follow, for creating such a pleasant atmosphere in our workplace, for giving to all of his collaborators the freedom and motivation necessary to develop their skills, my sincere thank goes to my supervisor: Professor Günter Quast.

I want also to thank professor Thomas Müller for accepting to be the co-referee of this thesis, for his extraordinary way of managing the whole institute and for giving me the possibility to spend more than one year of my PhD studies at CERN.

My sincere thank goes also to Dr. Klaus Rabbertz, for all that he taught me and for being a real guide and source of precious input for this work.

I must confess that I don't know how I could have achieved my results without the collaboration of all the members of the group. I thank Michael Heinrich, for being the first person believing in my abilities in the field of spoken German and for all the incredibly fruitful conversations we had, Oliver Oberst for helping me when I was lost, Dr. Andreas Oehler for the brilliant and frank exchanges of opinions and all the advice he gave about this text, Christophe Saout and Fred Stober for sharing with me their unprecedented technical skills, Manuel Zeise for his practical sense and kindness in everything, Dr. Volker Büge and Dr. Armin Scheurer for their friendly guidance during these three years, Christoph Hackstein for his vast knowledge of physics and calm in every occasion, Dr. Grégory Schott, Matthias Wolf, Joram Berger and Thomas Hauth for their new ideas, proposals and hard work.
My thank goes also to Dr. Lorenzo Moneta, Dr. Wouter Werkerke and all the RooStats team.
I am grateful to Dr. Roberto Salerno and Dr. Guillelmo Gomez Ceballos, the first power users of RooStatsCms.
I thank Dr. Pietro Govoni, professor Marco Paganoni and all the members of the group of Milano Bicocca for the productive collaboration.

I also thank the administrators of the institute, for keeping the computing infrastructure always running and the National Analysis Facility (NAF) team for the high quality service provided.

I thank also all my friends in Milan, for being always there. I am grateful to my family and my parents, for always endorsing my choices. Finally I thank Lara Bochsler, for being who she is.

Contents

Introduction

The scope of particle physics is the description and understanding of phenomenological manifestations of the fundamental entities of matter and forces. The key to reach this knowledge is the interplay between theory and experiment, joint to the development of technology. Theories are proposed to elucidate and predict observations of natural phenomena and, in return, more and more advanced experiments are performed to corroborate or falsify those assumptions. During the past century, this kind of scientific progress brought the Standard Model of particle physics to its present form: An extremely successful theory that describes the fundamental particles and their interactions with unprecedented precision.

An eminent example among the achievements of the Standard Model is the discovery of the predicted carriers of the weak force, the W and Z bosons, that took place at the Super Proton anti-Proton Synchrotron at CERN ("Conseil Européen pour la Recherche Nucléaire") in 1983. The study of the Z boson then became the driving factor of the physics program of the Large Electron Positron collider. In the 1990s, the experiments at this machine delivered the most accurate measurements of the Z boson properties.

In the context of present collider experiments, the role of particles like W and Z bosons, which were object of the discoveries of the past, evolves into the one of reliable standard candles. The Z boson can now be exploited for crucial detector commissioning, alignment and calibration purposes.

Nevertheless, the Standard Model leaves several open questions. Indeed, there is no experimental evidence for the Higgs boson, which is the entity designated by the Standard Model theory for the description of the mass assignment to particles. In addition, anomalies which are not explained in the Standard Model were already observed. For instance, neutrinos are assumed to be massless, yet their observed oscillations imply that they are not. Astronomical observations strongly indicate distributions of invisible matter in the universe, the Dark Matter, which cannot find a description in the Standard Model. Moreover, the accelerating expansion of the universe s the point about an agent which could drive such a process, the Dark

Energy. These days are very exciting since the community might be on the edge of a scientific revolution which could transform radically the current paradigms of particle physics. Being the most powerful accelerator ever built, the Large Hadron Collider ("LHC") has a crucial role in this process.

This machine, which collides protons as well as heavy ions, started its operations at the end of 2009. The interactions between these particles mostly involve quantum chromodynamics ("QCD") processes. Such phenomena are interesting per se since new insights about the proton structure and the dynamics of its constituents, quarks and gluons, can be gained at the LHC energy scale. Furthermore, these processes represent the predominant background for several analyses.

Due to their confinement, quarks and gluons which are scattered outside the protons after a collision, hadronise and originate a collimated spray of particles observed in the detectors. In order to group these particles and infer the four-momenta of the initiators of such showers, the concept of jets is introduced. However, the energy reconstruction of jets in the detector is not equal to the one of the originating quark or gluon because of several effects.
A precise estimation of the jet energy scale is fundamental and can be a crucial factor for a discovery. This is the reason why several methods to calibrate jets were studied up to now. In this thesis, the calibration of the absolute jet energy scale which makes use of events featuring a Z boson produced in association with jets is discussed. For the first time, this strategy is studied with the data acquired by the CMS ("Compact Muon Solenoid") experiment at the LHC.

In an LHC event, every high energetic scattering between two of the components of the colliding protons is accompanied by several soft interactions among other spectator gluons and quarks in the colliding protons (underlying event).
Moreover, the dense beams circulating in the LHC give rise to several proton-proton collisions (pileup) each time they are crossed. All this additional activity results in an unwanted energy supplement in all jets in an event. A strategy for the subtraction of this additional contributions on an event-by-event and jet-by-jet basis which takes advantage of the concept of jet areas is discussed and, for the first time, evaluated through its application to the data acquired by CMS.

In chapter 1, the basic concepts of the Standard Model are discussed. The motivations for the existence of a Higgs boson and for the masses of the weak bosons W and Z are depicted. In addition, a concise introduction to QCD is given. Eventually, the production mechanisms of the Z boson in association with jets and of the Higgs boson are summarised. After this theoretical introduction, in chapter 2, the features of the CMS detector at the LHC are characterised.

The CMS software framework is described in chapter 3. The reconstruction of tracks, muons and jets is discussed in detail. In addition, particular emphasis is given to the software components that profited from contributions originated from

the work described in this thesis.

The discovery or exclusion of the Standard Model Higgs boson can only take place if different decay channels are combined and advanced statistical methods are deployed. Chapter 4 is endowed to the description of such statistical methods, the treatment of systematic uncertainties and the combinations of analyses that took place in the context of this thesis.

The approach adopted by CMS for the jet energy corrections is described in chapter 5. The absolute jet energy scale calibration exploiting the balance of a Z boson decaying into two muons and a jet is discussed. A description of the cuts applied to isolate the required topology, the possible backgrounds and the systematic uncertainties is given. The prospects of such a calibration are presented using Monte Carlo simulations. Using CMS data, cross-checks of the quality of the physics objects involved in the analysis are performed and the first estimation of the jet energy scale using the $Z(\rightarrow \mu^+\mu^-)$+jet topology is carried out.

In chapter 6, a jet area based strategy for underlying event and pileup subtraction acting event-by-event and jet-by-jet is deployed exploiting jets formed by tracks reconstructed in the detector. For the first time, a proof of concept of the strategy is given with CMS acquired data.

Chapter 1

The Standard Model of Particle Physics

1.1 Introduction

The Standard Model of particle physics is a theory that describes the electroweak and strong interactions of elementary particles in a quantum relativistic frame. It is the result of the interplay between theory and experiments that took place during the 20^{th} century. Its predictions have been verified with high precision during decades of observations with experiments i.a. at particle colliders (see [1] for a summary).
According to the Standard Model, all matter is made out of three kinds of elementary particles, namely leptons and quarks, force mediators (the vector bosons) plus one scalar boson, the Higgs particle which is involved in the mechanism that accounts for particle masses (figure 1.1).
An impressive amount of ideas, experimental techniques and discoveries involving many of the greatest scientists of the past century coalesced to form the theory in its present formulation, from the postulation of the light quantum to describe black body radiation up to the discovery of W and Z bosons at CERN [2].

Standard Model limitations

Despite its great success, quite a number of hints suggest that in many cases the Standard Model is not the final word on the description of nature. The theory has many adjustable parameters, for example the values of the particle masses. Their values are not predicted, but rather parameters provided by experiments. Moreover, the pattern of masses is very irregular. In addition to that, neutrinos are predicted to be massless, but many oscillation measurements [3] prove that their mass is definitively not zero. Furthermore, cosmological observations demonstrate that only about 4 % of the mass-energy in the universe consists of ordinary baryonic matter [4]. Another weakness of the Standard Model is that gravity is not described in the current theoretical framework.

The LHC could revolutionise our understanding of these frontier topics and also

open new paths beyond the present paradigms.

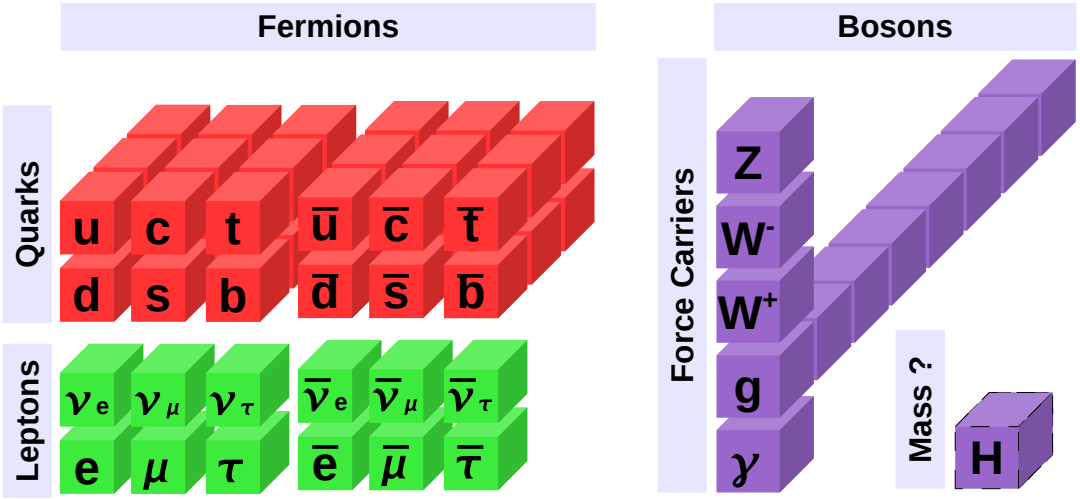

Figure 1.1: The particles in the Standard Model are 12 leptons and anti-leptons, 36 quarks and anti-quarks, 12 force mediators plus 1 particle responsible for the mass of the fields describing particles, the Higgs boson.

1.2 Relativistic Field Theories and Lagrangian Formalism

In classical mechanics, the Euler-Lagrange equations for the generalised coordinates labelled with the index i can be written as

$$\frac{d}{dt}\left(\frac{\partial L}{\partial \dot{q}_i}\right) - \frac{\partial L}{\partial q_i} = 0 \tag{1.1}$$

where the Lagrangian L as a function of time, the generalised coordinates $\mathbf{q}$ and their derivatives $\dot{\mathbf{q}}$ is defined as

$$L = L(\mathbf{q}, \dot{\mathbf{q}}, t) = T(\dot{\mathbf{q}}) - V(\mathbf{q}, \dot{\mathbf{q}}, t) \tag{1.2}$$

where T represents the kinetic energy and V the potential.
A modern relativistic description of microscopic entities is provided by quantum field theory (QFT). In this context, particles are not described as point-like objects but as fields, i.e. complex functions of position and time:

$$\phi = \phi(x_\mu), \mathbb{R}^4 \rightarrow \mathbb{C} \tag{1.3}$$

where x_μ is the position time four-vector. The formalism used to describe these fields is also Lagrangian since it allows to treat space and time on equal footing.

Lagrangian Density

It is convenient to replace the Lagrangian function by the Lagrangian density

$$L(\mathbf{q}, \dot{\mathbf{q}}, t) \rightarrow \mathcal{L}\left(\boldsymbol{\phi}, \partial_\mu \boldsymbol{\phi}, x_\mu\right), \quad L = \int d^3x\, \mathcal{L} \tag{1.4}$$

therefore the Euler-Lagrange equations generalise to

$$\partial_\mu \left(\frac{\partial \mathcal{L}}{\partial(\partial_\mu \phi_i)} \right) - \frac{\partial \mathcal{L}}{\partial \phi_i} = 0. \tag{1.5}$$

It should be noted that the Lagrangian density in equation 1.4 is often also called simply Lagrangian, even if this terminology is not completely correct.

The Dirac equation describes free Spin-$\frac{1}{2}$ fermions as complex Dirac spinors with four components (see appendix A for the details about the notation):

$$(i\gamma^\mu \partial_\mu - m)\psi = 0, \ \psi = \begin{pmatrix} \psi_1 \\ \psi_2 \\ \psi_3 \\ \psi_4 \end{pmatrix} \ (\hbar = c = 1) \tag{1.6}$$

and the corresponding Lagrangian is:

$$\mathcal{L} = \bar{\psi}(i\gamma_\mu \partial^\mu - m)\psi. \tag{1.7}$$

Two other examples of Lagrangians can be the Klein-Gordon one describing a free scalar (Spin-0) field ϕ:

$$\mathcal{L} = \frac{1}{2}\left[(\partial_\mu \phi)(\partial^\mu \phi) - m^2 \phi^2\right], \tag{1.8}$$

and the Proca Lagrangian for a free massive Spin-1 vector field A^μ:

$$\mathcal{L} = -\frac{1}{4} F^{\mu\nu} F_{\mu\nu} + \frac{1}{2} m^2 A^\mu A_\mu \tag{1.9}$$

where $F^{\mu\nu}$ is the electromagnetic tensor $\partial^\mu A^\nu - \partial^\nu A^\mu$.

1.3 Local Gauge Invariance

A gauge transformation of a complex field is a unitary transformation changing its phase. When this change is dependent on space-time coordinates, the transformation is referred to as *local gauge transformation*. Transformations among gauges form gauge or symmetry groups, for example the $U(1)$ group [5]. Spinor fields, under the effect of the elements of $U(1)$ transform as

$$\psi \rightarrow \psi' = e^{iq\theta(x_\mu)}\psi \tag{1.10}$$

where q is called *generator* of the transformation. A fundamental requirement in QFT, is the local gauge invariance of Lagrangians. The Dirac Lagrangian (equation 1.7), after the transformation 1.10, presents an unphysical extra term dependent on the arbitrary function $\theta(x_\mu)$

$$\mathcal{L} = \bar{\psi}(i\gamma_\mu\partial^\mu - m)\psi - q\bar{\psi}\gamma^\mu\psi\partial_\mu\theta(x_\mu). \tag{1.11}$$

Restoring the invariance involves the cancellation of this new term.

Covariant Derivatives

To achieve this goal, covariant derivatives are introduced:

$$\partial_\mu \rightarrow D_\mu = \partial_\mu + iqA_\mu \tag{1.12}$$

where A_μ is called *gauge field*, a vector field which transforms as follows:

$$A_\mu \rightarrow A'_\mu = A_\mu - \frac{1}{q}\partial_\mu\theta(x). \tag{1.13}$$

This leads to the new invariant Lagrangian

$$\mathcal{L} = \bar{\psi}(i\gamma^\mu\partial_\mu - m)\psi - q\bar{\psi}\gamma^\mu\psi A_\mu \tag{1.14}$$

where the first term describes the free Spin-$\frac{1}{2}$ fermions and the second their interaction with a vector field which has to be massless since an extra mass term $1/2m^2A^\nu A_\mu$ like in equation 1.9 would spoil the gauge invariance. This field can be consequently identified as the electromagnetic field, the photon, and the generator q as the electric charge of the fermions. The presence of the vector field is not a coincidence: In general one vector field per transformation generator is necessary to preserve gauge invariance.
Therefore, requiring the Dirac Lagrangian (equation 1.7) to be invariant under local gauge transformations of the $U(1)$ group led to the introduction of a massless vector field which interacts with the spinor fields. This new construction, together with the quantisation of the electromagnetic field, can naturally accommodate the description of Quantum Electrodynamics (QED) if a free photon term is added to form the Lagrangian

$$\mathcal{L}_{\mathcal{QED}} = \bar{\psi}(i\gamma^\mu\partial_\mu - m)\psi - q\bar{\psi}\gamma^\mu\psi A_\mu - \frac{1}{4}F^{\mu\nu}F_{\mu\nu} \tag{1.15}$$

$\alpha_m(\mathbf{Q})$

Measuring the value of the electric charge of the electron is a way by which QED can be tested. In high energy physics, a well known process of electrodynamics takes place, namely the effect of vacuum polarisation that partially screens the charge of the electron. The screening and, thus, the measured charge and, ultimately, the coupling constant $\alpha_{em} = e^2/4\pi$, depend on the distance between

the interacting particles, which is correlated with the transferred momentum in the scattering process, Q. The higher the energy regime, and, thus, the smaller the distances between the entities taking part in the interactions, the higher are the values of the electric coupling constant α_{em} . For example, $\alpha_{em}(0) \simeq \frac{1}{137}$ and at energies of the order of the Z mass, $\alpha_{em}(M_Z) \simeq \frac{1}{128}$.

Bilinear Covariants and Projectors

It is interesting to observe that the term $\bar{\psi}\gamma^\mu\psi$ in the Lagrangian represents a vector. More formally, it is part of the family of the bilinear covariants that can be built combining the Dirac spinors and the Dirac matrices (see table 1.1).

Table 1.1: The bilinear covariants that can be built with the Dirac spinors.

Expression	Name
$\bar{\psi}\psi$	scalar
$\bar{\psi}\gamma^5\psi$	pseudo-scalar
$\bar{\psi}\gamma^\mu\psi$	vector
$\bar{\psi}\gamma^\mu\gamma^5\psi$	pseudo-vector
$\bar{\psi}\sigma^{\mu\nu}\psi$	antisymmetric tensor

Another use of the Dirac matrices, is the representation of a special class of operators that act on spinors, the projectors. They can be expressed as

$$P_L = \frac{1}{2}(1-\gamma^5) \qquad P_R = \frac{1}{2}(1+\gamma^5), \tag{1.16}$$

respectively for the left handed and right handed projection. Such objects allow to express the spinors as sum of a left handed and right handed component:

$$\psi = I\psi = (P_L + P_R)\psi = \psi_L + \psi_R = \frac{1}{2}(1-\gamma^5)\psi + \frac{1}{2}(1+\gamma^5)\psi. \tag{1.17}$$

This formalism turns out to be very useful in the description of the weak interactions (see section 1.5).

1.4 Spontaneous Symmetry Breaking and Higgs Mechanism

Another important concept in the Standard Model is the spontaneous symmetry breaking. To illustrate this procedure, it is useful to start from the $U(1)$ gauge invariant Klein-Gordon Lagrangian to which a fourth degree potential is added

$$\mathcal{L} = (D^\mu\phi)^*(D_\mu\phi) - \mu^2\phi^*\phi - \lambda(\phi^*\phi)^2 - \frac{1}{4}F^{\mu\nu}F^{\mu\nu}, \tag{1.18}$$

where $\lambda > 0$ and $\mu^2 < 0$. The complex field ϕ can be expressed as the composition of the two real fields ϕ_1 and ϕ_2, so that

$$\phi = \frac{\phi_1 + i\phi_2}{\sqrt{2}} \qquad \phi^*\phi \ = \frac{\phi_1^2 + \phi_2^2}{2}. \tag{1.19}$$

The potential has a symmetrical "Mexican hat" shape (formula 1.2), and has a set of minima with a radius

$$v = \sqrt{-\frac{\mu^2}{\lambda}}. \tag{1.20}$$

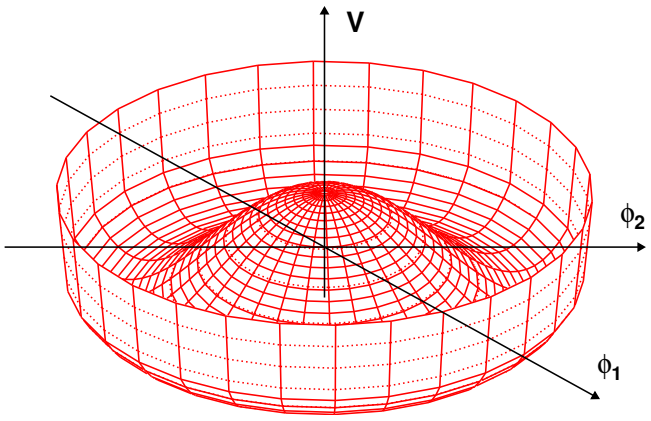

Figure 1.2: The potential $V(\phi) = \mu^2\phi^*\phi + \lambda(\phi^*\phi)^2$, for $\mu^2 < 0$ and $\lambda > 0$. A circle of degenerated minima is present.

Since all the calculations carried out in the Standard Model are perturbations around the vacuum ground energy, it is convenient to write the field ϕ as:

$$\phi(x) = \frac{1}{\sqrt{2}}\left[v + \eta(x_\mu) + i\xi(x_\mu)\right], \ \eta(x_\mu), \xi(x_\mu) \in \mathbb{R} \tag{1.21}$$

therefore changing the origin of the coordinates. This gives rise to an expression of the form

$$\begin{aligned}\mathcal{L}' =& \frac{1}{2}(\partial_\mu\xi)^2 + \frac{1}{2}(\partial_\mu\eta)^2 - v^2\lambda\eta^2 + \frac{1}{2}q^2v^2A_\mu^2 \\ & - qvA_\mu\partial^\mu\xi - \frac{1}{4}F_{\mu\nu}F^{\mu\nu} + \text{interaction terms.}\end{aligned} \tag{1.22}$$

Goldstone Bosons

It becomes clear that this Lagrangian describes a system with a massless boson ξ, called *Goldstone boson*, a massive scalar field η with mass $m_\eta = \sqrt{2\lambda v^2}$ and

a massive vector field A^μ with mass $M_A = qv$. The presence of the Goldstone boson is not accidental. Indeed, for every continuous symmetry of a Lagrangian which is spontaneously broken, the theory must contain massless scalar particles (*Goldstone Theorem* [6]). The Lagrangian 1.18 implied a continuous symmetry transformation, namely a rotation around the potential axis. After expanding the scalar fields as perturbations around the chosen ground state (equation 1.21), the Lagrangian lost its rotational invariance and a term associated to the Goldstone boson appeared.

Higgs Mechanism

A real improvement to this construction is the Higgs Mechanism. Indeed, through this procedure it is possible to re-absorb the spurious degree of freedom represented by the Goldstone boson, making it disappear from the Lagrangian. This solves the problem of the non-observation of a massless particle which could represent a candidate for the Goldstone boson. Another difficulty which is removed is the unphysical term $qvA_\mu\partial^\mu\xi$ representing an interaction that turns ξ into A. A transformation at first order equivalent to the one in equation 1.21, can be introduced to make the Lagrangian more physical, eliminating the two aforementioned issues

$$\phi(x_\mu) = \frac{1}{\sqrt{2}}\left[v + h(x_\mu)\right]\exp[i\frac{\theta(x_\mu)}{v}] \tag{1.23}$$

where the h field is real and represents the Higgs boson. The new Lagrangian will therefore read

$$\begin{aligned}\mathcal{L}'' =& \frac{1}{2}(\partial_\mu h)^2 - \lambda v^2 h^2 + \frac{1}{2}q^2v^2A_\mu^2 - \lambda v h^3 - \frac{1}{4}\lambda h^4 \\ &+ \frac{1}{2}q^2A_\mu^2h^2 + vq^2A_\mu^2h - \frac{1}{4}F_{\mu\nu}F^{\mu\nu}\end{aligned} \tag{1.24}$$

where only two massive particles appear, a vector boson A and the scalar Higgs boson h. The degree of freedom represented by the Goldstone boson is still present in the theory, but as the third polarisation of the A_μ field.
Therefore, the sacrifice of a manifest symmetry of the Lagrangian allowed to make the physical content of the Lagrangian more transparent.

1.5 The Glashow-Weinberg-Salam Model

The idea underlying the Glashow-Weinberg-Salam (GWS) theory is to choose a Lagrangian invariant under transformations of the $SU(2) \times U(1)$ group, in order to unify the description of weak and electric forces. This group foresees four generators, to be able to account for the four vector bosons fields $W^\pm$, Z and γ, and describe their interactions. In the following, only leptons will be considered. A full treatment involving quarks can be found in [5].

Weak Isospin and Hypercharge

First of all, values of weak hypercharge Y and weak isospin T are assigned to the fermions. These two quantities can be thought as mere analogues of the electric charge. The hypercharge of a fermion is defined as $Y = 2(Q - T_3)$ where T_3 is the third component of the isospin vector. Then, lepton generations are arranged in doublets and singlets. The doublets are called left-handed weak isospin doublets and are indicated with the ψ_L symbol. They have hypercharge $Y_L = -1$ and isospin $T = 1/2$: The third component of the isospin is therefore either $1/2$ or $-1/2$

$$\begin{array}{lccc} & & & T_3 \\ T = 1/2: & \begin{pmatrix} \nu_e \\ e^- \end{pmatrix}_{\mathrm{L}} & \begin{pmatrix} \nu_\mu \\ \mu^- \end{pmatrix}_{\mathrm{L}} \quad \begin{pmatrix} \nu_\tau \\ \tau^- \end{pmatrix}_{\mathrm{L}} & \begin{matrix} +1/2 \\ -1/2 \end{matrix} \end{array} \tag{1.25}$$

The singlets are called right handed isospin singlets, indicated with the symbol ψ_R and have isospin $T = 0$ and hypercharge $Y_R = -2$

$$T = 0: \quad e^-_{\mathrm{R}} \quad \mu^-_{\mathrm{R}} \quad \tau^-_{\mathrm{R}}. \tag{1.26}$$

The Dirac Lagrangian is therefore required to be invariant under local gauge transformations of the type

$$\psi_{\mathrm{L}} \rightarrow \psi'_{\mathrm{L}} = \exp\left[i\left(\alpha(x_\mu)\mathbf{T} + \beta(x_\mu)Y\right)\right]\psi_{\mathrm{L}} \tag{1.27}$$

for left-handed doublets and of the type

$$\psi_{\mathrm{R}} \rightarrow \psi'_{\mathrm{R}} = \exp\left[i\beta(x_\mu)Y\right]\psi_{\mathrm{R}}, \tag{1.28}$$

for right-handed singlets. Where generators of the $SU(2)$ and $U(1)$ transformations are $\mathbf{T} = (T_1, T_2, T_3)$ and Y. The covariant derivative to be introduced to preserve local gauge invariance is

$$D_\mu = \partial_\mu + ig\mathbf{T}\cdot\mathbf{W}_\mu + i\frac{g'}{2}YB_\mu, \tag{1.29}$$

where $\mathbf{W}_\mu = (W^1_\mu, W^2_\mu, W^3_\mu)$. This differential operator therefore results to be for left-handed doublets

$$\mathbf{T} \equiv \frac{\boldsymbol{\tau}}{2}, \quad Y = -1 \quad \Rightarrow \quad D_\mu = \partial_\mu + i\frac{g}{2}\boldsymbol{\tau}\cdot\mathbf{W}_\mu - i\frac{g'}{2}B_\mu, \tag{1.30}$$

where $\boldsymbol{\tau}$ represents the Pauli matrices (see appendix A), and for right handed singlets

$$T = 0, \quad Y = -2 \quad \Rightarrow \quad D_\mu = \partial_\mu - ig'B_\mu. \tag{1.31}$$

The resulting Lagrangian therefore reads

$$\begin{aligned}\mathcal{L}_1 =&\bar{\psi}_{\mathrm{L}}\gamma^{\mu}[i\partial_{\mu} - \frac{g}{2}\boldsymbol{\tau}\cdot\mathbf{W}_{\mu} + \frac{g'}{2}B_{\mu}]\psi_{\mathrm{L}} \\ &+ \bar{\psi}_{\mathrm{R}}\gamma^{\mu}[i\partial_{\mu} + g'B_{\mu}]\psi_{\mathrm{R}} - \frac{1}{4}\mathbf{W}_{\mu\nu}\cdot\mathbf{W}^{\mu\nu} - \frac{1}{4}B_{\mu\nu}B^{\mu\nu}.\end{aligned} \tag{1.32}$$

Higgs Potential

Keeping that in mind, the Higgs mechanism can be put in place adding to the Lagrangian 1.32 the following gauge invariant potential

$$\mathcal{L}_2 = \left|(\partial_{\mu}\phi + i\frac{g}{2}\boldsymbol{\tau}\cdot\boldsymbol{W}_{\mu}\phi + i\frac{g'}{2}YB_{\mu}\phi)\right|^2 - \mu^2\phi^{\dagger}\phi - \lambda(\phi^{\dagger}\phi)^2 \tag{1.33}$$

for the scalar field doublet ϕ

$$\phi = \begin{pmatrix}\phi_{\alpha}\\ \phi_{\beta}\end{pmatrix} = \frac{1}{\sqrt{2}}\begin{pmatrix}\phi_1 + i\phi_2\\ \phi_3 + i\phi_4\end{pmatrix}. \tag{1.34}$$

which has a hypercharge $Y = 1$. To make the masses for the vector bosons appear like in Lagrangian 1.24, a vacuum expectation value of the ϕ field is chosen

$$\phi_0 = \frac{1}{\sqrt{2}}\begin{pmatrix}0\\ v\end{pmatrix} \tag{1.35}$$

for the Higgs doublet and substituted into 1.33. Carrying out the necessary calculations, the relevant term results in

$$\begin{aligned}&\left|\left(i\frac{g}{2}\boldsymbol{\tau}\cdot\mathbf{W}_{\mu} + i\frac{g'}{2}B_{\mu}\right)\frac{1}{\sqrt{2}}\begin{pmatrix}0\\ v\end{pmatrix}\right|^2 \\ &\qquad = \frac{1}{4}v^2g^2W_{\mu}^{+}W^{-\mu} + \frac{1}{8}v^2\left[g^2(W_{\mu}^3)^2 - 2gg'W_{\mu}^3B^{\mu} + g'^2B_{\mu}^2\right]\end{aligned} \tag{1.36}$$

where the field combination $W^{\pm} = (W^1 \mp iW^2)/\sqrt{2}$ is employed to describe the W bosons. Comparing the first term with the expected mass term for a charged gauge boson $M_{\mathrm{W}}^2W^+W^-$ yields in

$$M_{\mathrm{W}} = \frac{1}{2}vg. \tag{1.37}$$

The γ and Z^0 bosons are described by linear combination of the B_{μ} and W_{μ}^3 to form a massless and a massive field:

$$\begin{aligned}\frac{1}{8}v^2\left[g^2(W_{\mu}^3)^2 - 2gg'W_{\mu}^3B^{\mu} + g'^2B_{\mu}^2\right] &= \frac{1}{8}v^2[gW_{\mu}^3 - g'B_{\mu}]^2 + 0[g'W_{\mu}^3 + gB_{\mu}]^2 \\ &= \frac{1}{2}M_{\mathrm{Z}}^2Z_{\mu}^2 + \frac{1}{2}M_A^2A_{\mu}^2,\end{aligned} \tag{1.38}$$

where, normalising the fields,

$$\begin{aligned} M_A = 0 \qquad &\Rightarrow \quad A_\mu = \frac{g' W_\mu^3 + g B_\mu}{\sqrt{g^2 + g'^2}} \\ M_{\mathrm{Z}} = \frac{v}{2}\sqrt{g^2 + g'^2} \qquad &\Rightarrow \quad Z_\mu = \frac{g W_\mu^3 - g' B_\mu}{\sqrt{g^2 + g'^2}} \end{aligned} \tag{1.39}$$

Weinberg Angle

The fields A_μ and Z_μ can be re-expressed in terms of an angle

$$\theta_{\mathrm{W}} = \arctan\frac{g'}{g} : \qquad \begin{aligned} A_\mu &= \cos\theta_{\mathrm{W}} B_\mu + \sin\theta_{\mathrm{W}} W_\mu^3 \\ Z_\mu &= -\sin\theta_{\mathrm{W}} B_\mu + \cos\theta_{\mathrm{W}} W_\mu^3. \end{aligned} \tag{1.40}$$

The parameter θ_{W} is the mixing angle between the W_μ^3 and B_μ fields. It is called Weinberg angle and is an important parameter of the Standard Model. Its value has been measured with exceptional precision and is

$$\sin^2(\theta_W) = 0.23119(14) \text{ [7]}. \tag{1.41}$$

Comparing 1.37 and 1.40 leads to the mass relation

$$\frac{M_{\mathrm{W}}}{M_{\mathrm{Z}}} = \cos\theta_{\mathrm{W}} \tag{1.42}$$

which has been verified with extremely high precision.

Electroweak Interactions

The interactions of the electroweak bosons with the leptons are described in the resulting Lagrangian by the terms

$$\begin{aligned} \gamma \textit{ coupling} &: i g_e \bar{\psi}\gamma^\mu\psi, \; g_e = \sqrt{4\pi\alpha} \\ W \textit{ coupling} &: i\frac{g_w}{\sqrt{2}}\bar{\psi}\gamma^\mu\frac{1}{2}(1-\gamma^5)\psi = i\frac{g_w}{\sqrt{2}}\psi_L\gamma^\mu\psi_L, \; g_w = \frac{g_e}{\sin\theta_W} \\ Z \textit{ coupling} &: i g_z \bar{\psi}\gamma^\mu\frac{1}{2}(c_V - c_A\gamma^5)\psi, \; g_z = \frac{g_e}{\sin\theta_W\cos\theta_W} \end{aligned} \tag{1.43}$$

which correspond to the vertices in figure 1.3. The photon couples to leptons with a pure vector current term. In the couplings of Z and W bosons the axial and the vector currents are distinguishable. In the case of W, charged weak currents, the two components have the same strength, while in the case of Z, neutral weak currents, the factors c_V and c_A are present. Their values vary according to the nature of the fermions, namely

$$c_V = T_3 + 2Q\sin^2\theta_W \qquad c_A = T_3. \tag{1.44}$$

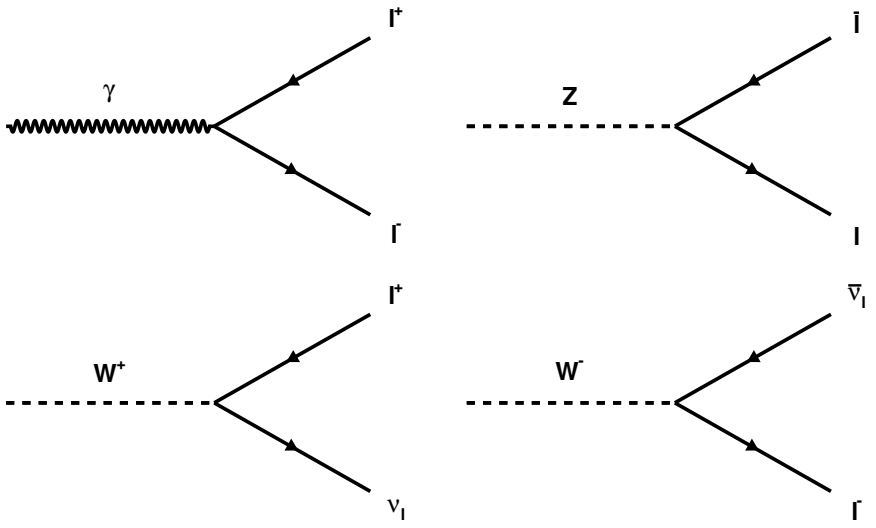

Figure 1.3: The Feynman diagrams of the couplings of the electroweak bosons to the leptons.

Moreover, it is important to observe that the couplings of the W boson can be read as exclusively involving the left handed components of the fermions. The description of the couplings of the W and Z to quarks is not as simple as with leptons only, but involves the formalism of the Cabibbo-Kobayashi-Maskawa (CKM) matrix [8].

Lepton Masses

To account for the masses of the leptons, a gauge invariant term has to be added to the Lagrangian

$$\mathcal{L}_M = -G_l \left[(\bar{\nu_l}, \bar{l})_L \begin{pmatrix} \phi_\alpha \\ \phi_\beta \end{pmatrix} l_R + \bar{l_R}(\phi^*_\alpha, \phi^*_\beta) \begin{pmatrix} \nu_l \\ l \end{pmatrix}_L \right]. \tag{1.45}$$

With some ordinary manipulations [9], it can be simplified to

$$\mathcal{L}_M = -m_l \bar{l} l - \frac{m_l}{v} \bar{l} l h, \tag{1.46}$$

which not only shows the familiar mass term for the lepton, but also its coupling to the Higgs boson, which is proportional to m_l.

1.6 Cross Section

Before proceeding to the characterisation of the strong interaction, a clarification of the widly employed concept of cross section is necessary. The cross section, indicated by the Greek letter σ, is a measure of the probability for a certain

process to happen in a scattering experiment. It is defined as the reaction rate R divided by the incoming flux Φ

$$\sigma = \frac{R}{\Phi}. \tag{1.47}$$

This quantity has the dimensions of a area and, in presence of two colliding particle beams, an appropriate measure unit is the barn (b), equivalent to 10^{-28} m^2.
From the theoretical point of view, the rate R can be expressed by *Fermi's golden rule* [10]:

$$R = 2\pi \cdot |\mathrm{M}_{if}|^2 \cdot \rho_f. \tag{1.48}$$

The ρ_f is a kinematical quantity and denotes the phase space available for the final state, while the dynamics involved in the scattering process is contained in the matrix element M_{if}, which connects the initial and final states

$$\mathrm{M}_{if} = < \psi_f | H_{int} | \psi_i >, \tag{1.49}$$

where the operator H_{int} is the interaction Hamiltonian. Fermi's Golden rule is a great achievement since it allows the complete factorisation of the dynamical and kinematical information in the description of a scattering process. A treatment to justify the validity of expression 1.49 in a relativistic context is given in [9].

1.7 Strong Interactions: Quantum Chromodynamics (QCD)

A treatment similar to the one shown for the electroweak sector in section 1.5 can also be formalised to describe the carriers of the strong force, the eight gluons, together with their couplings to the quarks. The strategy is to define three kinds of charges for quarks, called colour charges, traditionally referred to as red, blue and green, and then to make the strong component of the Standard Model Lagrangian invariant under transformations of the $SU(3)$ group. The resulting Lagrangian assumes the form:

$$\begin{aligned} \mathcal{L}_{QCD} &= \bar{\psi}(i\gamma_\mu \partial^\mu - m)\psi - \frac{1}{4} F^a_{\mu\nu} F^{\mu\nu}_a + g_s \bar{\psi} \gamma^\mu T_a \psi G^a_\mu \\ F^a_{\mu\nu} &= \partial_\mu G^a_\nu - \partial_\nu G^a_\mu - g_s f^{abc} G^b_\mu G^c_\nu \end{aligned} \tag{1.50}$$

where $G^a_{\mu\nu}$ represents the eight gluon fields, f^{abc} and T_a are the structure constant and the generators of the SU(3) group respectively (see appendix A for the details).

QCD and QED

The main difference between QCD and QED is that the gluons carry colour charge, one positive colour and one negative, while the photon is electrically neutral. This characteristic has an influence on the energy dependence of α_s, the strong coupling constant. The behaviour is indeed reversed with respect to the one of α_{em} (see section 1.3). Quark-antiquark clouds resulting from the splitting of gluons cause a screening of the quark colour-charge but unlike in QED, the charge contribution of the vector bosons, the gluons, has to be considered. It turns out that the gluon contribution is big enough to result in an anti-screening of the colour charge. It is therefore only at high energies that α_s becomes small, in what is called the regime of *asymptotic freedom*. The evolution of α_s is shown in figure 1.4.
Moreover, according to this construction, the more two quarks are driven apart, the bigger is the amount of energy contained in the fields connecting them, exactly the opposite of the electromagnetic case. At some point the energy becomes high enough to create quark-antiquark pairs that recombine with the original quarks. Quarks and gluons are therefore constrained inside hadrons, i.e. composite states formed by quarks, anti-quarks and gluons. This phenomenon is referred to as *confinement*. After a highly energetic collision, such a process can also occur on a large scale, when a single quark or gluon can give rise to a chain of creation of quark-antiquark pairs that can be detected as a spray of particles, a *jet*.

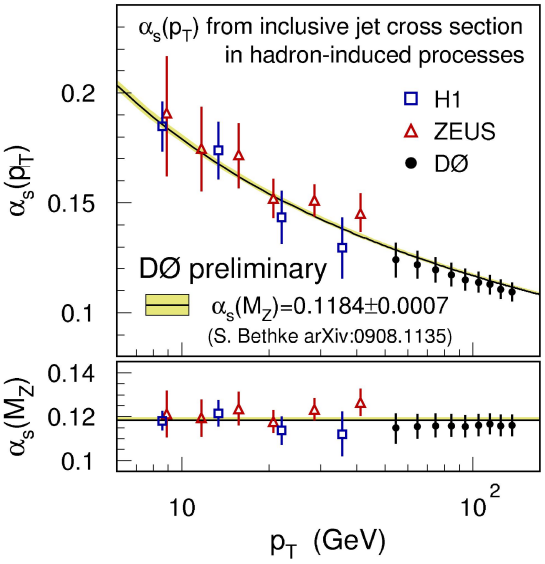

Figure 1.4: The α_s coupling constant measured as a function of the highest jet momentum in an event. The value decreases as the energy regime in the events increases. Taken from [11].

Even though direct observation of quarks, anti-quarks and gluons, generally called partons, outside hadrons could not be achieved, numerous experimental

proofs of their existence have been provided starting from the late sixties [12, 13, 14].

Parton Distribution Functions

The compositeness of the proton has important consequences on the physics at proton colliders. The cross section of a process involving the scattering of highly energetic protons must be expressed as a function of the partonic interactions:

$$\sigma_{pp}(P_1, P_2) = \sum_{i,j \in partons} \iint dx_i dx_j \; f_i(x_i, \mu^2) f_j(x_j, \mu^2) \sigma_{ij}^{partonic}(Q^2, \mu^2, x_i, x_j) \tag{1.51}$$

where P_1 and P_2 are the momenta of the colliding protons, x_i and x_j represent the fractions of the proton momentum carried by the interacting partons, μ is the factorisation scale and Q^2 is the characteristic scale of the hard scattering. The functions f_i and f_j are the Parton Distribution Functions (PDFs) and represent the probability for the parton to carry a fractional proton momentum x. Therefore, these parton density functions are important ingredients for any kind of cross section prediction for proton-proton interactions. For this reason, many experimental determinations of their values were carried out. The combination of all the available inputs coming from the experiments is done by independent groups like CTEQ [15] and MRST [16]. An example of a parton distribution function, together with some of the most relevant cross sections at the LHC, is shown in figure 1.5

1.7.1 The Underlying Event

As mentioned above, the protons are composite objects made of partons: quarks, anti-quarks and gluons. All the particles that share this compositness are called *hadrons*. Protons are made of three *valence quarks* (two up quarks and one down quark), and gluons that hold them together. The gluons are continuously splitting themselves in virtual gluon pairs and quark-antiquark pairs, called *sea quarks*.
Such a composite structure makes the collisions between protons much more complex than the simple electron-positron annihilation. Beyond the hard scattering that can take place between two highly energetic partons, other processes can occur at the same time, giving rise to the underlying event (UE). Since the underlying event is an unavoidable background for most collider observables, it is very important to understand its features in order to be able to study interesting physics phenomenona.

UE Components

Several descriptions of the underlying event are possible. In this work, the underlying event will be abstractly divided into different parts, which unfortunately cannot clearly be distinguished on an event-by-event basis (see figure 1.6).
The first component is represented by the additional semi-hard parton interactions

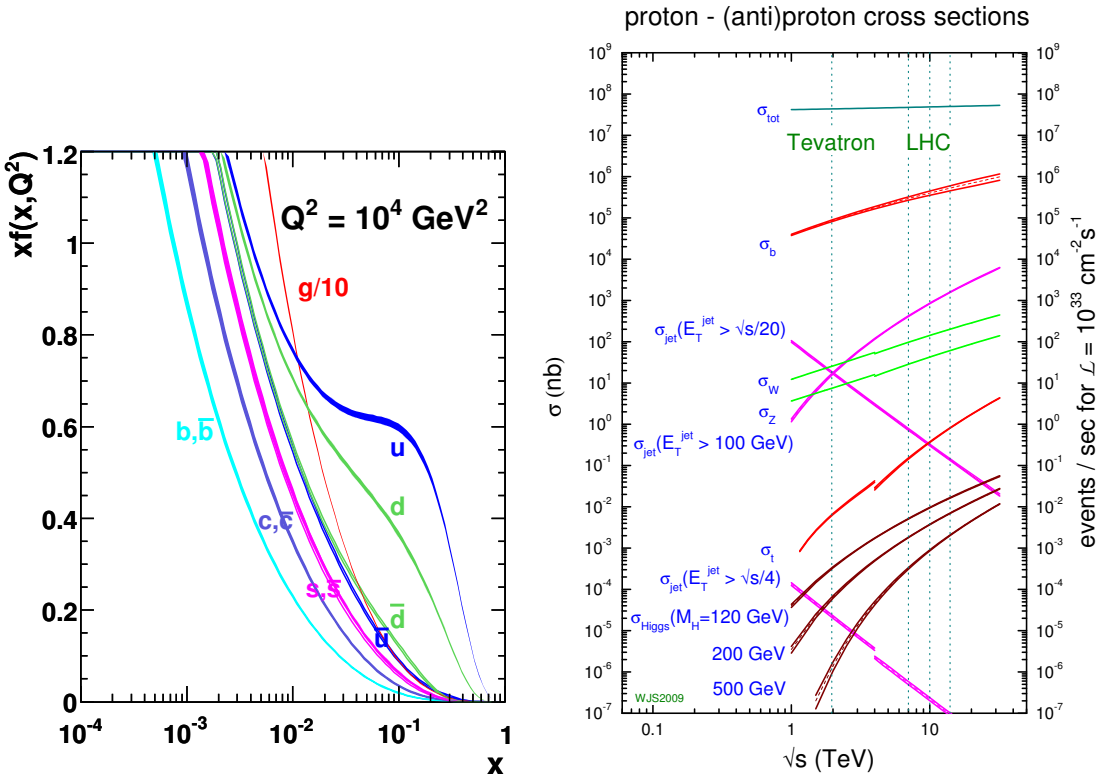

Figure 1.5: On the left, an example of PDF is shown. On the right, a scheme of the most important production cross sections involving proton-(anti)proton collision are shown. Both taken from [16]. A description of the LHC accelerator will be given in chapter 2.

that take place during a proton-proton collision, which are called multiple parton interactions (MPI).

Moreover, the portions of the protons which are not expelled by the main parton-parton collision, the *protons remnants*, are unstable. Since they carry a colour charge, the strong force make them interact and recombine with other colour-charged products of the collision in order to produce colourless states.

In addition to that, in every process that contains electrically or colour charged objects in the initial or final state, photon or gluon radiation in the form of initial state radiation (ISR) and final state radiation (FSR) may cause large contributions to the overall topology of events. Starting from a basic $2 \to 2$ process, this kind of corrections will generate $2 \to 3$, $2 \to 4$, and so on, final-state topologies. As the available energies are increased, hard emission of this kind becomes increasingly important.

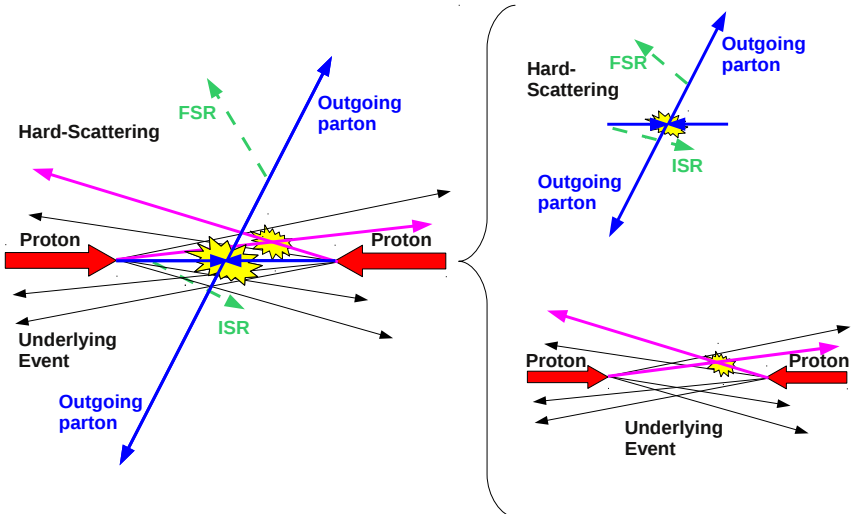

Figure 1.6: The hard-scattering component of the event consists of particles that result from the hadronisation of the two outgoing partons (i.e. the initial two jets, upper right) plus the particles that arise from initial and final state radiation (i.e. multi-jets). The underlying event consists of particles that arise from the beam-beam remnants and from multiple parton interactions which, however, cannot be differentiated unambiguously from other partons in the event (bottom right).

1.8 Z and Higgs Boson Production and Decays at Hadron Colliders

1.8.1 Z Boson Production

The first Z boson directly produced was observed at CERN, with a hadron collider, the Super Proton Antiproton Synchrotron (Sp$\bar{\text{p}}$S) in 1983 [2]. Since then, the properties of the boson were investigated by different experiments, above all the four experiments of the Large Electron Positron Collider (LEP). The machine was a veritable Z factory that lead to the detection of approximately 17 millions Z bosons [17]. Presently, the best estimates of the mass and width of the boson are extremely precise [7]:

$$M_Z = 91.1876(21)\ \text{GeV} \qquad \Gamma_Z = 2.4952(23)\ \text{GeV} \tag{1.52}$$

Despite the relatively large invariant mass, a huge number of Z bosons will be created also at the LHC. The most recent calculations for the process $pp \to Z \to l^+l^-$, predict a cross section of 0.97(4) nb at 7 TeV centre of mass energy [18].

Such an amount of candidates can be exploited to carry out precision measurements of Standard Model parameters and for essential calibration of detector components. For example, the absolute energy scale calibration of the CMS electromagnetic calorimeter, can be carried out with the electrons deriving from the Z boson decay [19]. On the other hand, events which contain a Z boson accompanied by an additional jet, open the field for a variety of relevant measurements. Processes like

$$\begin{aligned} qg &\rightarrow Z + q \\ \bar{q}g &\rightarrow Z + \bar{q} \\ q\bar{q} &\rightarrow Z + g \end{aligned} \tag{1.53}$$

present a particular topology, in which the transverse momentum of the jet which originates from quarks and gluons is balanced by the momentum of the Z boson. These events are excellent candidates for jet energy calibration procedures since the momentum of the Z decaying in two muons can be reconstructed with high accuracy and therefore be used to calibrate the jet originating by the balancing parton. Figure 1.7 shows the first event detected by CMS where such a topology is present. The Feynman diagrams of the production mechanisms that lead to a

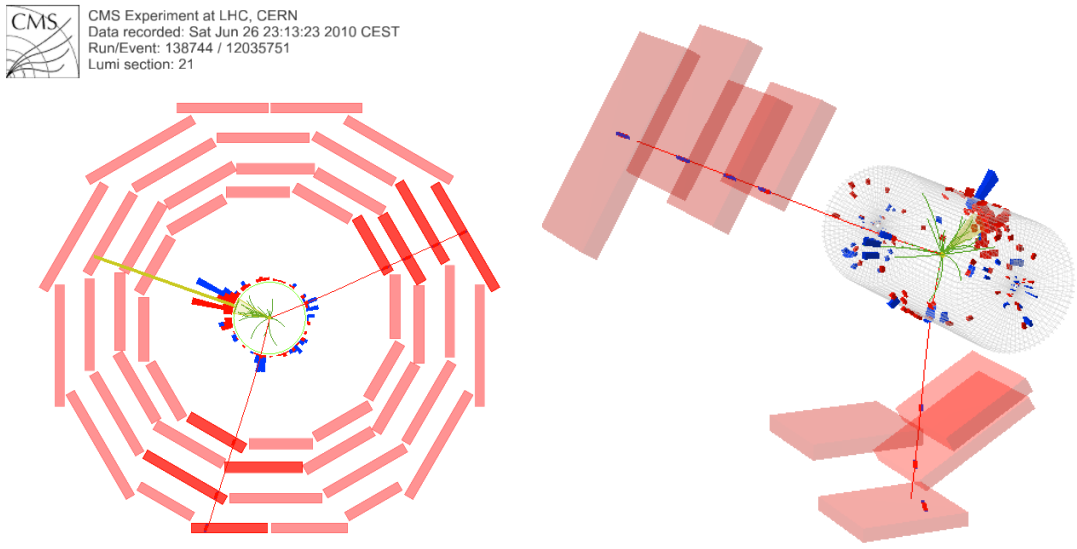

Figure 1.7: A bi-dimensional (left) and tri-dimensional (right) view of the first $Z(\rightarrow \mu\mu)$ + jet event detected by CMS. The transverse momentum of the jet is 46 GeV and the one of the Z boson 49.82 GeV.

Z boson and a jet balanced in transverse momentum are shown in figure 1.8.

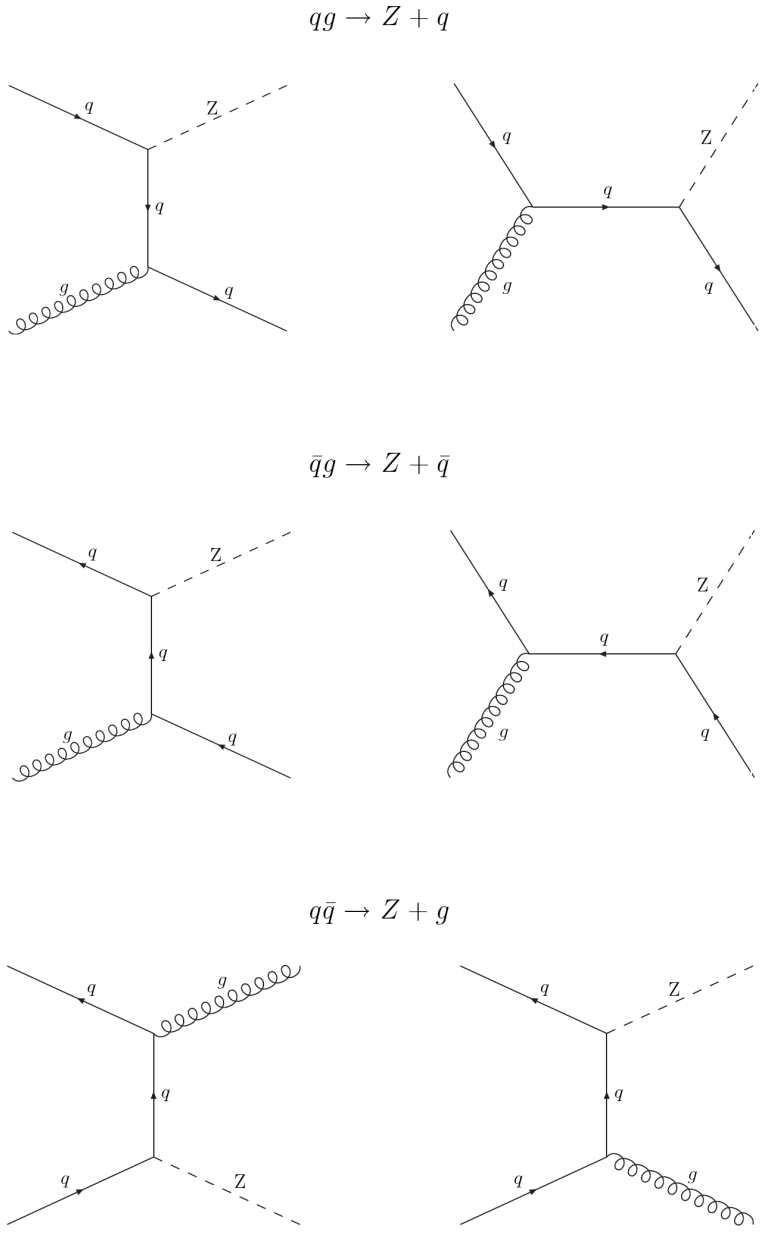

Figure 1.8: The Z+jet Feynman diagrams for different production processes, in the t and s channels.

1.8.2 Higgs boson Production

Given that the coupling of the Higgs boson to particles is proportional to their masses (equation 1.46), its production mechanisms preferentially involve heavy objects like the top quark.
For what concerns proton-proton collisions, the main production processes are (figure 1.9):

a) Gluon-gluon fusion: $gg \rightarrow H$

b) Vector boson fusion (VBF): $qq \rightarrow VV \rightarrow qqH$

c) Associated production with W and Z: $q\bar{q} \rightarrow VH$

d) Associated production with heavy quarks: gg,$q\bar{q} \rightarrow Q\bar{Q}H$

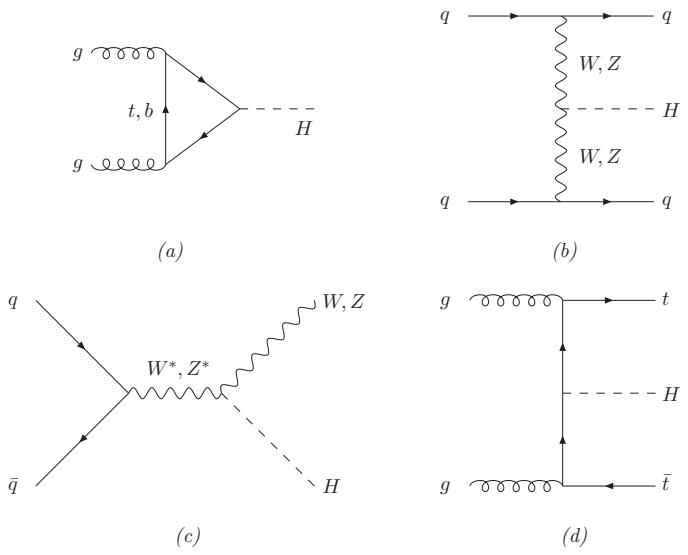

Figure 1.9: Typical leading order Feynman diagrams for Higgs boson production mechanisms relevant at the LHC: Gluon fusion (a), weak boson fusion (b), Higgsstrahlung (c), associated production (d) [20].

Gluon Fusion

The dominant process at the LHC is the gluon fusion. It is characterised by a quark triangle loop in which mainly a virtual top is exchanged among the vertices. This channel is affected by a huge scale dependence, which requires very precise theoretical calculations to predict its cross-section.

Vector Boson Fusion

The second largest contribution to the cross section stems from the fusion of two weak bosons which are radiated from the initial-state quarks. The cross section of this process is reduced by roughly a factor of ten as compared to the gluon initiated process. However, it represents an interesting alternative since it features a very distinct signature.
In vector boson fusion events, the initial quarks which emit the bosons are scattered to high pseudorapidities and can be detected as two energetic jets in the forward region of the detector (the *tagging jets*), thereby yielding a very pronounced signature quite different from generic QCD induced high-p_T signals. Furthermore, a feature of vector boson fusion events is the lack of colour exchange between the initial quarks, leading to reduced jet activity in the central detector region.
A precise knowledge of the jet energy in the different regions of the detector is crucial in order to take advantage of the tagging jets and veto of jet activity in the central region.

Higgs-strahlung

In the associated production with W and Z bosons, also called Higgsstrahlung, the Higgs boson is radiated by an off-shell weak boson resulting from the annihilation of a quark and an antiquark. The contribution of this process to the total cross section resulting from proton-proton collisions is minor.

Associated Production

Associated production with heavy quarks, especially with top quark pairs plays a significant role at the LHC for a light Higgs boson. Since the top quark decays to bottom quarks with an extremely high probability, a good handle on this type of events is given by the presence of jets originated by b quarks. These objects can be recognised with high efficiency through the application of b-tagging procedures [21].

Higgs Boson Decays and Mass Constraints

A very important aspect regarding the Standard Model Higgs Boson, besides the production mechanisms, are the decay modes. The probabilities of Higgs boson decays in a particular channel, the branching ratios, as a function of its mass, are displayed in figure 1.10. The present constraints on the Higgs boson mass indicate a value lying in the low mass region (figure 1.11).

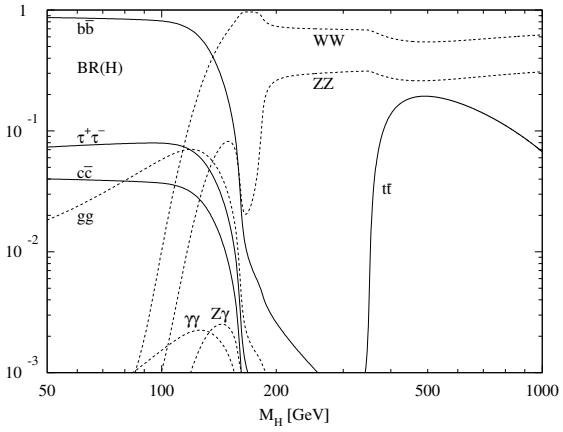

Figure 1.10: The branching ratios of the Standard Model Higgs boson as function of its assumed mass. Taken from [20].

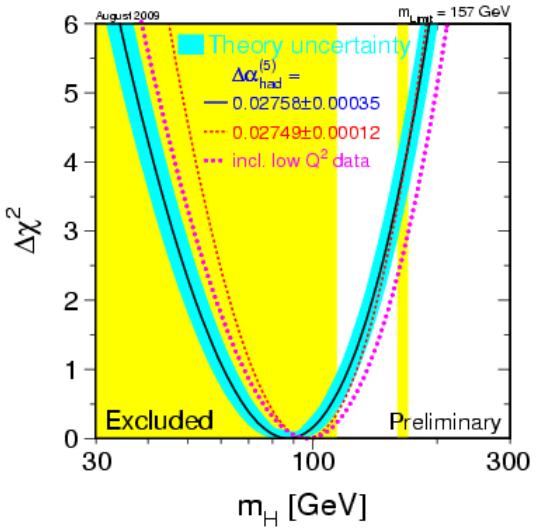

Figure 1.11: The best fit value of the Standard Model Higgs mass with the combined results of the four LEP experiments, together with the excluded regions featured by the LEP and Tevatron experiments. The "light" Standard Model Higgs boson seems to be favoured. The curve is obtained with the profile likelihood method and the exclusion bands with the conservative frequentist approach (see chapter 4). Taken from [22].

Chapter 2

The CMS Experiment at the Large Hadron Collider

The European Council of Nuclear Research (CERN) in Geneva is one of the world's largest and most respected centres for scientific research. CERN is born from the need in the late 1940s to relaunch fundamental research and promote cooperation and peace among states in Europe after the War. Back then, joining the efforts of the single states was also the only way to put together enough resources to compete with the laboratories that had been founded in the United States [23].

The LHC [24] is the most powerful particle collider built up to now and is operational at CERN, hosted in the old Large Electron-Positron Collider (LEP) [25] tunnel. Both protons and heavy ions can be accelerated, paving the way for a new era of exciting discoveries.
In the initial phase, the LHC collides protons at a centre of mass energy of $\sqrt{s}$ =7 TeV, which is the highest presently reached. The centre of mass energy will then rise in the coming years to the design value of $\sqrt{s}$ =14 TeV.
An overview of the machine and of the requirements for the LHC experiments is given in section 2.1. The general description of the CMS detector is then presented in section 2.2.

2.1 The Large Hadron Collider

The LHC proton-proton synchrotron is part of the CERN accelerator complex (figure 2.1). Old yet reliable machines at CERN, like the SPS and PS (Proton Synchrotron), are used to pre-accelerate protons to 400 GeV before their injection into the LHC. Two beams circulate simultaneously in the machine in opposite directions and they are crossed, via a bunch crossing procedure, at four interaction points where the main experiments are located. The accelerator is 27 km long and in total composed of more than one thousand magnets, all employing

superconductive wirings. It comprises superconductive radio frequency cavities to accelerate the protons and ions and dipole magnets able to generate a 8.3 T magnetic field to bend them. Moreover, other magnet types are employed for the machine optics, participating in the focussing and squeezing of the beams. The superconductivity regime is reached in the magnets at an operational temperature of about -271 ℃, which makes the LHC one of the coldest object in the universe. Such extreme conditions are reached with an advanced cryogenic system which exploits the properties of helium [24].

Colliding Protons

The LHC is a discovery machine, conceived to give access to a large range of physics opportunities, from the precise measurement of the properties of known objects to the exploration of high energy frontiers. To enhance the discovery potential of the LHC, protons were chosen. Given their composite nature, their total momentum is distributed according to the parton distribution functions among partons, which in the energy regime of the LHC are the scatterers taking part in the collisions. The centre of mass energy of the fundamental scatterers is therefore not known a priori, like for an electron-positron collider, allowing to explore every possible region of the phase space without varying the energy of the beams. At LEP the main limiting factor for the centre of mass energy was the synchrotron radiation [26] emitted by the accelerated electrons [27]. Protons, which have a much larger mass than electrons, do give rise to a much smaller significant synchrotron radiation, which allows to reach higher energies.

Luminosity

Beyond the energy of the particles circulating in the machine, another relevant parameter to consider is the luminosity, i.e. the factor of proportionality between the event rate and the interaction cross section. Hence, to accumulate the maximum number of events in a given amount of running time, a high luminosity is of crucial importance (figure 2.2). The design luminosity of the LHC is unprecedented for a proton machine: 10^{34} cm^{-1} s^{-1}. This quantity can be calculated as a first approximation by the formula

$$\mathcal{L} = \frac{N^2 k f \gamma}{4\pi\epsilon_n \beta^\star} F \tag{2.1}$$

where N is the number of particles in each of the k circulating bunches, the "packages" of protons into which the beam is divided, f the revolution frequency, $\beta^\star$ the value of the betatron function at the crossing point and ϵ_n the emittance corresponding to one σ contour of the beam, contracted by a Lorentz factor γ. F is a reduction factor due to the crossing angle between the beams. Thus, to achieve high luminosity, the LHC beam is made of a high number of bunches, filled with $\approx 10^{10}$ protons, which collide at an extremely high frequency (the nominal value is 40 MHz) with well focussed beams (small emittance and $\beta^\star$). The main machine parameters (design values) are listed in table 2.1.

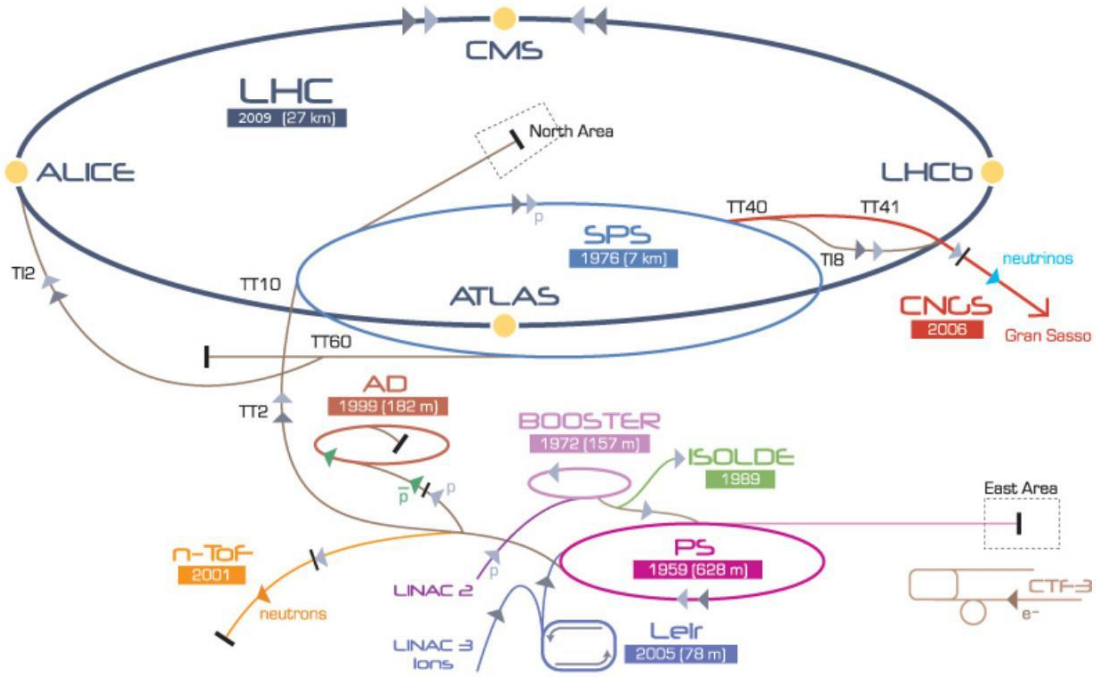

Figure 2.1: The CERN accelerator complex. The names of the machines are accompanied by the starting year of their operation. Several machines are used to pre-accelerate the protons before the injection into the LHC. Taken from [28].

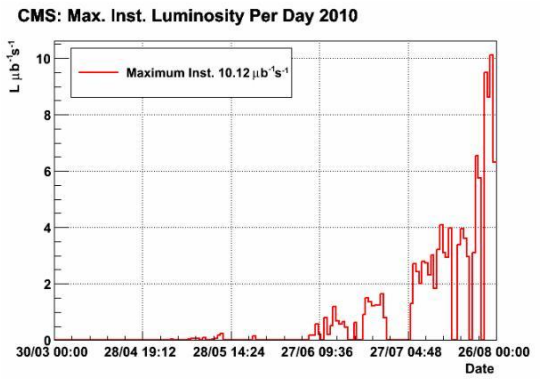

Figure 2.2: The plot shows the maximum luminosity reached by the LHC up to the end of August. Taken from [29].

Table 2.1: Some of the nominal machine parameters relevant for the LHC detectors.

Parameter	pp	$Pb-Pb$	Dimensions
Energy per nucleon	7	2.76	TeV
Dipole field at 7 TeV	8.33	8.33	T
Design Luminosity	10^{34}	10^{27}	$\mathrm{cm}^{-2}\ \mathrm{s}^{-1}$
Bunch separation	25	100	ns
No. of bunches	2808	592	–
No. particles per bunch	1.15×10^{11}	7.0×10^{7}	–
β-value at IP	0.55	0.5	m
RMS beam radius at IP	16.7	15.9	μm
Luminosity lifetime	15	6	h
Number of collisions/crossing	≈ 20	–	–

Detectors at LHC

The detectors at the LHC must face a wide range of unprecedented challenges. Firstly, the interaction rate is extremely high and only a very small part of the events contain interesting processes. This makes a fast and efficient trigger and data acquisition system necessary. Moreover, a fast response (of the order of 25-50 ns) is needed to resolve the signals coming from two subsequent events and a fine granularity is necessary to separate the large number of particles that are originated in the collisions.
The radiation environment that the detectors must withstand for the whole running period of the machine, due to the high flux of particles, is extreme. The components of the detectors which are directly next to the beam will receive a radiation dose of 10 kGy per year. Radiation resistance for all the detector components is therefore required for the good operation of the devices.

The main four detectors installed at the LHC, namely CMS (Compact Muon Solenoid) [30], ATLAS (A Toroidal LHC ApparatuS) [31], LHCb [32] and ALICE (A Large Ion Collider Experiment) [33].

CMS and ATLAS are two general purpose experiments, featured by complementary characteristics and detector choices. On the other hand, the LHCb collaboration aims above all to perform precision measurements in the sector of B mesons to reveal possible indications for new physics. ALICE is dedicated to heavy ions physics and the goal of the experiment is the investigation of the behaviour of a type of strongly interacting hadronic matter, called quark-gluon plasma, resulting from high energy lead and calcium nuclei collisions.

In the following sections a brief description of the CMS detector is given.

2.2 The CMS Detector

CMS is a general purpose detector that is installed at the interaction point number five along the LHC tunnel. The detector has a cylindrical shape, symmetric around the beam and is divided in two endcaps and a barrel. The overall dimensions of CMS are a length of 21.6 m, a diameter of 14.6 m and a total weight of 12,500 t. It is characterised by a layered structure: Starting from the beam pipe its subdetectors are a silicon tracking device, an electromagnetic calorimeter, a hadronic calorimeter and an advanced muon detection system.
In every particle detector, the magnetic field plays a fundamental role, since it is necessary for the momentum measurement of charged particles. An important aspect driving the detector design and layout is the choice of the magnetic field configuration. At the heart of CMS sits a 13-m-long, 5.9 m inner diameter, 3.8 T superconducting solenoid. Such a high magnetic field was chosen in order to achieve good momentum resolution within a compact spectrometer. The core of the magnet coil is large enough to accommodate the inner tracker and the calorimetry subsystems, with the exception of its very-forward part. An iron return yoke routes back the magnetic field generated by the solenoid, avoiding its spread into the cavern. The return field is so intense (1.5 T) to saturate three layers of iron, in total 1.5 m thick. Each of these layers is installed between two layers of muon detectors. The redundancy ensured by the muon measurements therewith obtained, ensures robustness as well as full geometric coverage.

The overall layout of CMS is shown in figure 2.3 and a slice of it can be inspected in figure 2.4.

The Coordinate Systems

The coordinate frame used to describe the detector is a right handed Cartesian system with the x axis pointing toward the centre of the LHC ring, the z axis directed along the beam axis and the y axis directed upward. Given the cylindrical symmetry of CMS, a convenient coordinate system is given by the triplet (r, ϕ, η), being r the distance from the z axis, ϕ the azimuthal coordinate with respect to the x axis and η the pseudorapidity, which is defined as $\eta = -ln(tan(\theta/2))$, where θ is the polar angle with respect to the z axis.

2.2.1 The Inner Tracking System

The CMS inner tracking [34] device is entirely constructed using silicon technology [35] and is the sub-detector closest to the beam line. Its role is to provide the information to precisely reconstruct the vertices of interactions and the tracks of the charged particles. It is divided in a silicon pixel and silicon strip tracker. The pixel detector is divided into three barrel layers and two endcap discs on each side of them. It comprises 66 million pixels. The strip detector is made of 9.6 million silicon strips and is divided into four sub-components, the Tracker Inner Barrel

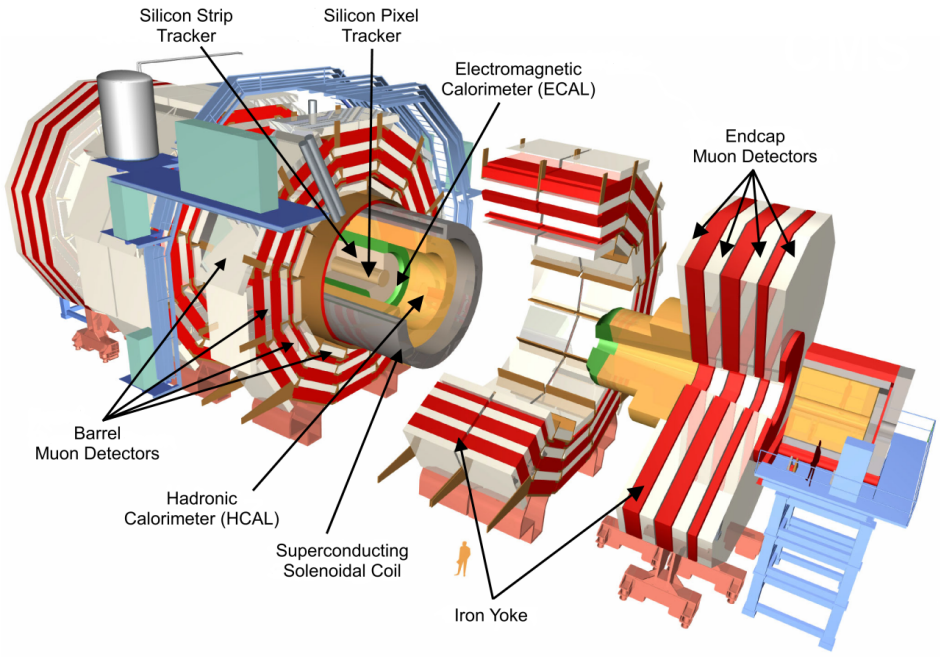

Figure 2.3: An overall view of CMS. The detector is symmetrical around the beam line and designed to be hermetic.

(TIB) made of four layers, the Tracker Outer Barrel (TOB) made of six layers, the Tracker End Cap (TEC) comprising nine disks per side, and the Tracker Inner Disks (TID), made of three small disks filling the gaps between the TIB and the TEC (see figure 2.5). The density of charged particles and the different levels of radiation in the different regions drove the choice of the different types of silicon devices installed in each subsystem:

- The region closest to the interaction vertex ($r \approx 10\,\text{cm}$), where the particle flux is the highest, houses pixel detectors. Each pixel has a size of $\approx 100 \times 150\,\mu\text{m}^2$.
- In the intermediate region ($20 < \text{r} < 55\,\text{cm}$), the particle flux is low enough to enable the use of silicon microstrip detectors with a minimum cell size of $10\,\text{cm} \times 80\,\mu\text{m}$.
- The outermost region ($\text{r} > 55\,\text{cm}$), where the particle flux has dropped sufficiently, larger-pitch silicon microstrips with a maximum cell size of $25\,\text{cm} \times 180\ \mu\text{m}$ are used.

To operate it without degrading its performance in the hard radiation environment of the LHC, the temperature of the whole device must be kept very low, ideally slightly below -10 ℃.

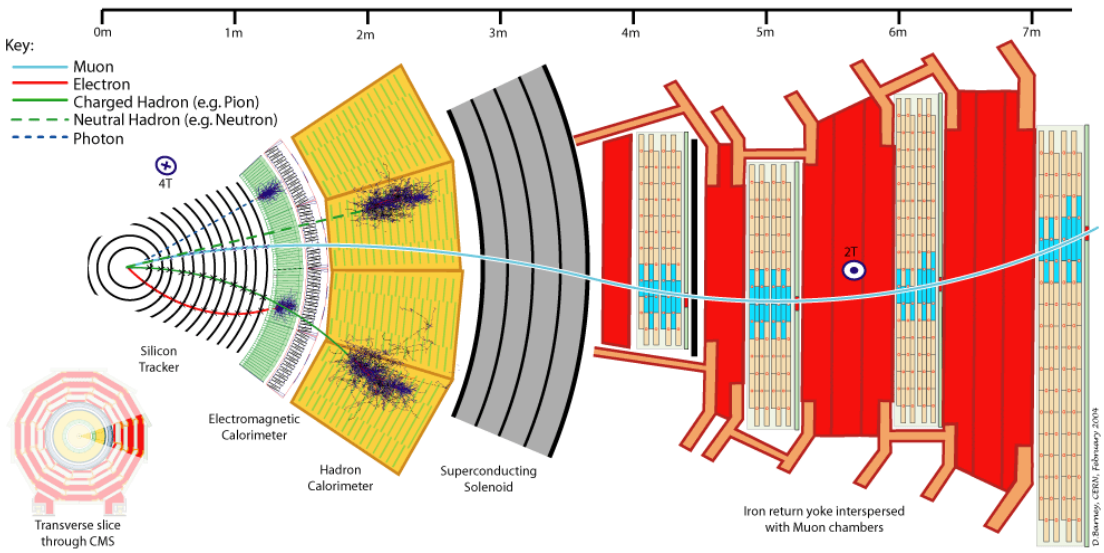

Figure 2.4: Slice of the CMS detector with the tracks of an electron, a photon, a hadron (e.g. a pion) and a muon. The electron and the photon deposit their whole energy in the electromagnetic calorimeter generating an electromagnetic shower. Despite a low energy deposit in this region, the hadron reaches the hadron calorimeter where it is stopped. Only muons are able to escape the whole detector and are detected by the tracker and the muon system.

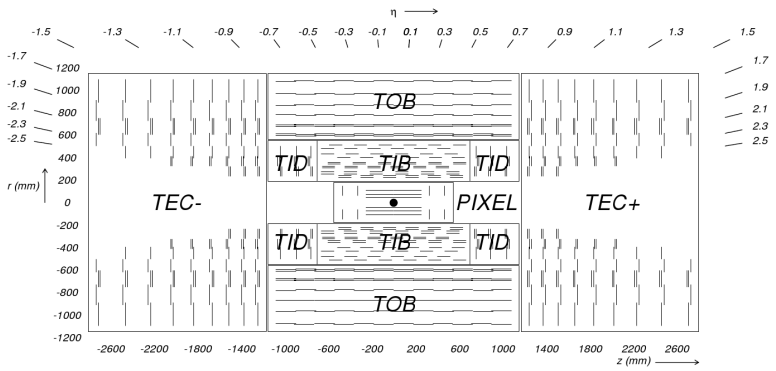

Figure 2.5: The structure of the silicon tracker, showing the different components.

Being 2.7 m long and with a radius of nearly 1.1 m, the overall volume of the CMS Tracker is bigger than 10 m^3 and its layers cover in total a surface of 196 m^2. Such unprecedented dimensions make it the biggest silicon device ever built up to now. The outer layers of the CMS tracker before its installation is shown in figure 2.6.

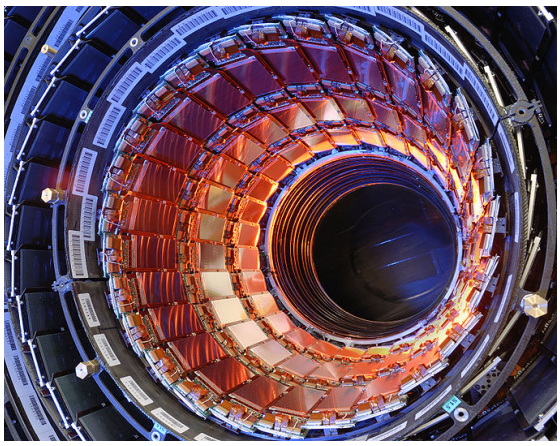

Figure 2.6: The CMS barrel strip tracker before its installation in the detector.

Tracker Budget Material

A rather complex cooling system, a huge number of electronic boards, readout and power supply cables are coupled to the silicon detectors that form the tracker. All these inactive components are referred to as budget material. One challenge that the tracker design had to comply with, are severe budget material constraints which have to be met in order to not degrade the excellent energy resolution of the electromagnetic calorimeter (see figure 2.7).

Tracker Operation

The operation of the CMS tracker, as well as all other subdetectors, requires constant supervision of operators. In particular, the strip tracker requires one shifter that with the aid of a very advanced monitoring system [37], checks the good functioning of the device in terms of temperature, power supply and quality of the acquired data. In the scope of the work described in this thesis, a large number of tracker-shifts were performed in the CMS control room discussed at the LHC interaction point 5 in Cessy (France).

2.2.2 The Electromagnetic Calorimeter

The electromagnetic calorimeter ECAL [38] is a hermetic, homogeneous calorimeter composed of 61,200 lead tungstate ($PbWO_4$) scintillating crystals mounted in the central barrel part, closed by two endcaps including 7324 crystals each. The endcaps cover the pseudorapidity range $1.479 < |\eta| < 3$ and consist of identically

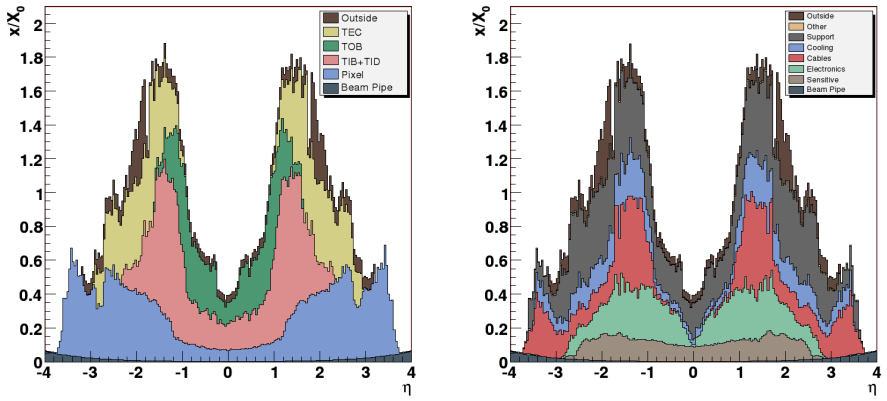

Figure 2.7: The CMS tracker material budget can be several interaction lengths thick [36].

shaped crystals, grouped into carbon fibre structures of 5×5 elements, called supercrystals. The barrel granularity is 360-fold in ϕ and (2×85)-fold in η. The ECAL allows to measure the energy of electrons and photons with high precision and with a resolution $\Delta E/E$ that reaches 0.5%. A view of the calorimeter during its assembly is given in figure 2.8.
The high density of lead tungstate (8.3 g/cm^3), short radiation length (0.89 cm) and a small Molière radius (2.2 cm) result in a fine granularity and allow the dimensions of the detector to be small. Moreover, the scintillation decay time is of the same order of magnitude as the LHC bunch design crossing time: About 80% of the light is emitted in 25 ns. The energy calibration of ECAL is a challenging task, that began during the test beam campaign of 2006 [39, 40]. Presently, an *in situ* calibration process is ongoing to equalise the responses of each crystal and to align them to a known reference, exploiting well understood physics channels, like for example Z boson decaying in two electrons.

2.2.3 The Hadron Calorimeter

The hadron calorimeter (HCAL) [41] is the CMS component endowed to the measurement of the energy of hadrons, both charged and neutral (see figure 2.9). Like the ECAL, the HCAL is divided into a barrel and two endcaps and its design is strongly influenced by the choice of the magnet. The device is almost entirely located inside the coil. Two exceptions are a layer of scintillators, referred to as the "hadron outer" detector, located in the barrel outside the coil and the "hadron forward" calorimeter, placed outside the magnet yoke, 11 m away along the beam direction from the nominal interaction point.
Like many other hadron calorimeters, the HCAL is composed of an absorber

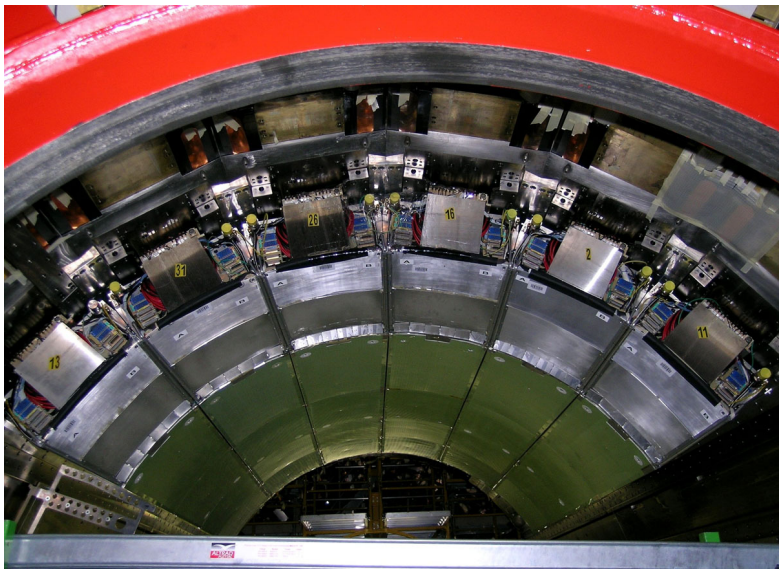

Figure 2.8: A view of the electromagnetic calorimeter barrel slices, the supermodules, during the assembly of the detector.

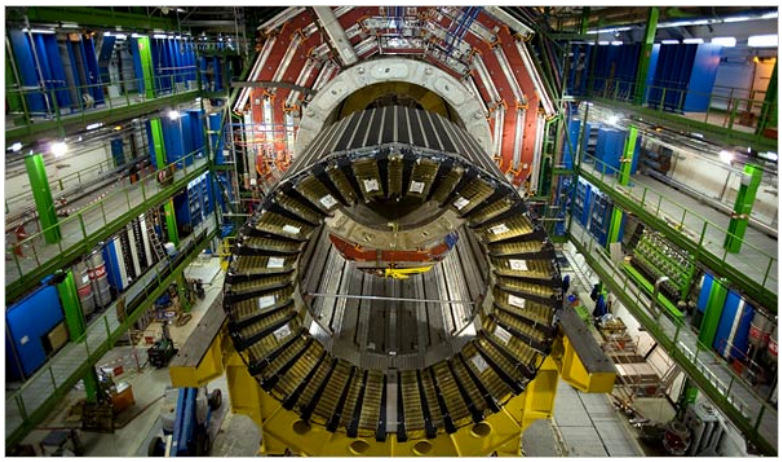

Figure 2.9: The barrel of the HCAL calorimeter just before its insertion in the CMS detector. The whole subdetector fits inside the superconducting solenoid.

medium and an active medium. Brass has been chosen as absorber material as it has a reasonably short interaction length, is easy to machine and is non-magnetic. In addition, both copper and zinc have a fairly low nuclear number Z, a positive factor since muon scattering is prominent in materials with high Z values.
This subdetector has to completely absorb the particle showers originating from the interactions of the hadrons in the hadronic calorimeter. Thus, the amount of interaction lengths in front of the magnet had to be maximised keeping the amount of space devoted to the active medium to a minimum. The tile/fibre technology represents an ideal choice, being implemented as 3.7 mm thick active layers of plastic scintillators read out with embedded wavelength-shifting fibres.

2.2.4 The Muon System

The design of CMS is characterised by the emphasis on the precise measurement of muons properties. A huge muon detection system [42], composed of three types of gaseous detectors, is integrated in the iron return yoke of the magnet.
The detector technologies have been chosen considering the large surface to be covered and the different radiation environments. In the barrel region, where the neutron induced background is small and the muon rate is low, drift tube chambers (DT) are used. In the two endcaps, where the muon rate as well as the neutron induced background rate is high, cathode strip chambers (CSC) are deployed and cover the region up to $|\eta| < 2.4$. In addition to these, resistive plate chambers are used in both the barrel and the endcap regions for trigger purposes and time measurements, for example cosmic muons rejection. The usage of these detectors allows also to improve the geometrical coverage. The barrel muon chambers during the installation are shown in figure 2.10.

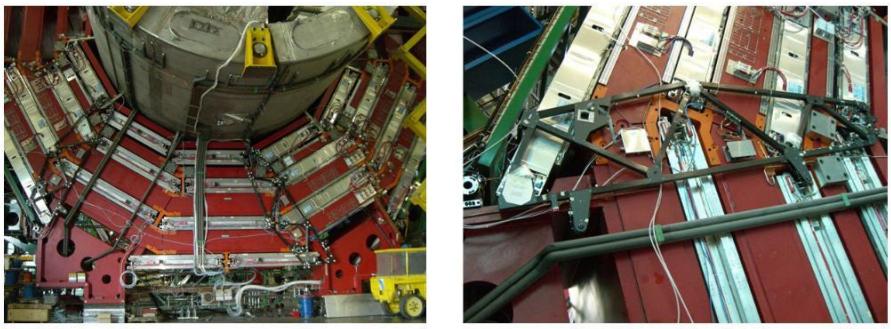

Figure 2.10: The outer part of one CMS slice of the barrel. The muon chambers were being assembled between the iron wheels.

Redundant measurements of the properties of the muons produced at the interaction point are effectuated by CMS, in a first step by the tracker and then by the muon system. The combination of these complementary measurements together

with the very intense but simply structured magnetic field allows to achieve an unprecedented precision in the measurement of muon kinematic quantities (see section 3.5.2).

2.2.5 The Trigger and the Data Acquisition

At the nominal luminosity, when collisions take place at a frequency of 40 MHz, the data acquisition system will have to cope with $\approx 10^9$ events per second. Since every event as a size of ≈ 1.5 MB, the total flux of data would consist of 60 TB/s. Clearly this amount of information is impossible to process and above all no mass storage would be able to record it. A two level trigger system (figure 2.11) is used by CMS to achieve the adequate output data flow. This data reduction has been carefully designed since it is an inherent selection procedure for every physics analysis and must not contain any bias.

L1 and HLT

The Level-1 trigger [43] reduces the event rate from 40 MHz to 100 kHz. It is implemented on custom hardware in order to decide very rapidly ($3\,\mu$s) to reject or keep an event. The decision is taken for example upon the presence of energy deposits in the calorimeters, silicon detectors and muon systems. To detect such signatures in an efficient way a reduced detector granularity and resolution are employed.
Once accepted by the Level-1 trigger the events are filtered through the High Level Trigger (HLT) system [44]. The HLT consists of a farm of about a thousand commercial processors and has access to the information coming from the whole detector. The time spent for the analysis of one single event is of the order of one second. Each processor runs the same HLT software code to reduce the Level-1 output rate of 100 kHz to 150 Hz.

Thus CMS presents challenges not only in terms of the physics programme, detector operation and maintenance, but also in terms of the acquired data volume and the computing infrastructure to process it. Datasets and resource requirements are at least an order of magnitude larger than in previous experiments. More details about how to overcome these challenges are located in chapter 3.

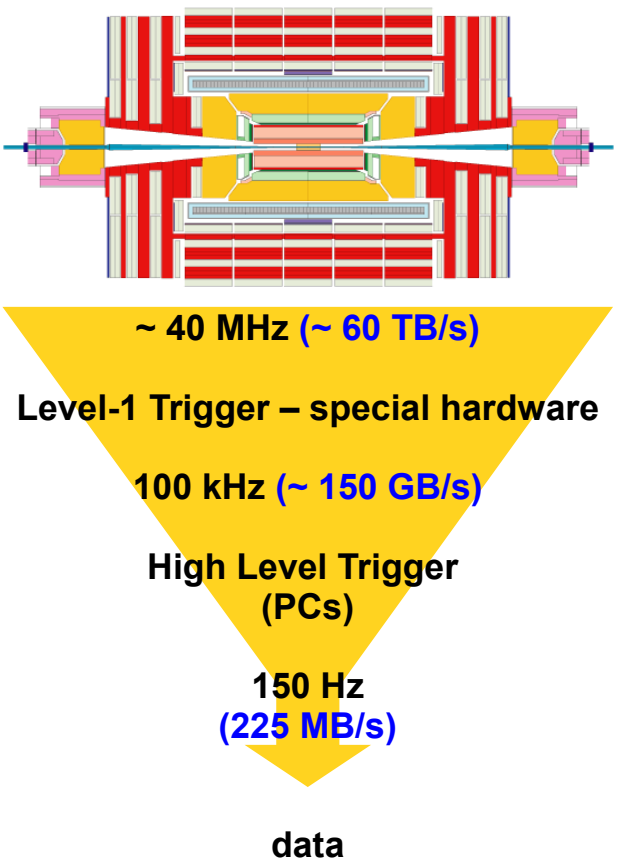

Figure 2.11: Overview of the CMS trigger system. A huge reduction is achieved via the deployment of a two level system.

Chapter 3

Software Framework and Computing Infrastructure

Each event recorded by the CMS detector consists of the raw electronic response of its components. Before any analysis could start, these signals have to be treated and converted in order to analyse the underlying physics processes that took place inside the detector. This operation is called reconstruction. Moreover, a huge amount of Monte Carlo pseudo-data has to be delivered to compare the distributions of the measured observables to their expected values in order to both monitor the good functioning of the detector and underline the presence of signals proving new physics phenomena. The CMSSW application framework has been developed to achieve these objectives [45].
In the following the description of the software components and the computing infrastructure used to carry out the studies described in this thesis are presented. At the beginning the ROOT [46] data analysis framework is outlined, with a particular emphasis on the RooFit [47] and the RooStats [48] toolkits. At the end the CMSSW framework is characterised through the description of the procedures of Monte Carlo generation, detector simulation and event reconstruction steps. Finally, a short summary of the basic concepts of the Worldwide LHC Computing Grid and the CMS computing model are given.
In the context of the work described in this thesis, beside the complete design and development of RooStatsCms (see appendix B) and other major contributions to RooStats, many improvements to RooFit were conceived (see sections 3.1.1 and 3.1.2). These contributions spanned from new functionalities to optimisation of the memory management and CPU usage. Moreover, the responsibility for the integration of the statistical software and new ROOT components in the CMSSW framework were taken over.

3.1 ROOT

ROOT is an object oriented framework written in C++ [49] that appeared in 1995 and is officially supported by CERN since 2003. Successor of PAW [50], during the years, it became the most frequently used tool for data analysis in High Energy Physics. It offers more than 1200 classes in a layered hierarchy, grouped in about 60 libraries. This hierarchy is organised in a mostly single-rooted class library, i.e. most of the classes inherit from the same base class called *TObject*. ROOT offers, among other tools, an advanced mathematical library, matrix calculus classes, various interfaces to the most popular minimisation packages and the implementation of a neural network toolkit. Moreover, it provides a full collection of classes for data representation, e.g. graphs and histograms in different dimensions. Many other graphics primitives like lines, arrows, geometrical shapes, text and legends are as well built-in. This allows the user to create plots that carry a dense amount of information without overhead and on a short time scale (figure 3.1).

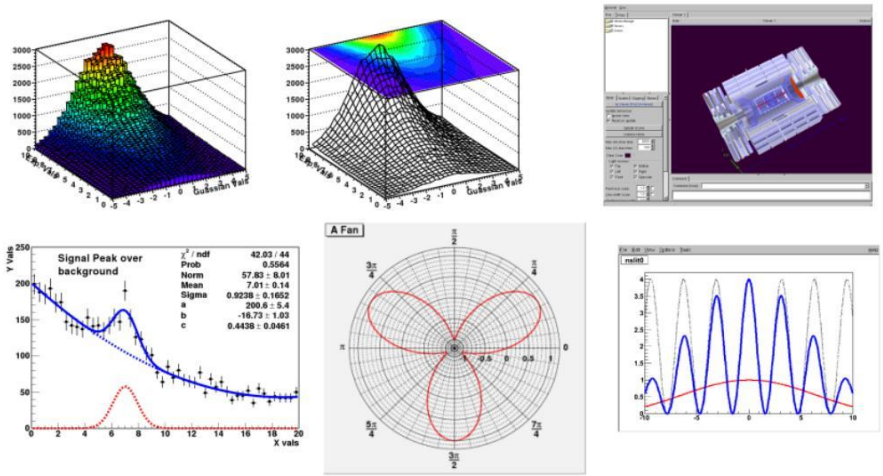

Figure 3.1: Some examples of ROOT plots, taken from [51, 46]. Almost every type of graphical representation is possible, up to complex three-dimensional objects representations.

Persistence on disk

A very important aspect of ROOT is the infrastructure offered for the persistence of data on disk. This is not a trivial issue since a large amount of programming languages, like C++ for example, do not offer native mechanisms to make memory structures and data types persistent. Therefore all the objects describing a collision event can be compressed and written in ROOT files by a certain application and retrieved by another program. The most powerful class used to store the information relative to physics events is the *TTree*, or simply ROOT tree. It extends the popular concept of row-wise and column-wise ntuples to complex

objects and data structures. A ROOT tree can contain a huge amount of objects and can be separated and saved in multiple ROOT files if needed. The ambitious idea is to exploit the same language, the same data model and the same style of data access across all available data sets in an experiment. CMS, as well as the other three main LHC experiments, opted for ROOT as the tool for recording their data and as the preferred environment for their analyses.

3.1.1 RooFit

The RooFit toolkit provides a framework for modelling the expected event data and measured distributions of one or more observables of a physics analysis. Models can be used for example to perform unbinned maximum likelihood fits or produce plots. RooFit was conceived as a tool for the analyses of the BaBar [52] collaboration and is now a part of ROOT.

RooFit Technology

The RooFit technology is based on an object oriented design, according to which almost every mathematical entity is represented by a class. A feature of this granular design philosophy is that all RooFit models always consist of multiple objects. For example, function parameters and variables are treated on equal footing and can be expressed as objects, representing real intervals, holding an (asymmetric) error and a fit range. Probability density functions (PDFs) are as well represented by classes and, exploiting their methods, sophisticated mathematical objects can be described. The numerous simple models natively provided in the package such as Gaussians, polynomials or Breit-Wigner distributions can be combined to build the elaborate shapes needed for the analyses. Many manipulations involving PDFs like simple addition, convolution or external product are supported. The communication between the class instances, bound together in elaborated structures that result from such operations, is granted by an advanced reference caching mechanism. Moreover, RooFit automatically takes care of the normalisation of the PDFs with respect to all of their parameters and variables within their defined ranges. In addition, the integration procedures are highly optimised, combining analytical treatment with advanced numerical techniques like the VEGAS [53] algorithm. Simultaneous and disjointed fits can be carried out with the possibility of interfacing RooFit to MINUIT [54] or other minimisation packages.

Persistence with Workspace

Every model and dataset can be made persistent on disk with the *RooWorkspace* class. A model representing the expected and measured distributions for a certain observable in data and Monte Carlo can therefore be built only once and later circulated in an electronic format and exploited by different users in association with different analysis procedures.

3.1.2 RooStats

RooStats is a framework for statistical analysis initiated as a fusion of the RooStatsCms [55] package (for more details see appendix B) and some prototype code presented at Phystat 2008 [56]. Overlooked by both the ATLAS and CMS statistics committees, RooStats is part of the official ROOT releases since December 2008 and rapidly became a widely accepted tool in the high energy physics community. It has been built on top of ROOT and RooFit: In some sense RooStats provides high-level statistical tools, while RooFit provides the core data modelling language as well as many low-level aspects of the functionality. The goal of the tool is to feed different statistical methods with the same input model and compare their outcome. The software also needs to be versatile in order to be able to cope with both simple analyses, like those based on number counting, and complex ones which use the parametrisation of experimental distributions.

3.2 CMSSW: The CMS Application Framework

CMSSW is a framework based on a collection of software packages, on an event data model (EDM) and services taking care of calibration, simulation and detector alignment together with modules for the reconstruction. The primary goal of CMSSW is to ease the development of reconstruction and analysis software by the physicists working at the experiment. The architecture of CMSSW foresees one single executable, the same for Monte Carlo and data samples, called `cmsRun`, and many plug-in components that encapsulate units of precise event-processing functionalities in form of algorithms. The parameters needed to specify the operations to be carried out by `cmsRun` are contained in a configuration file, interpreted at runtime and written in the Python programming language [57]. This file can also be user-specific and contains all the information about which data to use, which modules are needed, their specific parameters and their order of execution. The complete chain of modules is then executed for every event processed. An important feature of CMSSW is that only the modules specified in the configuration file are loaded, which reduces the overall memory footprint of the program.

The EDM Event

The central concept in the CMS EDM is the *event*. An event is concretely a very general C++ object container for all the raw and reconstructed data of a single collision. Reading and writing information in the event is the only way in which modules can pass information to each other. The event also contains metadata which describing for instance the configuration of the software used to produce the data present in the event. During the data processing, all or part of the objects accumulated in the events can be written to ROOT files exploiting the tree technology. Nevertheless, not all the information should be stored in the event. For example, detector calibrations, pedestals, run conditions, the status of the

accelerator or the geometrical description of the detector are updated periodically and not suited to be made persistent in event trees. Due to that a mechanism to read them, for example from databases, has been put in place, namely the EventSetup system.

CMSSW module types

Presently, six different types of modules are available in CMSSW, with different functionalities:

- **Source**: Reads the event and EventSetup data from DAQ or ROOT files, or in case of Monte Carlo production, generates empty events to be filled subsequently.
- **EDProducer**: Reads information in the event and elaborates them in order to write new objects into it.
- **EDFilter**: Evaluates the properties of objects in the event and returns a boolean value that can be used to stop the execution of the modules chain and skip to the next event.
- **EDAnalyzer**: Used to study the properties of an event and write some output, e.g. a histogram. Analyzers are neither allowed to write information to events nor to stop the execution of modules.
- **EDLooper**: These modules control the multi-pass looping over data.
- **OutputModule**: After the execution of all other modules, reads the data from the Event and stores it, also selectively, on external media.

The diagram in figure 3.2 shows a typical execution pattern of `cmsRun`.

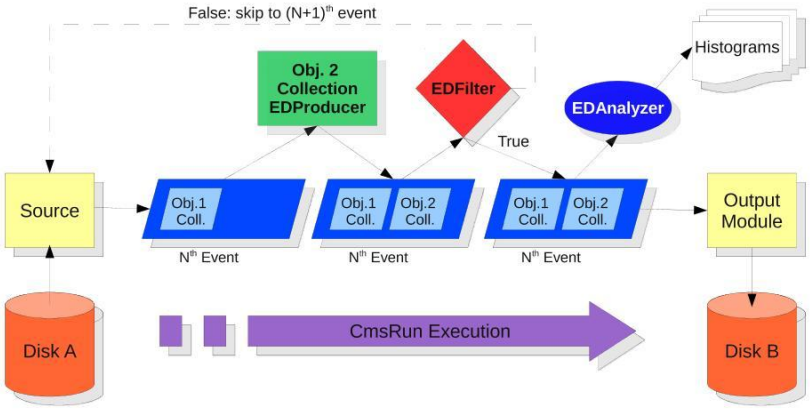

Figure 3.2: A sample module chain composed of a Source, one EDProducer, one EDFilter, one EDAnalyzer and an Output. The Source reads in this case the ROOT files where the trees with the events are stored. The EDProducer writes in every event a collection of objects of the type "2". Thus the EDFilter is configured to allow to proceed with the modules execution only if a certain condition concerning the objects in the events is satisfied. Before writing the updated events on disk, an EDAnalyzer fills a collection of histograms.

3.3 Monte Carlo Event Generation

Monte Carlo event generators are highly sophisticated computer programs that are used to produce "pseudo-events" of high energy collisions starting from the interactions between two incoming particles. The purpose is to provide a representation as accurate as possible of all the observables for a large range of reactions, with the aid of Monte Carlo techniques. For this, the latest possible theory developments have to be implemented within Monte Carlo generators. In many cases, like QCD phenomena, not all calculations can be carried out analytically, and therefore supplementary models are embedded. A large number of Monte Carlo event generators are available like Pythia [58], Herwig [59], Alpgen [60], MadGraph [61] or Sherpa [62]. CMSSW offers interfaces to many of them. The Monte Carlo events used for this thesis were generated with Pythia, versions 6.4 and 8.

3.3.1 Pythia

At the moment, Pythia is the principal Monte Carlo generator used by CMS. It contains implementations of models describing many physics aspects, as hard interactions, fragmentation and the underlying event.

The initial highly energetic interaction between two partons is referred to as hard scattering. Pythia provides leading order precision matrix elements for about 300 processes, within and beyond the SM. Radiative effects and higher order corrections are taken into account through the concept of parton showers starting from the particles resulting from the hard interaction.

The generator also offers a model for the hadronisation of the coloured states that might result from the hard scattering or parton showering, called the Lund string fragmentation model. Quarks and anti-quarks are thought as endpoints of one-dimensional relativistic objects, called "strings", which represent the colour field between them. As a result of the progressive separation of its endpoints, the string can break into hadron sized pieces via new quark-antiquark pair production, emulating QCD confinement [63]. This procedure is carried out until a colour neutral final state is reached. In this picture, outcoming gluons can result in kinks on strings, giving rise for example to angular momentum distributions of the produced hadrons.

UE, ISR and FSR

Furthermore, Pythia is able to describe the underlying event (multiple parton and beam-beam remnants interactions, see section 5.1.4) together with the initial and final state radiation. Such characterisation needs a large amount of parameters to configure the event generator. A complete set of such parameters is called *Tune* [58]. Many tunes to describe the extra activity in proton-(anti)proton collisions at different energies are available. They are the summary of the measurement campaigns that took place in the past, for example at LEP and Tevatron [64], and

the observations with the first LHC data.

All tunes used for this work, provide different descriptions of the non-diffractive component of the scattering process and are in agreement with CDF [65] data at $\sqrt{s} = 630\,\text{GeV}$ and $1.8\,\text{TeV}$ [66, 67]. The Pythiatune D6T [68, 69] has been adopted as default within the CMS collaboration. It is based on the CTEQ6LL [70] parton distribution functions and incorporates the information of the measurements of the UA5 collaboration at the $\text{Sp}\bar{\text{p}}\text{S}$ collider [71]. All other tunes like DW [69], Pro-Q20 [72], and Perugia-0 (P0) [73] conventionally use the CTEQ5L [74] PDFs. In addition, an improved description of hadron fragmentation based on LEP results has been taken into account for Pro-Q20 and P0. Tune P0 also uses the new PythiaMPI model [75], which is interleaved with a new p_T ordered parton showering.

As a consequence of the observed generally higher particle multiplicities in LHC collision data at $0.9\,\text{TeV}$ [76] and $7\,\text{TeV}$ [77] compared to model predictions, the new tune CW [78] was derived from DW. This tune manages to increase the UE activity while remaining consistent with the CDF results.

In addition, a new tune was recently created [79], the Z1 tune. This tune incorporates the knowledge acquired by the CMS QCD working group with the analysis of the $0.9\,\text{TeV}$ and $7\,\text{TeV}$ data.

The scheme in figure 3.3 illustrates the main steps carried out by a multi-purpose Monte Carlo generator like Pythia. In this work (according to the Pythia definition) particles are considered stable if they have an average proper lifetime of τ such that $c\tau > 10\,\text{mm}$.

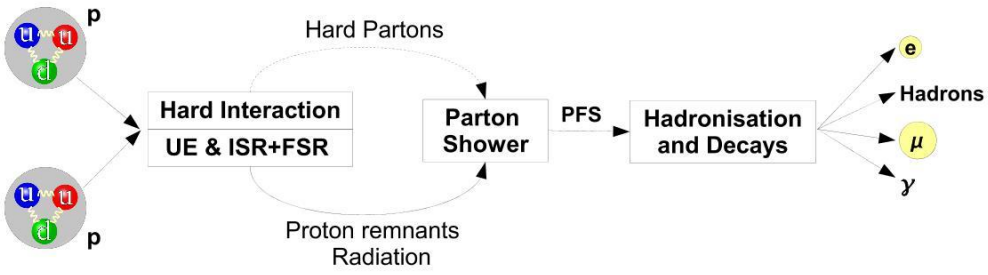

Figure 3.3: Overview of the different steps from the incoming protons to the colourless final state of a multi-purpose Monte Carlo generator. The hard interaction and the extra activity in the event is calculated. Subsequently the parton showering takes place. The partonic final state (PFS) is then hadronised. Eventual decays of the hadrons are simulated as well leading to the final stable particles [80].

3.4 Detector Simulation

The physics events coming from the event generator can be seen as what could be measured with a perfect detector. This is far from reality since the detection of particles is possible only via the measurement of the energy they deposit during their passage through matter, which is dominated by stochastic processes, and depends on attributes like the particles mass and electric, weak or strong charge. Moreover, the detector electronic scan suffer from malfunctioning or non-operational channels and its behaviour varies as a function of time depending also on external factors like temperature or machine conditions. In addition to that, every detector contains a certain amount of inactive material which is not able to detect particles. These materials are for example the power supply cables, cooling systems or the electronic readout boards.

CMSSW Simulation: three steps

CMSSW provides a very advanced detector simulation, that allows to compare Monte Carlo samples with the measured data. The detector simulation can be divided into three steps, namely the simulation of the interaction of particles in matter, the simulation of the signals coming from the readout of CMS and, finally, the simulation of the L1 trigger.
The energy deposition and interactions of particles in the CMS detector is achieved with the aid of the Geant4 [81] simulation toolkit. This package is able to describe with a complete range of functionalities including tracking, geometry and various physics models the electromagnetic and nuclear interactions in presence of a magnetic field, basing all the calculations on the CMS geometrical and material descriptions.
The simulation of the electronics behaviour leading to a digital signal is called *digitisation* and reproduces the real response of the readout components of the detector in presence of certain kinds of energy deposit. Finally the emulation of the L1 trigger takes place.

Pileup: Events Mixing

In the acquired data, given the high luminosity of the LHC, frequently more than one pair of protons interact, giving rise to pileup. The simulation is achieved through the superposition of event contents coming from different samples, usually a signal sample and a second one containing only the most frequent events to mimic the pileup contribution.

3.5 Reconstruction of Physics Objects

The reconstruction step leads to higher level physics objects starting from the output of the detector DAQ (see section 2.2.5) or simulation. The reconstruction of tracks, muons and jets performed by the algorithms implemented in CMSSW is characterised in the following.

3.5.1 Track Reconstruction

The reconstruction of particle tracks in a densely populated environment needs an efficient search of hits during pattern recognition and a fast propagation of trajectory candidates. The track reconstruction is decomposed in five logical steps:

- Hit reconstruction
- Seed generation
- Pattern recognition or trajectory building
- Ambiguity resolution
- Final track fit

First of all, the neighbouring clusters of electric charge deposited by the charged particles in the silicon strips and pixels are reconstructed to form *hits*.

Then the initial trajectory candidates for track reconstruction are provided by seed generation. Each seed consists of at least three hits that stem from one particle track. Alternatively, at least two hits and the nominal interaction point can be used.

Combinatorial Kalman Filter

The pattern recognition is based on the Combinatorial Kalman Filter (CKF) method [82], the main algorithm used in the track reconstruction of CMS. It basically consists of a least-squares minimisation. More specifically it is a local and recursive procedure, namely one track is reconstructed at a time and the estimates of its parameters are improved upon with every successive hit added. The filter therefore starts its operations from the seed level providing a coarse estimate of the track parameters. Progressively it includes the information of the successive detection layers, one by one, updating and improving the precision of the track parameters with every added measurement. Since multiple hits on a particular tracker layer may be compatible with the predicted trajectory, one additional trajectory is created per hit. The exponential increase of the number of such candidate trajectories is avoided discarding the least probable of them according to a certain criteria, for example comparing their reduced χ^2 with respect to a threshold value. The procedure stops when the whole available information has been integrated.

The ambiguities resulting from the fact that one track can arise from different seeds or that one seed is associated to multiple tracks are then resolved. The discriminating variable is the fraction of shared hits, defined as follows

$$f_{shared} = \frac{N_{hits}^{shared}}{min\left(N_1^{hits}, N_2^{hits}\right)}, \tag{3.1}$$

where $N^{hits}_{1(2)}$ is the number of hits in the first (second) track candidate. If f_{shared} is greater than 0.5, the track with the least number of hits is discarded. In case both tracks have the same number of hits, the track with the highest χ^2 value is discarded.

Finally, the remaining trajectories are refitted and then declared real *tracks*.

3.5.2 Muon Reconstruction

The muon reconstruction software produces two kinds of muon objects: With the information provided by the muon system alone (standalone muons) or in combination with the tracking device input (global muons).

Regional Reco

This design is called "regional reconstruction" and implies important savings in term of CPU resources needed. The muon system, in presence of a signal, can isolate a restricted region of interest in the tracker where the associated tracker track is expected to be found. Once the muon track is reconstructed in that region of interest in the tracker, the global muon reconstruction allows the CKF to navigate up to the hits obtained in the drift tubes, cathode strip chambers and resistive plate chambers detectors (see section 2.2.4).

Such a muon reconstruction has an extremely good performance, see table 3.1. Challenging objects like muons with a transverse momentum in the TeV range, which are characterised by a significant energy loss in matter and originate severe electromagnetic showers in the muon system can still be accurately measured.

Table 3.1: Transverse momentum resolution of muons exploiting stand alone and global reconstruction [20].

Muon p_T [GeV]	$\Delta p_T/p_T$ standalone	$\Delta p_T/p_T$ global
10	8-15%	0.8-1.5%
1000	16-53%	5-13%

3.5.3 Jet Reconstruction

Partons coming from proton collisions cannot be detected as isolated entities but the streams of particles resulting from their hadronisation are well measurable by CMS. A link between the properties of these final states and the originating partons is established through the construction of a jet, achieved via the deployment of a jet algorithm. Jet algorithms define a set of rules to group particles, involving one or more parameters that indicate how close, depending on a given

distance measure, two particles must be located for them to belong to the same jet. Additionally they are always associated with a recombination scheme which indicates what momentum to assign to the combination of two particles, e.g. their four vector sum.

Jet Definition

Taken together, a jet algorithm with its parameters and a recombination scheme form a "jet definition".

An agreement about the required general properties of jet definitions, the "Snowmass accord", was set out in 1990 by a joint group of theorists and experimenters. It contains the following features [83]:

1. Simple to implement in an experimental analysis.
2. Simple to implement in the theoretical calculation.
3. Defined at any order of perturbation theory.
4. Yields finite cross sections at any order of perturbation theory.
5. Yields a cross section that is relatively insensitive to hadronisation.

The first issue is particularly relevant for LHC, since a proton-proton event at high luminosity can contain more than two thousands particles and a heavy ion collision can easily result in four thousands final state particles. The jet algorithms must be fast enough to cluster all these particles together in fractions of a second.

Infrared and collinear safety

Issue number four introduces the fundamental concept of infrared safety requirement for jet algorithms. An infrared safe behaviour implies that the addition of a soft object does not affect the jets resulting from the input clustering. Such additional low momentum particles can be gluons radiated from initial particles or components originating from pile-up or the underlying event. An infrared safe algorithm is for example fundamental for next to leading order (NLO) calculations of inclusive jet cross section and its comparison with measured data [84], see figure 3.4. Another crucial issue concerning jet algorithms is collinear safety. A collinear safe algorithm does not change its output if an input particle is split into two collinear particles. This situation can arise, among other configurations, in presence of collinear radiation of gluons, see figure 3.5.

Several algorithms have been developed during the past years. In the following the official ones chosen by the CMS collaboration, Iterative Cone, k_T and anti-k_T are illustrated. A brief description of the SISCone algorithm is also given. To reconstruct the jets, CMSSW does not use individual implementations of jet algorithms but relies to the FastJet [85] package to which it is interfaced. Hence, the concept of the jet area is illustrated and a brief description of the jet inputs used in CMS is given.

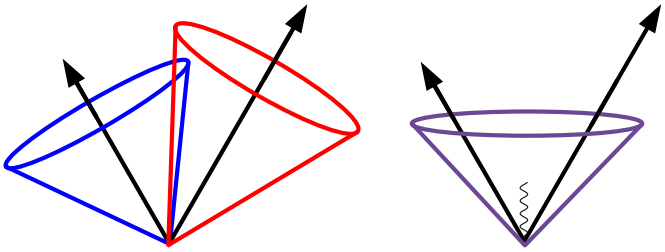

Figure 3.4: Illustration of the infrared sensitivity of a feebly designed jet algorithm. On the left, an event gave rise to two jets. When an additional soft parton is present, i.e. a radiated gluon, the particles are clustered into a single jet.

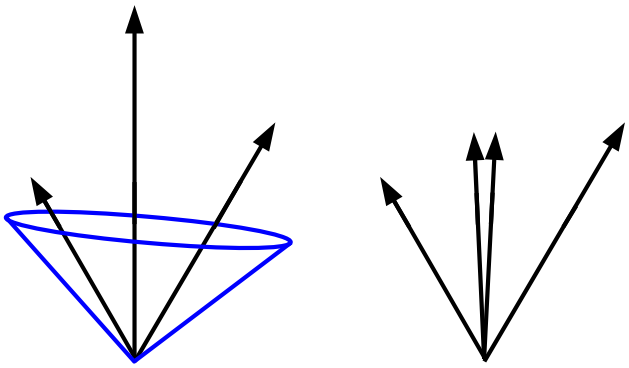

Figure 3.5: Illustration of the product of a collinear unsafe jet algorithm. The number of jets is changed in presence of collinear radiation and a transverse momentum threshold.

Iterative Cone Algorithm

The Iterative Cone (IC) algorithm is not collinear and infrared safe. Nevertheless, it is still present in the CMS official reconstruction chain since it is involved in high level trigger patterns (see section 2.2.5). The algorithm is suited for high level trigger code since it is fast and has a local behaviour, i.e. when its clustering is restricted to a portion of the detector, the jets are individuated as if the whole geometry were at disposal. To build IC jets, an iterative procedure is adopted. The object in the event with the biggest transverse energy is taken as seed and a cone with a radius $R = \sqrt{\delta\eta^2 + \delta\phi^2}$ is built around it. All the objects contained in that cone are merged into a proto-jet, the direction and transverse energy of which are defined as

$$E_T = \sum_i E_{Ti} \quad ; \quad \eta = \frac{1}{E_T} \sum_i E_{Ti} \cdot \eta_i \quad ; \quad \phi = \frac{1}{E_T} \sum_i E_{Ti} \cdot \phi_i. \tag{3.2}$$

The direction of the proto-jet is then taken as seed for a subsequent proto-jet. The procedure is repeated until the variation between two iterations of the proto-jet axis is below a certain threshold or a maximum number of iterations has been reached. At that point the proto-jet is declared a jet and its components are removed from the collection of objects to be clustered. The algorithm produces jets until all available objects in the event are exhausted.

SISCone Algorithm

Since April 2010, the algorithm is no longer in the CMS standard reconstruction chain. In any case it might be interesting to give a short description of its functionality since the Seedless Infrared-Safe Cone (SISCone) [86] was the first infrared and collinear safe cone algorithm that was developed. The SISCone does not rely on seeds like the IC algorithm, but seeks through an optimised procedure all stable cones in the event. This technique exploits the fact that a circle which encloses two input objects can be moved around such that two of the remaining objects lie on its circumference [85]. Reversing this allows the determination of all stable cones with a radius R by testing the circles defined by a pair of input objects and radius R (see figure 3.6). In case two cones share some input objects, a split-merge operation takes place. If the sum of the transverse energy of these objects is smaller than a parameter E_{split}, the objects are assigned to the nearest proto-jet in the η-ϕ plane. Otherwise, the two proto-jets are merged together. Unfortunately, even if the stable cone search is far more efficient than a brute force approach, it becomes too slow (order of seconds) for one typical LHC proton-proton event with pile-up.

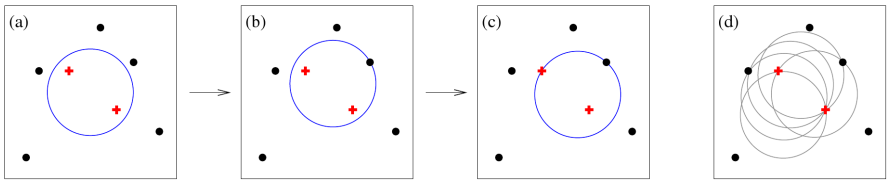

Figure 3.6: (a) Some initial circular enclosure; (b) moving the circle in a random direction until some enclosed or external point touches the edge of the circle; (c) pivoting the circle around the edge point until a second point touches the edge; (d) all circles defined by pairs of edge points leading to the same circular enclosure. Taken from [83].

Generalised k$_T$ Algorithm

In this paragraph, a description of the k$_T$, anti-k$_T$ and Cambridge-Aachen algorithms is given which groups them into a single class, the generalised k$_T$ algorithm. This approach has the advantage of underlying the different features of the algorithms from a unified perspective, but it does not follow the chronological order in which they were developed. The generalised k$_T$ algorithm represents a whole family of infrared and collinear safe algorithms depending on a continuous parameter, denoted as p. All these algorithms foresee sequential (pair-wise) recombination and are not based on a fixed geometrical shape. The basic feature underlying this class of jet algorithms is the dependence on the energy of the input objects beyond the radius parameter R. The fundamental quantities taken into account during the clustering procedure are the distance d_{ij} calculated for every object i with respect to every other object j and the distance d_{iB} calculated for every object i with respect to the beam. These distances are defined as

$$d_{ij} = min(k_{ti}^{2p}, k_{tj}^{2p})\Delta R_{ij}^2/R^2 \ , \ d_{iB} = k_{ti}^{2p}, \tag{3.3}$$

where k_t traditionally denotes the transverse momentum in this context. The steps in which the jets are created are the following:

1. For each pair of particles i, j calculate d_{ij} and, for each particle, d_{iB}.
2. Find the minimum d_{min} of all the d_{ij} and d_{iB}. If d_{min} is a d_{ij}, merge the i and j objects together summing their four momenta. Otherwise, declare the i object a jet and remove it from the input collection.
3. Repeat the procedure until no object is left.

k$_T$, C/A and anti-k$_T$

The cases where $p \in \{-1, 0, 1\}$ represent the most commonly used subset of clustering algorithms. Special names are assigned to them, k$_T$ [87] for $p = 1$, Cambridge/Aachen (C/A) [88] for $p = 0$ and anti-k$_T$ [89] for $p = -1$ (table 3.2). The

k_T algorithm merges at first the softest objects with the neighbouring harder ones, while the anti-k_T starts from the hardest and merges them with the neighbouring softer ones. As a result of this recombination, this latter algorithm behaves like a perfect cone algorithm (see the *Jet Area* section below for more details). On the other hand, the Cambridge/Aachen does not consider the energy of the cluster particles, but pure geometrical distances. It is interesting to observe that for $p \to -\infty$, the behaviour of the generalised k_T algorithm tends to the one of the iterative cone.

Table 3.2: The main features of the k_T, Cambridge/Aachen and anti-k_T infrared and collinear safe algorithms.

Name	p value	Description
k_T	1	Start from softer objects
C/A	0	d_{ij},d_{ib} pure geometrical quantities
anti-k_T	-1	Start from harder objects, cone shaped jets

The time necessary to produce the jets collection of a typical LHC event on a modern CPU is shown in figure 3.7 for several jet algorithms. The Fastjet implementation of the k_T algorithms provides the best performance.

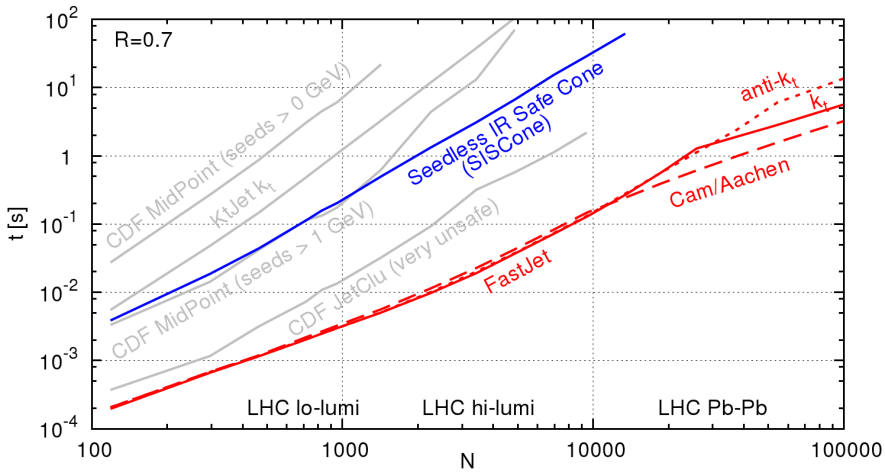

Figure 3.7: Timings for the clustering of a simulated 50 GeV dijet event, to which increasing numbers of simulated minimum-bias events have been added. Taken from [83].

Jet Area

The intuitive interpretation of the jet area is the measure of the $\eta-\phi$ plane portion covered by a jet. For a cone jet, the area is the size of the cone base. Nevertheless a precise definition is needed to make this important jet property unambiguous. Three different possible ways of defining the jet area are available [83], but in this thesis only *active areas* are considered. To obtain the active areas of jets a uniform background of extremely soft ($\simeq 10^{-100}$ GeV) and massless artificial objects, called *ghost particles*, are added to the event and allowed to be clustered with the physical objects. The area of a jet is then be proportional to the number of ghost particles which it contains and calculated. The areas of the jets obtained with k_T, C/A, SISCone and anti-k_T algorithms are shown in figure 3.8. The anti-k_T algorithm allows to obtain cone shaped jets [90]. It is important to note that the area of a k_T or C/A jet is influenced by the configuration of the neighbouring jets, reflecting in some sense the geometry of the event.

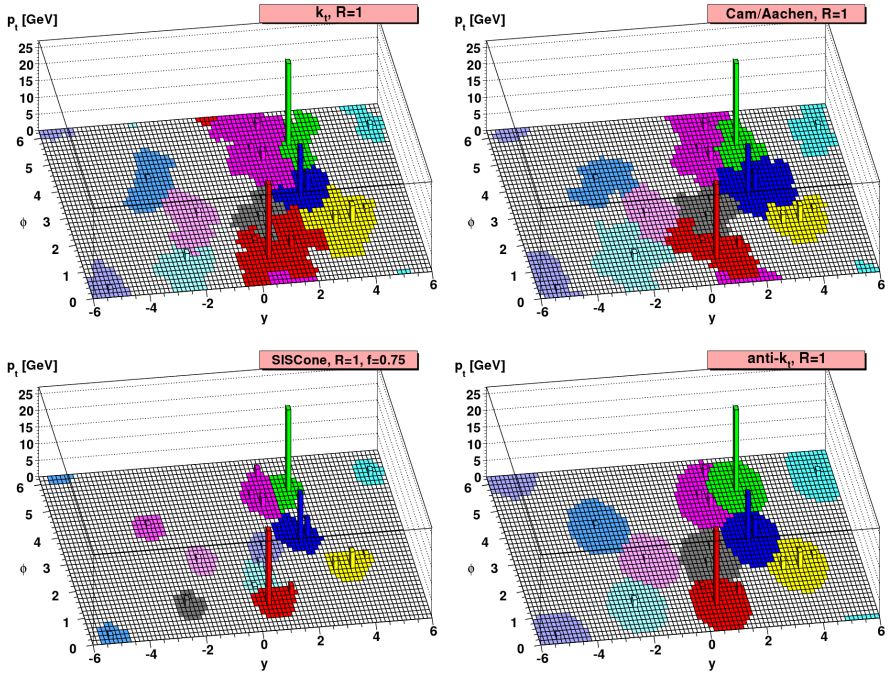

Figure 3.8: The jet areas obtained by clustering the particles in the same event using the k_T (upper left), C/A (upper right), SISCone (bottom left) and anti-k_T (bottom right) algorithms. Taken from [83].

Since all sensible algorithms are infrared safe, the presence of ghost particles does not influence the final collection of jets resulting from the clustering. The jets only made of ghost particles are called *ghost jets*. They are unphysical entities and normally not considered in physics analyses. It is interesting to remark that

in general the area of ghost jets is slightly smaller than the one of physical jets (all the details are described in [90]). Figure 3.9 shows the evolution of the jet area quantity for the leading jet in a Monte Carlo sample of $Z(\rightarrow \mu^+\mu^-)$+jet events versus the transverse momentum of the jet. For energies above 25 GeV, anti-k_T algorithm produces almost exclusively round shaped jets.

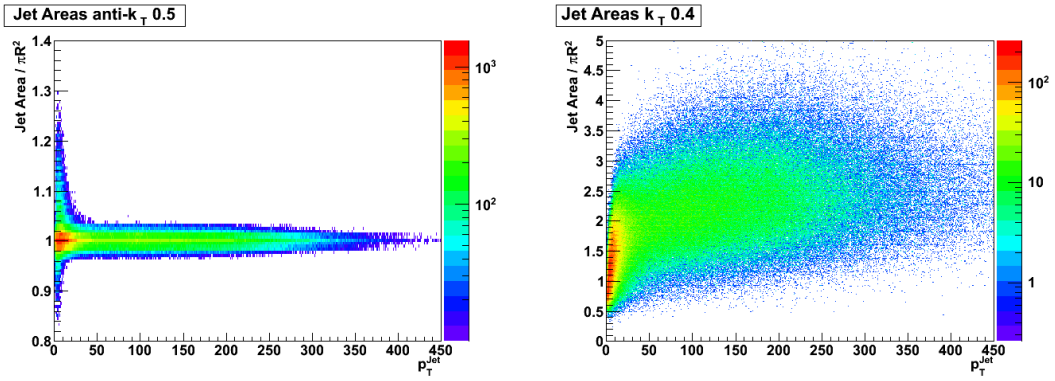

Figure 3.9: Distribution of the jet area versus jet pt for the leading jet in Monte Carlo $Z(\rightarrow \mu^+\mu^-)$+jet events. The anti-$k_T$ algorithm (left) behaves like a perfect cone algorithm for p_T^{jet} greater than 25 GeV in contrast to the k_T algorithm (right). Calorimeter jets were considered to produce these figures.

An event based variable based on jet areas which is useful to describe the procedure in chapter 6 is the *occupancy* C. This quantity is defined as the summed area $\sum_j A_j$ covered by all physical jets in an event divided by the considered detector region A_{tot}:

$$C = \frac{\sum\limits_{j \in \text{physical jets}} A_j}{A_{\text{tot}}} \,. \tag{3.4}$$

Generator Particle, Track, Calorimeter and Particle Flow Jets

CMS relies on five different types of objects as input collection for the jet algorithms. The first type consists of all the hadronic stable final states as produced by a Monte Carlo generator. The jets resulting from such objects are used for example for all the cross-checks necessary to design new analyses, accessing a source of "true" information. These jets are called *generator particle jets*.

Tracks of charged particles can also be used as input for jets. The resulting jets are called *track-jets* and are very useful for all those analyses that need to reach the lowest possible transverse momentum regions and for many validation purposes [91].

Jet algorithms can also be applied to other energy deposits in the CMS detector, for example the combination of calorimetric information, i.e. in HCAL and ECAL [92]. These jets are called *calorimeter jets*.

CMS provides also an algorithm to correct the energy of calorimeter jets using the momenta of charged particles measured in the tracker, the Jet Plus Tracks (JPT) algorithm [93]. The JPT corrections are not treated in this thesis.

A relatively recent development in the CMS community was the introduction of the Particle Flow Event Reconstruction [94]. This technique aims at reconstructing all stable particles in the event. Exploiting the redundant measurements of the CMS subdetectors, electrons, muons, photons and charged and neutral hadrons are reconstructed optimising the determination of particle types, directions and energies. The list of the objects obtained, can be used as input collection for the jet algorithms to construct the *particle flow jets*.

3.6 Grid Computing

The handling, analysis and distribution of the data coming from the four LHC experiments is a task of unprecedented magnitude and complexity. The storage, networking and processing resources necessary to analyse all and only the CMS data would exceed by far the capabilities of the central computing system at CERN. The solution to this issue is the Grid, an evolution of distributed computing, which became in the past decades one of the central concepts of HEP computing. The term Grid was born in the nineties to define an ensemble of worldwide distributed sites which offer resources as computing nodes, storage, specific software installations or data to users with specific credentials as if they were built into a simple desktop machine.

WLCG

The institution responsible for the deployment of the Grid services for LHC experiments is the Worldwide LHC Computing Grid (WLCG) [95]. WLCG is a global collaboration that besides the four LHC experiments involves several institutions from many countries and numerous national and international Grid projects. The mandate of WLCG is to build and maintain a data storage and analysis infrastructure for the entire high energy physics community that will use the LHC.

CMS computing model

The CMS computing model is designed for the seamless exploitation of Grid resources. It is based on the multitier structure established by the WLCG (figure 3.10). One single Tier-0 centre is built at CERN and accepts data from the CMS DAQ, archives it and performs a prompt first pass reconstruction. The Tier-0 distributes raw and processed data to a set of Tier-1 centres located in CMS collaborating countries, like GridKa [96], the German facility located at Karlsruhe or the CNAF[97] located at Bologna, Italy. These centres provide services for data

archiving on disk and tape storage systems, (re)reconstruction or calibration. A bigger set of Tier-2 centres, smaller than Tier-1s but pledging a substantial CPU power, offer resources for analysis, calibration activities and Monte Carlo samples production. The Tier-2 centres rely upon Tier-1s for fast access to large datasets and custodial storage of data. Tier-3 centres are smaller than Tier-2s and provide interactive resources for local analysis groups. The main analysis platform used to produce the results of this work is the German National Analysis Facility (NAF) [98], a computing centre coupled with the DESY Tier-2 in Hamburg.

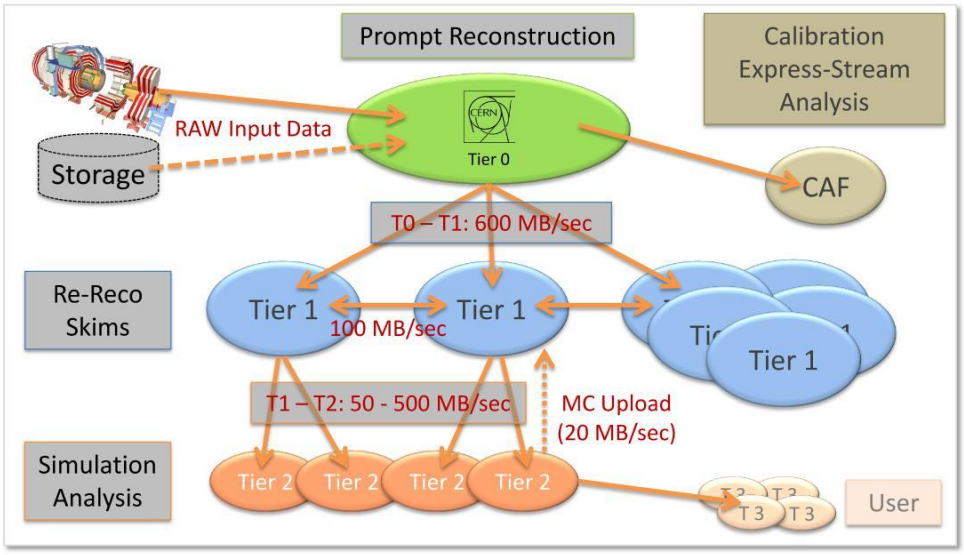

Figure 3.10: Schematic flow of data in the CMS Computing Model [99]. AOD stands for Analysis Object Data, a small subset of the complete collection of information produced by the reconstruction. Taken from [100].

Chapter 4

Statistical Combination of Higgs Channels

One of the primary goals of the LHC scientific program is the discovery or exclusion of the Standard Model Higgs boson (see chapter 1). No single Higgs decay channel has a discovery potential high enough to be used to discover the boson or to rule out its existence over the whole mass range. For this reason, the statistical combination of different decay channels joint to an accurate statistical treatment plays a fundamental role.

In this chapter, a brief introduction to statistical inference is given and two statistical techniques are described, the profile likelihood and hypothesis separation methods. Furthermore, the treatment of systematic uncertainties is briefly characterised. The first combinations and statistical treatments of Higgs discovery analyses performed in the scope of this thesis are then discussed. The software used to achieve this goal was the RooStatsCms package, also developed in the context of this thesis (see appendix B).
Even if the studies were performed assuming a centre of mass energy of 14 TeV, the combination exercises were extremely useful to establish a solid infrastructure made of statistical techniques, communication channels among the analysis groups and software tools.

4.1 Statistical Inference

Statistical inference is a very rich topic, which cannot be comprehensively treated in this thesis. An exhaustive discussion of this subject can be found e.g. in [101, 102]. When treating this subject, foundational differences can lead to different answers, and therefore one should consider them carefully. Two major classes of inference techniques can be identified: Classical/Frequentist and Bayesian [48].

Likelihood Function

A fundamental concept in every type of statistical inference is the likelihood function. In its simplest form the likelihood function is defined as:

$$L(\underline{x};\underline{\theta}) = \prod_{i=1}^{N} f(\underline{x}_i;\underline{\theta}) \tag{4.1}$$

where $\underline{x}_i = (x_a, x_b, x_c, \dots)$ are independent sets of N measured quantities, the distributions of which are described by a joint probability density function, $f(\underline{x}_i;\underline{\theta})$, where $\underline{\theta} = (\theta_1, \theta_2, \theta_3, \dots)$ is a set of K parameters. The f probability density function could also be different for every measurement i. The likelihood principle states that all the information about $\underline{\theta}$ obtainable from an experiment is contained in the likelihood function for $\underline{\theta}$, given a set of $\underline{x}_i$. Moreover, two likelihood functions for $\underline{\theta}$, from the same or different experiments, contain the same information if they are proportional to each other.

The likelihood function plays also an important role in parameter estimation. The principle of maximum likelihood states that the best estimate for the value of a parameter is the one which maximises the likelihood function. A complete treatment of parameter estimation can be found in [102].

4.1.1 Classical / Frequentist Inference

The classical or frequentist approach in statistics restricts itself to making statements of the form "probability to obtain the acquired data given the hypothesis". This approach is close to scientific reasoning, where probability means, given a large ensemble of samples, the relative frequency of something happening. Suppose a series of N events is considered, and n among them are of the type X: The frequentist probability that any single event is of the type X is then defined as the empirical limit of the frequency ratio

$$P(X) = \lim_{N\to\infty} \frac{n}{N} \tag{4.2}$$

This concept of probability is not related to any possible *a priori* belief. The price to be paid for this objectivity, is that this interpretation can be applied only in presence of repeatable phenomena.

4.1.2 Bayesian Inference

To define a probability that can be applied to non repeatable experiments, an alternative to the concept of frequency is needed. Among the various possibilities, the most relevant is the *degree of belief*, which is the basis of the Bayesian probability. The Bayesian approach does not need a certain phenomenon to be repeatable and is therefore valid also where frequentist reasoning cannot be applied. This kind of inference can be considered closer to everyday reasoning, where

probability is interpreted as a degree of belief that something will happen, or that a parameter will have a given value.

Bayes' Theorem

The name Bayesian derives from the extended use of the Bayes' Theorem [103] in this group of techniques:

$$P(A \mid B) = \frac{P(B \mid A) \cdot P(A)}{P(B)} \tag{4.3}$$

were:

- $P(A)$ is the *prior probability* of A. It represent how likely is A to be true, without taking into account any information on B.
- $P(A \mid B)$ is the *posterior probability*, i.e. the conditional probability of A, given B.
- $P(B \mid A)$ is the conditional probability of B given A.
- $P(B)$ acts as a normalising constant.

Taking into account the likelihood function, the theorem can be formulated with continuous probability density functions:

$$P(\underline{\theta} \mid \underline{x}) = \frac{L(\underline{x}; \underline{\theta}) \cdot P(\underline{\theta})}{\int L(\underline{x}; \underline{\theta}) \cdot P(\underline{\theta}) dx} \tag{4.4}$$

The Bayesian approach allows to make statements of the form "probability of the hypothesis given the data", which requires a *prior probability* of the hypothesis.

Formal Priors

A criticism that could arise against Bayesian inference is that subjectivism is one of its philosophical foundations. Indeed, a prior probability has to be *chosen*. Yet, most Bayesian analyses are performed with priors selected by "formal rules" (or "formal" priors [104, 105]). This strategy allows the viewpoint about the interpretation of priors to shift towards a *choice by convention*.

4.2 Intervals, Limits and Significances

In high energy physics, one of the possible expected outputs of a statistical method is represented by the regions in which the values of the parameters of interest are contained, or the exclusion regions where the parameter values are not contained.

Intervals and Limits

In the one-dimensional case, these regions become intervals or upper (and lower)

limits respectively. The interpretation of such intervals varies depending on the framework of reasoning used to obtain them: Frequentist or Bayesian. A frequentist interval is called *confidence interval* and is characterised by a *confidence level* (CL), a continuous parameter ranging from 0 to 1. The confidence interval represents the region in which the value of the parameter of interest would be located with a relative frequency of CL in the limit in which a very large number of repetitions of the measurement are performed.
A Bayesian interval is called *credibility interval* and represents the region in which the value of the parameter of interest is located with a certain probability.

Signal Significance

In addition to that, in an analysis aiming to a discovery, a typical studied quantity is the significance, S, of an observed signal [106]. Significance is usually understood as the number of standard deviations an observed signal is above the expected background fluctuations. Furthermore, S is implicitly understood to follow a standard Gaussian distribution with a mean of zero and a standard deviation of one. Therefore, to a given value of S (number of sigmas) corresponds the probability that the claimed signal is caused exclusively by fluctuations of the background, and this probability is obtained by performing the corresponding integrals of a standard Gaussian distribution. The general arbitrary agreement is that, in presence of a discovery, the value of S should exceed 5, which is equivalent to state that the probability that a Gaussian upward fluctuation of the background that mimics the signal is characterised by a probability of about $2.9 \cdot 10^{-7}$.

4.3 The Profile Likelihood Method

This frequentist method is used to estimate the best value and confidence interval or limits of a parameter. Moreover, it can be exploited to build an estimator of the signal significance.
In the following, the discussion will be limited to the one-dimensional case. The likelihood function, defined in equation 4.1, is here expressed as a function of a single parameter of interest, θ_0, and $K-1$ remaining parameters. The profile likelihood is defined as

$$\lambda(\theta_0) = \frac{L(\theta_0, \hat{\hat{\underline{\theta}}}_{i\neq 0})}{L(\hat{\theta}_0, \hat{\underline{\theta}}_{i\neq 0})} \tag{4.5}$$

The denominator $L(\hat{\theta_0})$ is obtained maximising the likelihood with respect to all parameters in the model, while for the numerator $L(\theta_0)$ the maximisation procedure is carried out after fixing θ_0 to a certain value and varying the remaining K-1 parameters. In the asymptotic limit (i.e. in presence of a large number of independent measurements) the likelihood function becomes a Gaussian centred

around the maximum likelihood estimator for θ_0 and therefore:

$$-2\ln\lambda(\theta_0) = -2(\ln L(\theta_0) - \ln L(\hat{\theta}_0)) = n_\sigma^2,\ n_\sigma = \frac{\theta_0 - \hat{\theta}_0}{\sigma} \tag{4.6}$$

where σ represent the Gaussian standard deviation of the parameter θ_0.

Intervals and Limits

With this construction, it is possible, in the asymptotic limit, to obtain the one- or two-sided confidence intervals with a graphical method, simply inspecting the intersections of the profile likelihood function with horizontal lines (figure 4.2), the y-coordinate of which is expressed by:

$$\begin{aligned} two-sidedintervals &= \sqrt{2}\cdot Erf^{-1}(CL) \\ one-sidedintervals &= \sqrt{2}\cdot Erf^{-1}(2\cdot CL - 1) \end{aligned} \tag{4.7}$$

where Erf is the error function:

$$Erf(x) = \frac{2}{\sqrt{\pi}}\int_0^x e^{-t^2}dt \tag{4.8}$$

Equations 4.7 are used since they allow to transform a confidence level into a number of sigmas, for a Gaussian with mean zero and a standard deviation of one. For the one-sided intervals, the Gaussian is considered to be defined only for positive values. Even if the asymptotic limit is not reached (i.e. in presence of a non parabolic $-2\ln\lambda$), it can be shown that this approach is still valid [102]. The main argument behind this statement is the assumption of the existence of a transformation $g = g(\theta_0)$ which makes the likelihood L Gaussian. Since the likelihood is a function of $\underline{x}$, no Jacobian is involved in going from $L(\underline{x};\theta_0)$ to $L(\underline{x};g(\theta_0))$, and hence $L(\underline{x};\theta_0) = L(\underline{x};g(\theta_0))$, see figure 4.1. A problem might be that such a

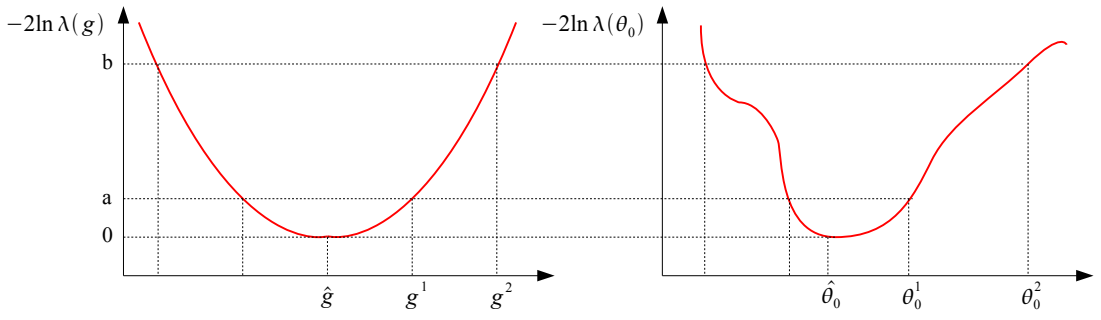

Figure 4.1: The direct usage of the transformation g is not necessary. The intervals for θ_0 can be found directly since no Jacobian is involved in the transformation $L(\underline{x};\theta_0) \rightarrow L(\underline{x};g(\theta_0))$.

g might not exist, but since the intervals for θ_0 can be estimated directly, without using any transformation, g can be only adopted as an assumption.

This method of interval estimation is sometimes also cited in the physics community as the "MINOS method" since it has been implemented by the MINOS algorithm of the Minuit [54] minimisation package. Figure 4.2 shows an example of profile likelihood function for the θ_0 parameter. The best-fit value for the θ_0 is represented by the abscissa of the profile likelihood function minimum.

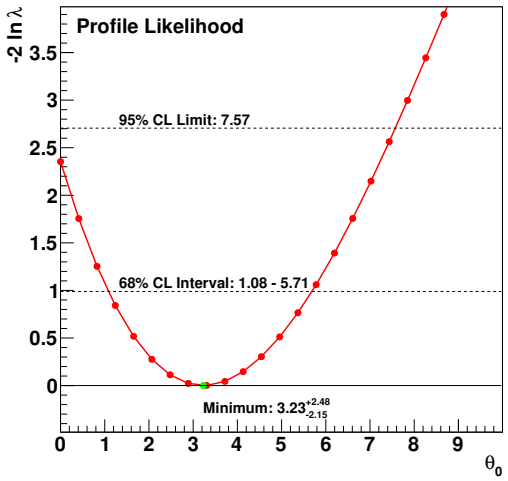

Figure 4.2: Likelihood scan over the parameter θ_0 in the case of a two-sided 68% CL and one-sided and one-sided 95% CL intervals. The minimum point abscissa is the best fit value for the θ_0 parameter. For each of the red points, the θ_0 parameter was fixed and the other parameters optimised through a fit procedure in order to maximise the likelihood. The figure was obtained with the RooStatsCms tool described in appendix B.

Significance

The profile likelihood method provides a well-behaved estimator of the signal significance [106], often quoted as S_L defined as

$$S_L = \sqrt{-2 \cdot \ln \lambda(yield^{signal})|_{yield^{signal}=0}} \tag{4.9}$$

If the θ_0 parameter in figure 4.2 is considered to be the signal yield, the value of S_L is therefore equal to the square root of the y-value of the intersection point common to the profile likelihood curve and the y axis.

4.4 Separation of Hypotheses

The interpretation of results for new particles and phenomena near the sensitivity limit of an experiment is a common problem in particle physics. Such a search

analysis can be formulated in terms of a hypothesis test [107]. The null hypothesis H_0 (or *background only* hypothesis) is that the searched signal is absent, the alternate hypothesis H_1 (or *signal plus background* hypothesis) simply that it exists. A test statistic, i.e. a function of the observables and model parameters, is to be defined, which could rank measurements from least to most signal-like and vice versa. Finally, the rules for discovery and exclusion should be formalised, in other words point out those test statistic intervals in which the observation leads to one or another conclusion. Most of the times, it is also desirable to quote an exclusion interval or a significance for the signal.

The test statistic is usually indicated with the symbol Q and the value of the test statistic for the data being observed with $Q_{\rm obs}$. A comparison of this quantity with the expected probability distributions dP/dQ for both null and alternate hypotheses allows to compute the confidence levels:

$$CL_{sb} = P_{sb}(Q < Q_{\rm obs}), \quad \text{where} \quad P_{sb}(Q < Q_{\rm obs}) = \int_{-\infty}^{Q_{\rm obs}} \frac{dP_{sb}}{dQ} dQ, \tag{4.10}$$

$$CL_b = P_b(Q < Q_{\rm obs}), \quad \text{where} \quad P_b(Q < Q_{\rm obs}) = \int_{-\infty}^{Q_{\rm obs}} \frac{dP_b}{dQ} dQ. \tag{4.11}$$

Small values of CL_{sb} point out poor compatibility with the H_1 hypothesis and favour the H_0 hypothesis and vice versa. In absence of measured data, the expected CL value can be calculated assuming the test statistic for the observed data to be numerically equivalent to the median of the signal plus background (background only) distribution.

Toy Monte Carlo Experiments

The functional form of the dP_{sb}/dQ and dP_b/dQ distributions is in general not known a priori. Therefore, histograms representing them can be built with a Monte Carlo re-sampling strategy. The technique consists in performing many *toy Monte Carlo* experiments, namely generating background only and signal plus background pseudo datasets, sampling the distributions of the null and alternate hypotheses. For each dataset, the value of the test statistic is then calculated providing one entry for the histograms representing dP_{sb}/dQ and dP_b/dQ. Since the number of toy Monte Carlo experiments which can be performed is limited, the calculated CL_{sb} and CL_b values are affected by a statistical uncertainty.
Given N toy Monte Carlo experiments with N greater than 20, a good description of this uncertainty is the binomial one [108]:

$$\sigma_{CL_x} = \sqrt{\frac{CL_x(1 - CL_x)}{N}} \tag{4.12}$$

where CL_x stands for CL_{sb} or CL_b.

A commonly used test statistic consists in the ratio of the likelihood functions for signal plus background and background only hypotheses: $Q = L_{sb}/L_b$. In this

case, usually, one refers to $-2\ln Q$ instead of Q as test statistic. An example plot of $-2\ln Q$ distributions is shown in figure 4.3.

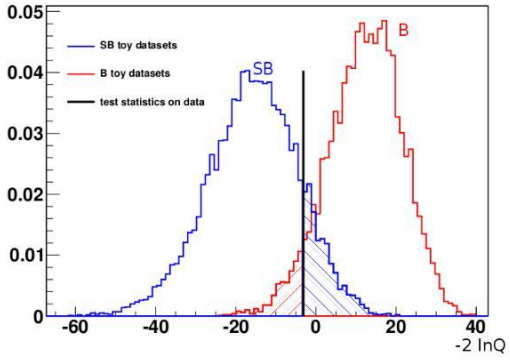

Figure 4.3: The distributions of $-2\ln Q$ in the background-only (red, on the right) and signal plus background (blue, on the left) hypotheses. The black line represents the value of $-2\ln Q$ on the tested data. The shaded areas represent $1 - CL_b$ (red on the left) and CL_{sb} (blue on the right). Each entry of the two histograms is the outcome of a toy Monte Carlo experiment.

Alternative choices of Q are possible, for example the number of signal and background events.

Significance

The signal significance can be obtained transforming CL_b in number of Gaussian sigmas using the equation:

$$S = n_\sigma = \sqrt{2} \cdot Erf^{-1}(2 \cdot CL_b - 1) \tag{4.13}$$

which quantifies in terms of Gaussian deviations how far the CL_b value is from the one expected in the background only hypothesis (CL_b=0.5).

Limits

A certain model can be excluded given the tested data at a certain CL if $1 - CL_{sb}$ is smaller than CL [1]. The signal plus background hypothesis can be altered multiplying the expected cross section of the signal σ_{sig}^{exp}, by a real variable R, obtaining what is called a "composite hypothesis" for the signal plus background case. Therefore, it is possible to scan the values of R so to find the value for which the signal plus background hypothesis can be excluded with a certain CL^{excl}, R^{excl} (figure 4.4). In other words, the smallest value of R satisfying the equation

$$1 - CL_{sb}(R) < CL^{excl} \Rightarrow CL_{sb}(R) > 1 - CL^{excl} \tag{4.14}$$

[1] Alternatively the CL_s prescription can be used, see section 4.4.1

The exclusion limit on the cross section will then be quantified by $R_{excl} \cdot \sigma_{sig}^{exp}$. When exclusion of standard model predictions of cross sections are involved, one typical way of referring to R is σ/σ_{SM}.

This method, known as *hypothesis test inversion*, involves the generation of a large amount of toy Monte Carlo experiments, a very expensive procedure in terms of computing time. An optimised search algorithm to minimise the number of tested R values is needed, like the one implemented for RooStatsCms [109]. An extensive treatment of the error estimation on the R value can be found in [108]. It should be observed that confidence intervals obtained with this technique, although widely used [110] do not have the same meaning as the ones obtained with the profile likelihood method or the Bayesian credibility intervals.

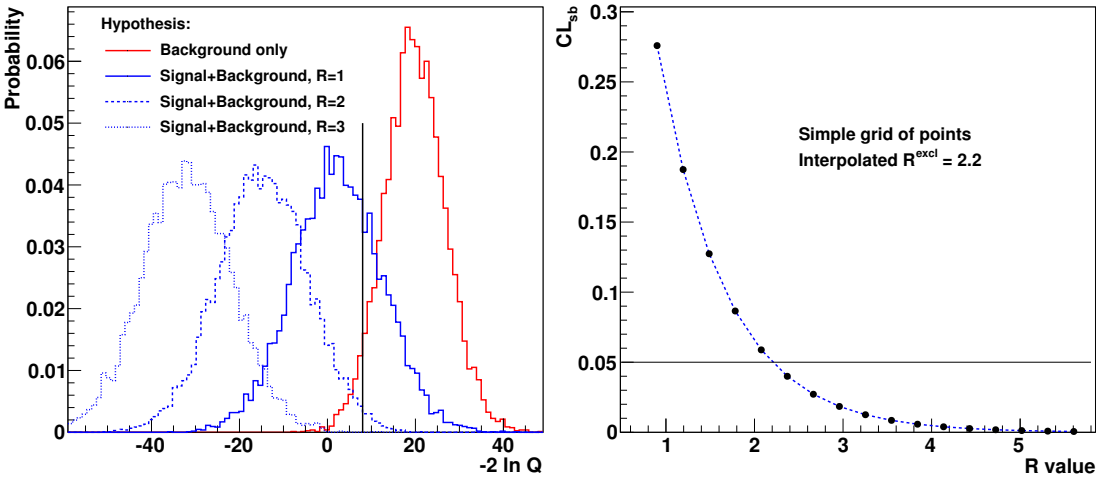

Figure 4.4: Illustration of a hypothesis test inversion. Left: The separation between the two hypotheses increases for increasing values of R. The black vertical line represents the value of the test statistic on the measured data. Right: The values of R are scanned until CL_{sb} is equal to 5% (95% confidence level exclusion). In this case, a simple grid of values is used together with a linear interpolation between the two points nearest to the desired CL_{sb} value.

4.4.1 The CL_s Prescription

Taking into account the presence of background in the data may result in a value of the estimator of a model parameter which is "unphysical". For example, observing less than the mean expected number of background events could be accommodated better if the signal cross-section was negative.

A possible strategy to treat this case is to follow the CL_s prescription, introducing the so called *modified*, or *conservative*, *frequentist approach*. The name refers to the fact that by construction the limits to which it leads are always less aggressive with respect to the ones obtained with the CL_{sb} quantity. This prescription consists in the normalisation of the confidence level observed for the H_1 hypothesis, CL_{sb}, with the one observed for the H_0 hypothesis, CL_b. The normalisation is simply defined as:

$$CL_s = CL_{sb}/CL_b \tag{4.15}$$

In this way, it is possible to obtain sensible limits on the number of observed signal events even if the observed yield is so low to compel the H_0 hypothesis. The CL_s quantity can be seen as the signal confidence level, as if every background event would have been discarded. Even if CL_s is not technically a confidence level, the signal hypothesis can be considered excluded with a certain confidence level CL when $1 - CL_s <$ CL. A more complete characterisation of the CL_s quantity can be found in [107].

4.5 Inclusion of Systematic Uncertainties: Hybrid Techniques

Every analysis must deal with systematic uncertainties. From the statistical point of view, this kind of uncertainty can be taken into account by diverse techniques. Two cases are discussed, namely the *profiling* and the *marginalisation* via Monte Carlo sampling.

For what concerns the profile likelihood method, a very convenient approach is to use the profiling procedure while in the hypothesis testing a Monte Carlo marginalisation technique is more suited.

Pragmatism in High Energy Physics

Both methods require to assume a probability distribution for the parameters affected by systematic uncertainties, or *nuisance* parameters. This probability distribution would be called the *prior probability* in a Bayesian context. It is therefore common practice in high energy physics to relax foundational rigour and pragmatically embrace *hybrid techniques*, for example plugging Bayesian concepts in purely frequentist methods. Indeed, incorporating nuisance parameters in the statistical treatment of data can lead to many difficulties if one sticks exclusively to one single class of statistical inference.

Profiling

The profiling (see section 4.3) takes place through the maximisation of the profile likelihood function built taking into account the systematic uncertainties and their correlations. Suppose that the θ_s parameter is affected by the systematic uncertainty described by the $h(\theta_s)$ probability density function: the joint pdf

describing the data and parameters can be written as

$$f'(\underline{x};\underline{\theta}) = f(\underline{x};\underline{\theta}) \cdot h(\theta_s). \tag{4.16}$$

In the log-likelihood function, $h(\theta_s)$ contributes as an additive penalty term. Therefore, the log-likelihood can be written as:

$$\ln L = \ln \prod_{i=1}^{N} f(\underline{x}_i;\underline{\theta}) \cdot g(\theta_s) = \ln g(\theta_s) + \sum_{i=1}^{N} \ln f(\underline{x}_i;\underline{\theta}) \tag{4.17}$$

The scan of this altered log-likelihood preserves the position of the minimum point but implies a smaller curvature and hence broader confidence intervals and less aggressive limits (see figure 4.5).

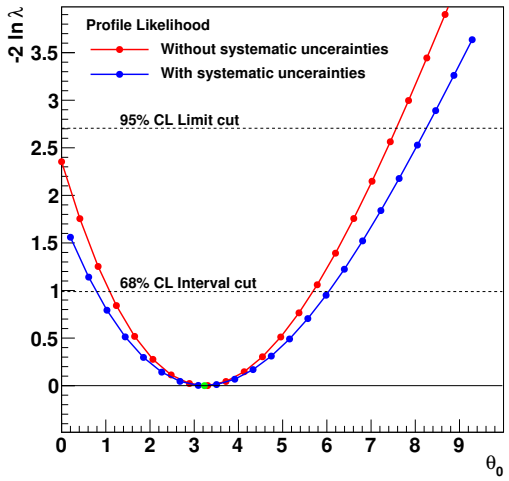

Figure 4.5: Illustration of the effect of systematic uncertainties on the profile likelihood function. The curvature decreases leading to less aggressive limits and broader intervals, while the minimum position is not altered.

In the following, two simple examples of penalty terms are illustrated, namely the representation of the penalty term in the log-likelihood for a Gaussian and in the multi-dimensional case with a multivariate Gaussian systematic uncertainty. If $h(\theta_s)$ distribution is Gaussian the penalty term looks like

$$\ln h(\theta_s) = \ln G(\theta_s;\mu_{\theta_s},\sigma_{\theta_s}) = -\frac{1}{2}\ln(2\pi\sigma_{\theta_s}^2) - \frac{1}{2}\left(\frac{\theta_s - \mu_{\theta_s}}{\sigma\theta_s}\right)^2 \tag{4.18}$$

Translating the problem in multiple dimensions and taking into account the multivariate Gaussian distribution in n dimensions, the penalty term assumes the

form:

$$\ln h(\underline{\theta_s}) = \ln G(\underline{\theta_s}; \underline{\mu_{\theta_s}}, \Sigma_{\theta_s}) = -\frac{n}{2}\ln(2\pi|\Sigma_{\theta_s}|^{1/n}) - \frac{1}{2}(\underline{\theta_s} - \underline{\mu_{\theta_s}})^T\Sigma^{-1}(\underline{\theta_s} - \underline{\mu_{\theta_s}}) \tag{4.19}$$

where Σ is the covariance matrix.

Marginalisation

The marginalisation approach, suited for the hypotheses separation technique, consists in varying for each toy Monte Carlo experiment the effective value of the nuisance parameters according to their distributions before generating the toy dataset itself. Hence, the whole phase space of the nuisance parameters is sampled through Monte Carlo integration. The final net effect consists in a broadening of the Q $(-2lnQ)$ distributions without a variation of their centres position. This leads to a degraded discrimination power between the two hypotheses (see figure 4.6).

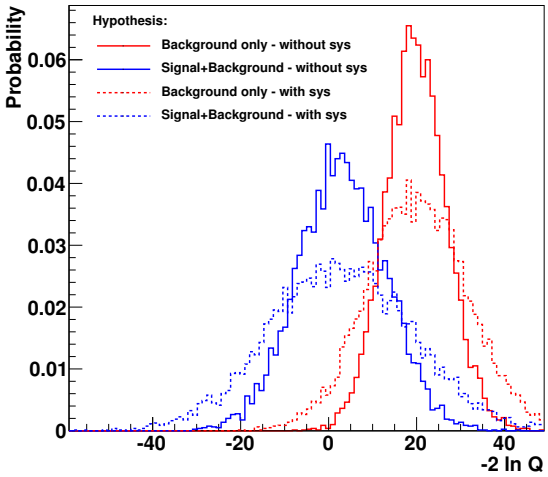

Figure 4.6: The systematic uncertainties degrade the discrimination power provided by the method. In particular, the width of the distributions increases while their central values remain basically unchanged.

4.6 First Higgs Analyses Combinations at CMS

Before the RooStats tool described in section 3.1.2 was released, RooStatsCms (see appendix B) was used in many occasions in the CMS collaboration. In this section, the results obtained in the scope of this thesis for the preparation of the Standard Model Higgs boson searches in τ pair [111] and W [112] boson pair decay channels are presented as an example of application of the afore-described

statistical methods. In addition, the first combination of different Higgs analyses performed within the CMS Higgs working group is presented [113]. Even if the integrated luminosity considered was 1 fb^{-1} and the centre of mass energy of the collisions 14 TeV, these exercises of combinations and statistical studies are of great relevance since they steered the collaboration towards common techniques and software tools and established communication channels among the analysis groups exchanging the results to be combined.

4.6.1 The Vector Boson Fusion $H \to \tau\tau$ Channels

An interesting alternative to the gluon-gluon fusion production of the Higgs boson is the vector boson fusion channel (see section 1.8.2). The SM Higgs boson produced via the vector boson fusion mechanism and decaying in τ lepton pairs is a crucial channel in the search for the Higgs boson in the mass range between 115 and 145 GeV. This interval is of primary importance since this region is suggested by the LEP combined measurements to be the one where the mass of the Standard model Higgs boson lies (section 1.8.2). Furthermore, this region is not yet excluded by the combined D0 and CDF combined limits [110]. The selection strategy conceived for this analysis concerned the final state where a τ decays hadronically, forming a jet, and the other decays leptonically, originating an electron or a muon. The events containing a muon or an electron in the final state are treated separately. The results in terms of expected signal significances and expected exclusion limits are obtained combining these two different channels and treating them using the hypothesis separation technique, marginalising the systematic uncertainties. A full description of the analysis can be found in [111].

The expected significance calculated according to equation 4.13 is always smaller than 1 (figure 4.7). No excess due to signal over the background can be therefore claimed over the whole mass spectrum investigated.

The limits on the cross section are also estimated. In figure 4.8, the limits obtained are shown. The bands are obtained fluctuating upwards and downwards by one (green) and two (yellow) standard deviations the number of background events. Such a plot can be only obtained exploiting a computer farm, indeed, the large number of toy Monte Carlo experiments performed costed in total 80 hours of computing time.

4.6.2 The $H \to WW$ Channels

The outcome of this analysis is relevant for the mass range between 140 and 190 GeV, with particular emphasis on the region around $\sim 2 \times M_W$ (see section 1.8.2). Here, three final state topologies are distinguished, namely e^+e^-, $\mu^+\mu^-$ and $e^\pm\mu^\mp$, all characterised by missing energy due to undetected neutrinos. All the details about this study can be found in [112]. The results for the

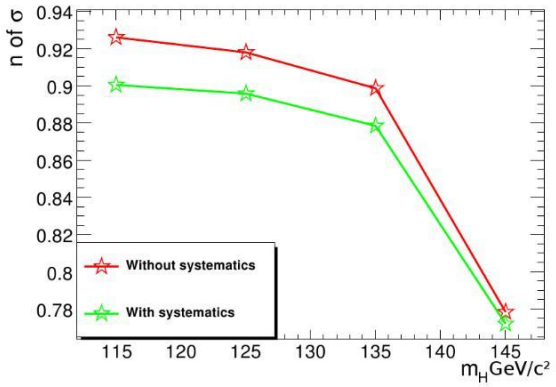

Figure 4.7: Expected signal significance for different Higgs mass hypotheses. No signal excess over the background can be claimed over the whole mass range.

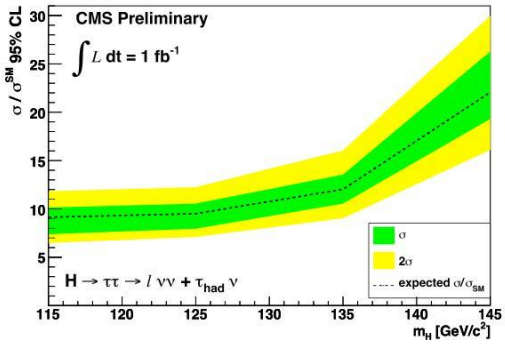

Figure 4.8: The 95% CL expected limit on the cross section as function of the Higgs mass hypothesis. The 1σ and 2σ bands are originated from the statistical fluctuation of the number of background events. The creation of this plot costed about 80 hours computing time with RooStatsCms (see appendix B).

expected significance and exclusion limits are obtained after the combination of these three channels, considering their background and signal yields affected by Gaussian systematic uncertainties, with an 80% correlation among them.

The expected significance as a function of the Higgs mass hypotheses was calculated using the S_L estimator (see section 4.3), incorporating the systematic uncertainties as penalty terms in the likelihood function. Figure 4.9 shows the expected significances as a function of the Higgs boson mass.

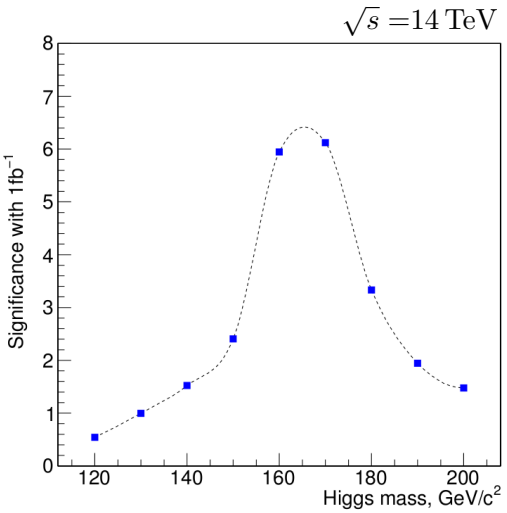

Figure 4.9: The expected signal significance of an event excess over the background as function of the Higgs mass hypotheses is shown.

The upper limits on the cross section were calculated both with the profile likelihood method and the conservative frequentist approach. A third Bayesian calculation was also performed as additional cross-check. The obtained values as a function of the Higgs boson mass is shown in figure 4.10.

The values of the limits in deserve some special attention. It is important to observe how three different approaches led to consistent results, and in particular how the profile likelihood method yielded slightly more aggressive limits with respect to the conservative frequentist method.

Green and Yellow Plot

Another way to visualise the results obtained through the hypotheses separation technique, is the so called "green and yellow" plot (figure 4.11). At glance, the expected separation power between the signal plus background and background only hypotheses is shown as a function of the different Higgs mass hypotheses. The black dashed line represents the expected value of the test statistic in the background only hypothesis (the median of the red distribution in figure 4.3). The green and the yellow band are the 68% and 95% confidence level intervals around this expected value. On the other hand, the red line represents the expected value

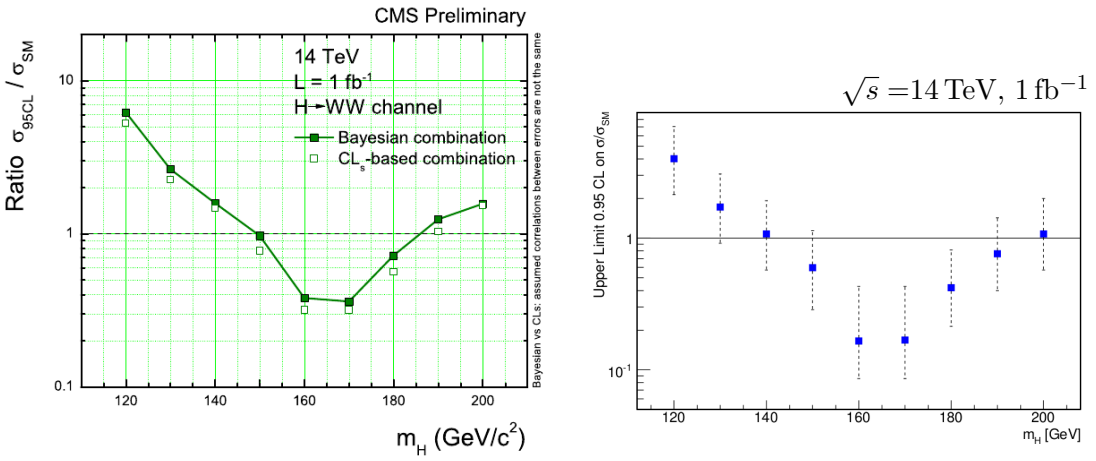

Figure 4.10: The 95% exclusion limits on the cross section as a function of the Higgs mass hypotheses using the conservative frequentist compared with a Bayesian tool (left) and the profile likelihood method (right). The error bars represent the one sigma statistical uncertainty on the limit values, estimated with many toy Monte Carlo experiments.

of the test statistics in the signal plus background hypothesis (median of the blue distribution in figure 4.3). As expected, the best expected separation between hypotheses is reached in the mass region around $2 \times M_W$.

4.6.3 The Combination of $H \to WW$ and $H \to ZZ$ Analyses

The combination of different decay channels was also performed in the CMS Higgs Working Group in the context of this thesis. The analyses considered were two, for a total of four channels combined, namely the $H \to WW$ (e^+e^-, $\mu^+\mu^-$ and $e^\pm\mu^\mp$ final states topologies), already discussed above, and the $H \to ZZ$ channel, for which the yields of the $4e$, 4μ and $2e2\mu$ final states were lumped together. The exclusion limits on the Higgs production cross sections were obtained with the conservative frequentist approach, treating the systematic uncertainties via marginalisation. An internal CMS Bayesian tool was used to cross-check the results. The exclusion limits on the Higgs production cross section are shown in figure 4.12. It is interesting to observe how different statistical methods produced compatible results.

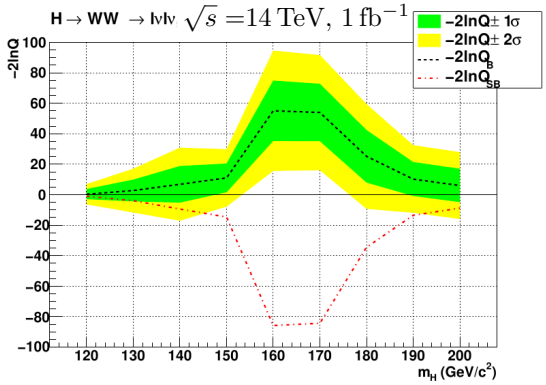

Figure 4.11: The expected separation power is shown in function of the Higgs mass. For every mass hypothesis, a plot like the one in figure 4.3 is produced. The dashed black line represents the mean of the test statistic distributions in the background only hypothesis while the dotted red line the mean of the signal plus background only one. The green (yellow) band represents the expected one (two) sigma fluctuation around the expected value in the signal plus background hypothesis.

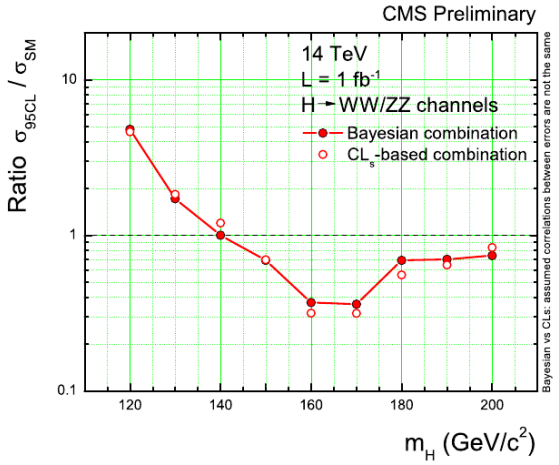

Figure 4.12: The 95% exclusion limits on the Higgs cross section as a function of the mass hypotheses using the conservative frequentist method. In the region from 160 to 190 GeV, one can see the improvement brought in by the $H \rightarrow ZZ$ analysis. A Bayesian CMS internal tool was used as a cross check.

4.7 Summary

In this chapter, the likelihood function and the basic principles of Bayesian and Frequentist inference were introduced. In addition, two procedures to obtain limits and signal significance were described, namely the profile likelihood and the hypotheses separation methods. In addition, two strategies for the inclusion of systematic uncertainties were discussed: Marginalisation and profiling.
These techniques were implemented in the RooStatsCms framework for analyses modelling, combination and statistical studies (see appendix B). The first large scale usage of this tool concerned a scenario featuring 14 TeV centre of mass energy and $1\,\mathrm{fb}^{-1}$ of integrated luminosity at CMS. Under those conditions, the combinations of the $H \to \tau\tau$, $H \to WW$ and $H \to ZZ$ decay channels were carried out.
The combination of analyses in the field of Higgs boson discovery is a necessary step to be carefully performed. Indeed, no single decay channel is expected to have enough statistical power for the Higgs boson discovery or exclusion. The combination exercises performed with the outcome of the different CMS Higgs analysis groups in the scope of this thesis paved the way for future Higgs decay channels combinations, having established common statistical approaches, tools and communication channels among the groups.

Chapter 5

Jet Energy Scale Determination using Z Balancing

The only way to measure the kinematical properties of partons, is to cluster all the particles they originate into a single entity, the jet, with the aid of a jet algorithm. Unfortunately, the reconstructed energy of a jet and the one of the originating parton are not equal for several reasons.
Firstly, detector effects like electronic noise, detection thresholds, inactive material and non linear response of the single components affect the measurement. In addition, other aspects like the reconstruction itself must be taken into account, namely the influence of initial and final state radiation and the additional activity originated by underlying event and pile-up which contributes to the alteration of reconstructed jet energies.
Every physics analysis that aims to investigate processes involving the measurements of jet properties must consider jets calibrated with a procedure of jet energy scale (JES) correction, before being able to compare its results to any kind of theoretical prediction.
In the following, the CMS approach to jet energy corrections is discussed. Furthermore, a comprehensive characterisation of a procedure to calibrate the absolute transverse momentum of jets exploiting the $Z(\rightarrow \mu^+\mu^-)$+jet topology is given. Calorimeter and particle flow jets reconstructed with four jet algorithms with different sizes are investigated. The data acquired by CMS during 2010 is considered for a preliminary study of this calibration and the validation of the basic physics objects involved.

5.1 Jet Energy Corrections: Factorised Approach

The transverse momentum of a jet can be estimated with the aid of a reference object. In case of simulated datasets in which Monte Carlo truth information is

present, the reference can be represented by the generator particle jet associated to the reconstructed one. In general, the reference can be an object balancing the jet in transverse momentum, like a Z boson or a photon. An estimator of the deviation of the measured transverse momentum of a jet with respect to its reference value is the single *jet response* defined as

$$R = \frac{p_T^{jet}}{p_T^{ref}}. \tag{5.1}$$

Here the label "jet" denotes the reconstructed jet while "ref" indicates a reference object.

Correction Levels

The CMS collaboration envisages a factorised multi-level jet calibration procedure [114]. Between uncorrected jets and fully calibrated ones, seven *correction levels* are foreseen (see figure 5.1):

1. **Offset:** corrections for pile-up, electronic noise and electronic signals thresholds.

2. **Relative:** corrections for variations in the jet response with pseudorapidity relative to a control region.

3. **Absolute:** corrections of the transverse energy of a jet with respect to a reference in the control region.

4. **EMF:** corrections of the jet response taking into account the energy fraction carried by electrons and photons.

5. **Flavour:** correction to particle jets in dependence of the flavour of the original parton (light quarks, c, b or gluon).

6. **Underlying Event:** correction for the energy excess in a jet due to the underlying event.

7. **Parton:** corrects the jet transverse momentum back to the energy of the parton.

The first three levels are considered as required, while the subsequent corrections are regarded as optional, depending on the analysis. One of the main advantages of a factorised approach is that the corrections of the different levels can be studied, calculated and refined almost independently. Moreover, the investigation of systematic effects can be restricted to the single levels, yielding a better understanding of the different uncertainties.

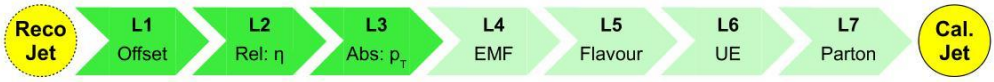

Figure 5.1: Schematic overview of the CMS jet energy scale correction levels. Only the first three levels are compulsory for all analyses.

5.1.1 Offset Corrections

The first step in the chain of factorised corrections is represented by the Level 1 offset correction [115]. Its task is to subtract the energy not associated with the main proton-proton collision. The excess energy to be subtracted out includes contributions from electronic noise in the calorimeter electronics, signal thresholds in the calorimeter towers, extra proton interactions within the same bunch crossing, the *in-time pile-up*, as well as additional energy integrated by the calorimeter read-out electronics from bunch crossings before and after the trigger event, *out-of-time pile-up*.

Methodology

The offset contribution is evaluated as the average energy deposited in the detector inside a cone of radius R as a function of η. This dependency accounts both for the different sorts of electronic noise in the different subdetectors of HCAL and ECAL and for the fact that pile-up activity is higher in the forward regions with respect to the central one. The method is therefore not envisaged for algorithms that produce jets with variable areas like k_T. The strategy for a dedicated jet-by-jet underlying event correction discussed in chapter 6, will be demonstrated to be also valid for the pile-up subtraction in case of every type of infrared and collinear safe algorithm.

5.1.2 Relative Corrections

The detector geometry introduces a pseudorapidity dependence of the reconstructed jets. Level 2 corrections allow to equalise the responses over the whole η range with respect to the response in the central part of the barrel (η <1.3). This particular region is chosen since it is the best part of the calorimeters to calibrate in absolute terms, it provides final states with highest p_T and within its boundaries the response is basically η independent.

Methodology

The Level 2 corrections are based on the principle of transverse momentum conservation. To derive them, a dijet sample is used [116]. Such a dataset can either consist of simulated Monte Carlo events or, in presence of enough recorded collisions, it can be extracted with appropriate selections from real data. The jets transverse momenta would be exactly balanced in presence of a perfect detector.

Indeed, they are are not balanced on average due to the variation of the jet response across the detector (non-uniformity in η). In order to derive the relative correction only events are considered in which one jet is observed in the central barrel region with $\eta < 1.3$ and the other, referred to as *probe jet*, at arbitrary η. The relative jet energy correction is defined as the mapping from the average observed transverse momentum of a jet at some η, to the average observed transverse momentum of the same jet in the barrel. The transverse momentum scale of the dijet event is defined as the average p_T of the two jets:

$$p_T^{di-jet} = \frac{p_T^{barrel} + p_T^{probe}}{2} \tag{5.2}$$

while the scale imbalance is expressed as

$$B = \frac{p_T^{probe} + p_T^{barrel}}{p_T^{di-jet}} \tag{5.3}$$

At this point the relative response can be constructed as

$$r = \frac{2+ < B >}{2- < B >} \tag{5.4}$$

where $< B >$ represents the average imbalance. For the final correction, the relative response is calculated for several η bins and is expressed as:

$$C\left(\eta, < p_T^{probe} >\right) = \frac{1}{r\left(\eta, < p_T^{probe} >\right)} \tag{5.5}$$

where $< p_T^{probe} >$ represents the average transverse momentum of the probe jet in that η bin.

A typical example of η dependence of the jet response is shown in figure 5.2.

5.1.3 Absolute Correction of Transverse Momentum

Once a flat response in η is obtained, the purpose of the Level 3 correction is to remove the residual dependence of the jet response on the transverse momentum of the jet.

Level 3 corrections are obtained through data driven procedures exploiting several physical processes like γ+jet or Z+jet associated production, or, in absence of consistent measured data, from Monte Carlo truth information.

The more integrated luminosity is recorded, the more events are recorded which contain a photon or a Z boson. Such candidates can be used to determine the absolute jet energy scale exploiting the transverse momentum balance of the parton and the vector boson involved in the hard process.

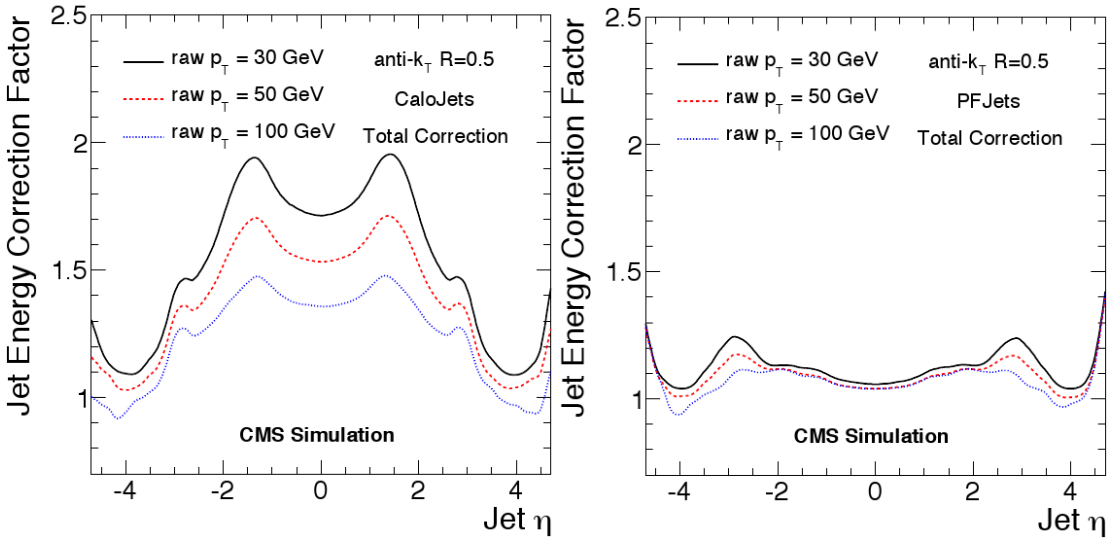

Figure 5.2: A typical trend of the jet response versus the pseudo-rapidity of the probe jet for calorimeter (left) and particle flow (right) jets. The barrel, endcap and forward regions of the CMS detector can be individuated. Taken from [117].

Photon+Jet

Whilst the measurement of the jet transverse momentum involves different subdetectors, the photon measurement can be performed exclusively with the electromagnetic calorimeter. Therefore its energy determination requires only ECAL information [118].
For this reason, the precise measurement of the photon transverse momentum, which can be performed after ECAL calibration (section 2.2.2), is a good estimator of the momentum of the balanced parton. Unfortunately, despite the tight isolation criteria of the photon, this channel suffers from a strong background contribution due to QCD events, especially in the low transverse momentum region.

Z+Jet

A new and promising candidate for jet energy scale determination and calibration are events where a Z boson is balanced by a jet [119]. Not exploited at previous Tevatron experiments because of the reduced number of events available [120], the detailed investigation of this process becomes possible at the LHC. In principle, the decay of the Z boson into a pair of electrons [121] or muons is suitable for such purpose, but the latter has several advantages.
The CMS detector offers an excellent coverage and reconstruction of muons for the precise determination of the kinematics of the Z boson. In addition, only information from the tracker and the muon chambers is considered for the reconstruction, which provides a measurement of the transverse momentum of the balanced Z boson without relying on any calorimetric information. A precise understanding

of both the tracker and the muon system has been achieved at first with cosmic muons during the commissioning phase of CMS and then with first collision data [122]. Furthermore, possible background processes can be suppressed due to the clean signature of this decay, leading to a negligible background. Although the cross section of the $Z(\rightarrow \mu^+\mu^-)$+jet process compared to the γ+jet is nearly one order of magnitude smaller, the precise reconstruction of the Z boson kinematics combined with the negligible background (see 1.8.1) makes this calibration the candidate of choice for the region of low transverse jet momenta [118].

5.1.4 Optional Corrections

Beyond the three required calibration levels, CMS foresees four additional optional corrections.

Electromagnetic Energy Fraction

When treating calorimeter jets, the jet response can be divided into two components, an electromagnetic one, measured by the ECAL, and a hadronic one, measured by the HCAL.
Since the fraction of energy deposited in the electromagnetic calorimeter (EMF) provides useful information for the improvement of the jet resolution, a correction based upon this component is envisaged.
This is an optional correction, which can be developed both from Monte Carlo truth and with a data driven approach which involves the measurement of the jet response as a function of the EMF.

Jet Flavour

The jet response of the detector on jets originating from gluons, u, d and s or b and c quarks is not identical. Due to differences in the jet fragmentation and the presence of neutrinos, the response for c and b jets is smaller with respect to the one for light quarks. Moreover, gluons give rise to more radiation than quarks since they carry more color charge. This behaviour results in broader jets, which are characterised by a smaller response (see section 5.4.4).
The flavour calibration factors proposed by CMS are referred as to Level 5 corrections and can be derived from Monte Carlo truth.

Underlying Event

The purpose of the Level 6 jet energy corrections is to subtract the contribution coming from the extra hadronic activity due to the underlying event from jet energies. Since the underlying event is luminosity independent the plan of CMS is to derive this correction through the analysis of the minimum bias collisions after

the application of the Level 1 offset corrections.
A new strategy for underlying event corrections is discussed in chapter 6. This technique involves the evaluation of the jet transverse momentum per unit area.

Parton

The Level 7 calibration is endowed to the correction of the jets back to the properties of the originating parton. This correction can only be performed starting with a Monte Carlo sample. In addition, it must be underlined that this last level is strongly dependent on the different Monte Carlo generators.

5.2 Absolute Correction of p_T Using Z($\rightarrow \mu^+\mu^-$)+jet Events

In the following the derivation of the Z($\rightarrow \mu^+\mu^-$)+jet calibration is presented, which assumes 200 pb$^-$1 of recorded integrated luminosity. Such an amount of collected data is expected to be reached during early 2011. Calorimeter and particle flow jets are considered for the study in this thesis and JPT jets will be added in the near future. The analysis setup is indeed very flexible, relying only on CMSSW and ROOT standard components, and can accommodate new jet types. Moreover, four jet algorithms with different sizes, including the five CMS official configurations, are treated (see table 5.1).

Table 5.1: The jet algorithms and the size parameters considered in the study of the jet energy scale using the Z($\rightarrow \mu^+\mu^-$)+jet topology.

Algorithm	Resolution	Name
Iterative Cone	0.5	Iterative Cone 0.5
k_T	0.4	k_T 0.4
k_T	0.6	k_T 0.6
anti-k_T	0.5	anti-k_T 0.5
anti-k_T	0.7	anti-k_T 0.7
SISCone	0.5	SISCone 0.5
SISCone	0.7	SISCone 0.7

Particle flow and calorimeter jets

The performance of particle flow and calorimeter jets is different. The new particle flow jets have been already demonstrated [117] to be characterised by a better response. Moreover, their behaviour is almost insensitive to the flavour of the originating parton. Nevertheless, a parallel accurate study of the calorimeter jets is necessary since their reconstruction only depends on calorimetric information.

Therefore, they can act as an irreplaceable partner for the particle flow jets, the reconstruction of which requires information coming from all CMS subdetectors (see section 3.5.3), allowing for mutual cross-checks and optimisation.

5.3 Event Selection and Reconstruction

The following study has been performed considering official CMS Monte Carlo samples, listed in appendix C. The event generator Pythia was employed to generate events where a Z boson decaying into two muons was balanced by a parton in the hard interaction (see figure 1.8).

Z+jet Cross section 7 TeV

To offer a sufficient number of events with large transverse momenta of the Z boson, the generation has been divided into ten bins of transverse momentum, called $\hat{p_T}$ bins. This is possible by the usage of Pythia internal cuts during the event generation, in order to restrict the lower and upper limit of the generated transverse momentum of the hard interaction.

To combine the events from different $\hat{p}_{\mathrm{T}}$ bins, the sub-samples are weighted according to their cross-section (see table 5.2).

Table 5.2: The predicted cross sections for the process $pp \rightarrow \mathrm{Z}(\rightarrow \mu^+\mu^-)$+jet at LHC with a centre of mass energy of $\sqrt{7}$ TeV.

$\hat{p}_{\mathrm{T}}$ bin [GeV]	Cross section [pb]
0 to 20	4,434
20 to 30	145.4
30 to 50	131.8
50 to 80	84.38
80 to 120	32.35
120 to 170	9.981
170 to 230	2.760
230 to 270	0.7241
270 to 300	0.1946
> 300	0.07627

5.3.1 Reconstruction of the Z Boson

Given momentum conservation, the $\mathrm{Z}(\rightarrow \mu^+\mu^-)$+jet topology implies that the kinematics of the balanced parton is linked to the Z boson momentum, which can be reconstructed with high precision. Thus, the calibration of the absolute jet

energy scale can be achieved by comparing the momentum of the boson with the balanced reconstructed jet.

The CMS detector provides an efficient and precise reconstruction of muons up to pseudorapidity values of $|\eta| = 2.4$, which is restricted by the geometry of the muon system. In order not to be influenced by border effects, the pseudorapidity of the muons must hold $|\eta| < 2.3$. To suppress fake candidates, only muons with a transverse momentum larger than 15 GeV are considered for this analysis. In addition, it is required that at least one muon in each event has been identified by the HLT.

To reject background processes, only muons are considered which are isolated from hadronic activity. This is realised by requiring that the sum of the transverse momenta of all tracks $\sum_{\text{track}} p_T$ within a cone of $\Delta R < 0.3$ around the muon is less than 3 GeV.

Finally, the Z boson is reconstructed as follows. Starting with the muons, which pass all selection criteria discussed above, the invariant mass of all possible pairs of muons with opposite charge is calculated. The di-muon system with the invariant mass closest to the Z boson mass [7] is selected and only events with a di-muon invariant mass closer than 20 GeV to the Z boson mass are considered.

5.3.2 Event Selection for Z Boson Balancing

Following the prescriptions of the modular jet energy correction procedure, only events in which the jet with the highest transverse momentum, the *leading jet*, falls in the pseudorapidity region $\eta < 1.3$ are considered. To avoid a bias in the analysis no cuts related to the transverse momentum of the leading jet are applied.

Topology Cuts

In order to exploit momentum conservation to infer the transverse momentum of the jet from the kinematics of the Z boson, it is necessary to restrict the analysis to events where the Z boson is balanced by exactly one jet of the hard process.

Therefore, a clean selection of the events with respect to these properties is required.

To enforce a good balance between the leading jet and the Z boson in the transverse plane, a limit on the back-to-back orientation of the event in azimuthal angle is imposed. In addition, it is required that a potential second leading jet does not exceed a maximal percentage of the bosons transverse momentum.

However, both criteria are correlated and the distribution of the difference of the azimuth of the leading jet and the Z boson versus the fraction of the momentum of the second leading jet compared to the Z boson is drawn in figure 5.3 for calorimeter jet and particle flow jets. The distributions for both jet types have similar shapes. Nevertheless, the reconstructed transverse momentum of the calorimeter jets is smaller than the one of particle flow jets. Moreover, due to

resolution effects, the former is wider. To select a pure sample of events where the Z boson is balanced by exactly one jet, only events which are in the area enclosed by the dashed lines are considered. This corresponds to request the fraction of transverse momentum carried by the second leading jet with respect to the Z boson to be smaller than 20%. In addition, the Z boson is only considered to balance the leading jet if the deviation from their back-to-back orientation in the azimuthal angle holds

$$|\Delta\phi(\mathrm{Z}, \mathrm{leadingJet}) - \pi| < 0.2. \tag{5.6}$$

Events after selections

After the application of the event selection discussed above, a large number of reconstructed Z plus one jet events is still expected for an integrated luminosity of $200\,\mathrm{pb}^{-1}$. Figure 5.4 shows the expected number of events after all selection cuts.

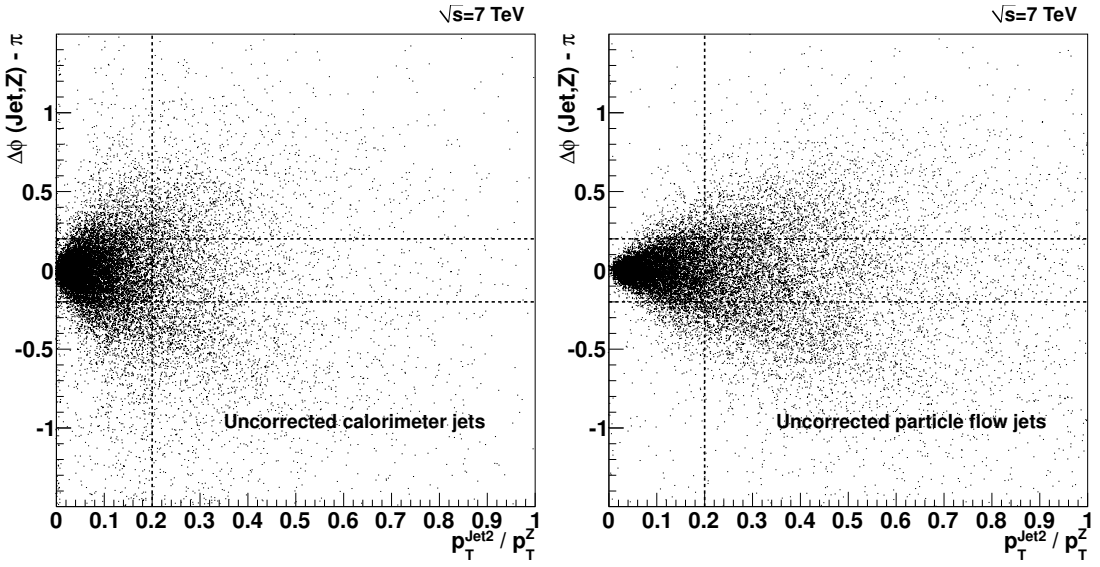

Figure 5.3: Correlation plots of the p_T^{Jet2}/p_T^Z and $\Delta\phi(Jet, Z)$ quantities for the $\mathrm{Z}(\rightarrow \mu^+\mu^-)$+jet events. The dashed lines represent the values of the cuts applied to isolate the topology.

Throughout this chapter, calorimeter jets quantities are displayed on the left and particle flow ones on the right if not stated differently. Especially in the region of interest of lower transverse momentum of the Z boson up to $100\,\mathrm{GeV/c}$, a large number of Z plus one jet events is available and a jet calibration using this process is feasible. In this study, the response was considered reliable only for those bins in which the expected number of events was greater than twenty for an integrated luminosity of $200\,\mathrm{pb}^{-1}$.

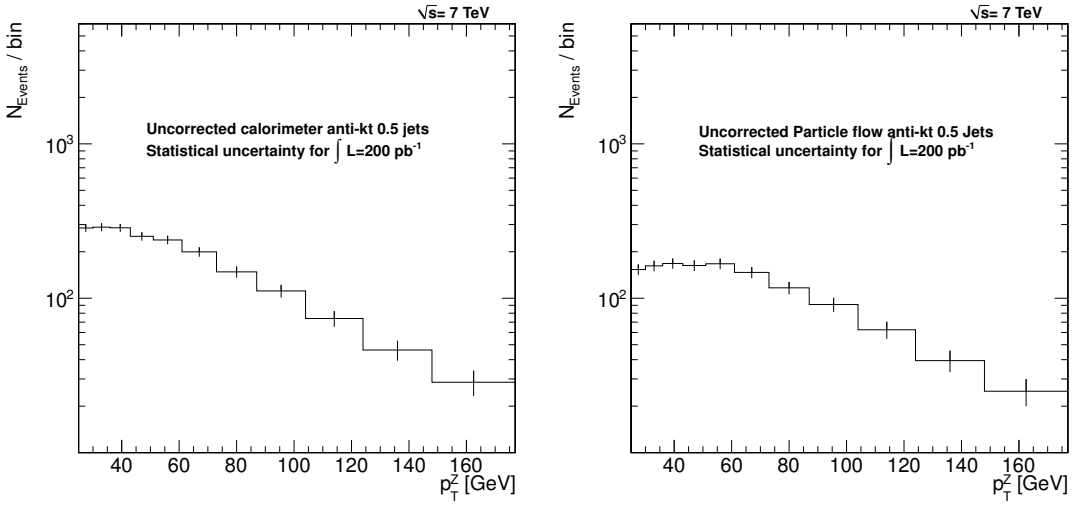

Figure 5.4: Expected number of reconstructed events containing a Z boson which is balanced by one jet for an integrated luminosity of $200\,\mathrm{pb}^{-1}$.

5.3.3 Summary of Reconstruction and Selection

The Z boson reconstruction foresees the following selections:

- Transverse momentum of the muons: $p_T > 15\,\mathrm{GeV}$.
- Pseudorapidity of the muons: $|\eta| < 2.3$.
- Isolated muons: $\sum_{\mathrm{track}} p_T < 3\,\mathrm{GeV}$ within $\Delta R < 0.3$.
- Matching of muons with opposite charge and an invariant mass M of the di-muon system closest to the mass of the Z boson.
- Only events with $|\mathrm{M} - \mathrm{M_Z}| < 20\,\mathrm{GeV}$.

On the other hand, the jet selection criteria were:

- No cut on transverse momentum.
- Pseudorapidity of the jet with the highest transverse momentum: $|\eta| < 1.3$

To isolate the Z boson plus one jet topology, the following requirements have to be fulfilled:

- Ratio of the transverse momentum of the second leading jet in p_T and the Z boson: $p_{\mathrm{T}}^{\mathrm{Jet2}}/p_T^Z < 0.2$.
- Deviation from the back-to-back orientation in the azimuthal angle of Z boson and the jet with the highest transverse momentum: $|\Delta\phi(\mathrm{Z}, \mathrm{leadingJet}) - \pi| < 0.2$.

5.3.4 Backgrounds

To give an estimate on the background contribution to this analysis, Monte Carlo samples of the following processes were investigated:

- $W \rightarrow \mu\nu_\mu$
- $Z \rightarrow \tau\tau$

These samples were filtered using the cuts presented in section 5.3.3 and the results are summarised in table 5.3. The clean signature of the di-muon system joint to the stringent topological constraints imposed by the Z($\rightarrow \mu^+\mu^-$)+jet topology leads to a clean sample for this analysis.
The background to the process Z($\rightarrow \mu^+\mu^-$)+jet was also investigated in the past for 10 TeV centre of mass energy in [119] and found to be negligible.

Table 5.3: Summary of the investigated background processes. The table shows the total number of simulated events as well as the number of events passing the Z plus one jet selection criteria. To compare these events with the signal, the background is scaled to an integrated luminosity of $200\,\mathrm{pb}^{-1}$. The upper limits quoted were calculated assuming a Poisson distribution of the number of background events.

Process	Evt. Available (Weight pb^{-1})	After Sel.	Expected evt. ($200 pb^{-1}$)
$W \rightarrow \mu\nu_\mu$	$1.9 \cdot 10^6$ (245)	0	< 3.7 (95% CL)
$Z \rightarrow \tau\tau$	$1.1 \cdot 10^6$ (708)	0	< 10 (95% CL)

5.4 Measure for the Balancing Quality

The observable, which is chosen as a measure for the quality of the balancing is the single jet response, where the reference is the transverse momentum of the Z boson

$$R(p_\mathrm{T}^\mathrm{Z}) = \frac{p_\mathrm{T}^\mathrm{Jet}}{p_\mathrm{T}^\mathrm{Z}}. \tag{5.7}$$

The estimator of the balancing quality has been chosen to be the mean of the $R(p_\mathrm{T}^\mathrm{Z})$ distribution. Figure 5.5 shows the distribution of $R(p_\mathrm{T}^\mathrm{Z})$ for generator particle jets with a transverse momentum of the Z boson between 25 GeV and 364 GeV. The fact that at generator level the response is centred around one demonstrates that the concept of the Z boson balancing is appropriate.

Response binning

The comparison of the transverse momentum of the jet with the Z boson is

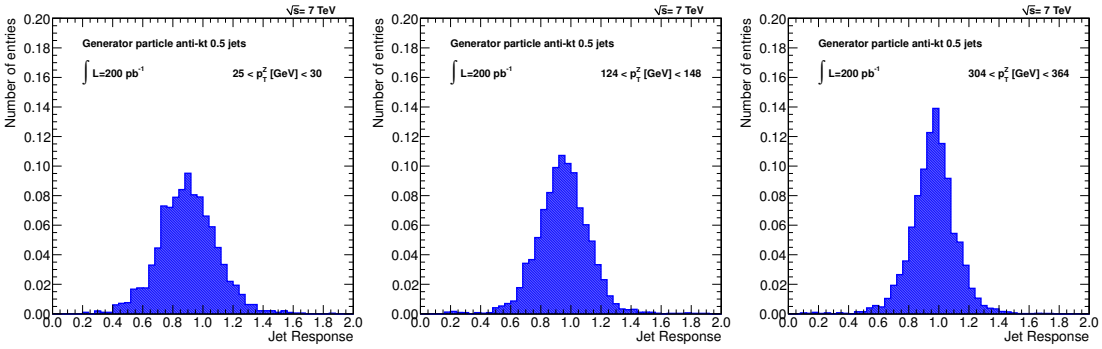

Figure 5.5: Distribution of the response for generator particle jets. Events with a transverse momentum of the Z boson between 25 GeV and 30 GeV (left), 124 GeV and 148 GeV (centre) and 304 GeV and 364 GeV (right) were considered. The values of the response are distributed around unity, which indicates that the concept of Z boson balancing works fine for particle jets.

performed in different bins of Z boson transverse momentum, called "$p_{\mathrm{T}}^{\mathrm{Z}}$ bins". In other words the jet response mean is displayed as a function of the mean of the transverse momentum of the Z boson in the corresponding $p_{\mathrm{T}}^{\mathrm{Z}}$ bin. The bin size is chosen to be variable, reflecting the steeply falling spectrum of the transverse momentum of the Z boson. The first bin covers the range from $p_{\mathrm{T}}^{\mathrm{Z}}$ between 0 and 20 GeV. The upper border of each following bin corresponds to a 20% incremental of the value of the lower border, rounded to the next integer for simplicity.

5.4.1 Particle Jets

In the event topology of this analysis, the momentum of the Z boson is balanced by the parton of the hard subprocess. Due to various effects reproduced by the Monte Carlo generator, namely initial and final state radiation, underlying event (see section 5.1.4), and out of cone effects, the momentum of the parton and the corresponding generator particles jet are not identical. Therefore, even if the kinematics of the Z boson enable a precise determination of the transverse momentum of the balanced particle jet, a difference between the momentum of the Z boson and the balanced particle jet is expected.
To provide a proof-of-concept for the consistency of this calibration method, the balancing of the Z boson and the jet at generator level is investigated. The result of the comparison among algorithms is shown in figure 5.6. In this scenario, the response of an ideal balancing would be equal to unity, and the four jet algorithms and different jet sizes investigated show only a small discrepancy from the perfect balancing. Only for the region of very small transverse momenta the deviation is about 5%. This is related to the presence of additional jets in the event below the

p_T^{jet2} threshold which spoil the ideal topology of a jet recoiling against a Z boson with exactly compensating transverse momenta.

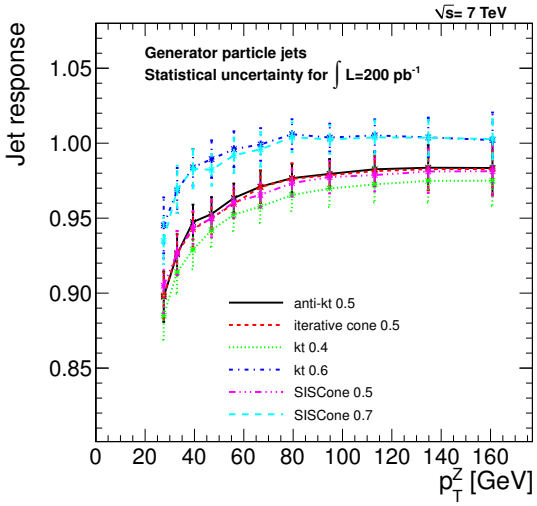

Figure 5.6: Mean of the ratios of the transverse momentum of a particle jet and the balanced Z boson. The error bars indicate the statistical uncertainty on the mean for a number of events corresponding to an integrated luminosity of $200\,\mathrm{pb}^{-1}$. For the region of very small transverse momenta the deviation is less than $\pm 10\%$ and becomes negligible for larger transverse momenta.

Jet algorithms behaviour

A slight difference between the behaviours of the algorithms can be observed which is due to the different jet sizes that are used for the clustering. The smaller the radius, the bigger is the underestimation of the jet transverse momentum. The reason for this effect is the increased probability for out-of-cone effects in case of smaller sizes. The lower set of parameters values, which is 0.5 for the cone-based and 0.4 for the k_T algorithm tend to underestimate the transverse momentum of the particle jet slightly more than for the larger R. More details on the influence of the jet size parameter on the balancing can be found in [123].

5.4.2 Uncalibrated Jets

The situation changes for uncalibrated calorimeter and particle flow jets as shown in figure 5.7. The transverse momentum of the jet is systematically underestimated when compared to the Z boson, especially for calorimeter jets, for which, in the region of small transverse momenta, the reconstructed energy of the jet is about 50% of the corresponding momentum of the Z boson. With increasing p_T, this effect decreases and the energy underestimation is about 25%. On the

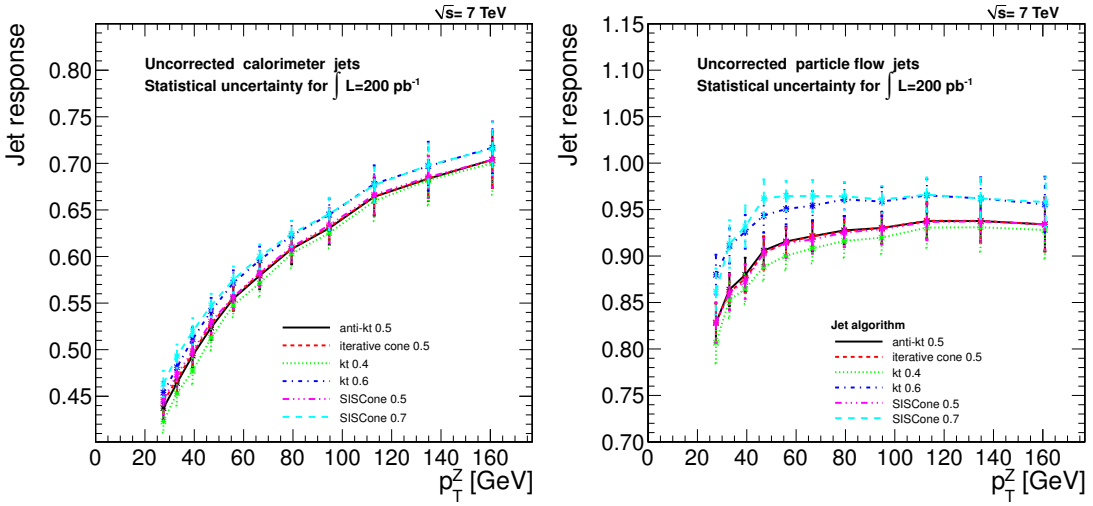

Figure 5.7: Mean of the ratio of the transverse momentum of uncorrected calorimeter and particle flow jets and the balanced Z boson. The error bars indicate the statistical uncertainty on the mean for a number of events corresponding to an integrated luminosity of $200\,\mathrm{pb}^{-1}$. For the region of small transverse momenta, the reconstructed jet energy with respect to the corresponding momentum of the Z boson is 50% and 15% for calorimeter jets and particle flow jets respectively. With increasing p_T, this effect decreases and for the bins of the largest transverse momenta, the energy underestimation is only about 30% and 10%.

other hand, the underestimation for the particle flow jets is of about 20% at low transverse momentum and decreases to 10% for the region of Z boson transverse momentum of about 160 GeV. The underestimation of the transverse momentum for larger jet size parameters is again slightly lower compared to smaller values of R. However, the difference becomes smaller with increasing transverse momentum.

5.4.3 Jets Corrected for the η-Dependence

As discussed in section 5.1.2, the Level 2 relative correction is intended to flatten the detector response as a function of the pseudorapidity with respect to the control region $|\eta| < 1.3$. Therefore in this context, the application of this correction is not expected to sensibly alter the response. This is demonstrated over the whole range of transverse momentum in figure 5.8. These corrected jets are taken as input for the calibration of the transverse momentum of the jets.

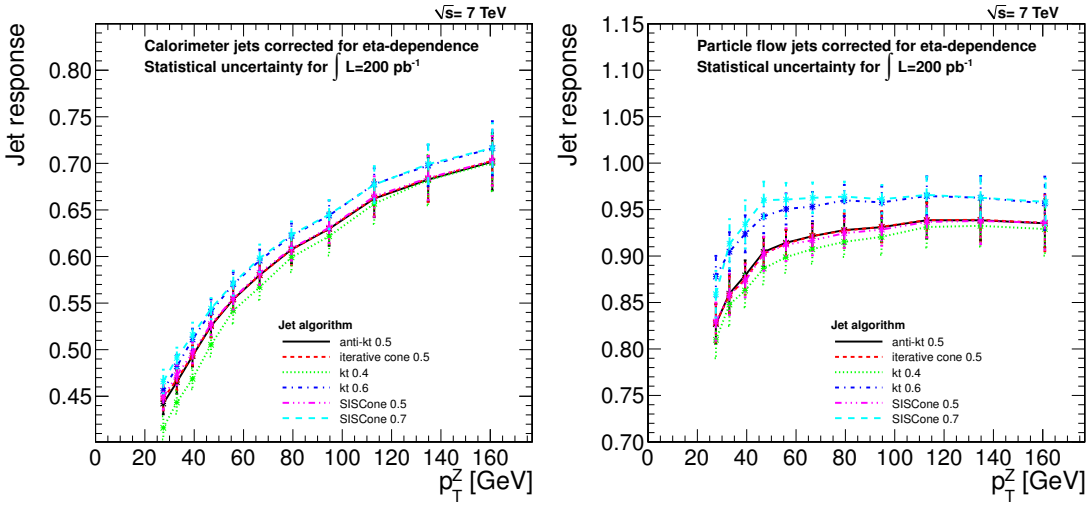

Figure 5.8: Same as figure 5.7 but for jets corrected for the η-dependence. As expected, no significative difference is introduced by the application of Level 2 corrections in the $\eta < 1.3$ region.

5.4.4 Dependence on the Quark-Gluon Fraction

The selection of a process on which a particular calibration is based, influences the fraction of quark and gluon jets present in the available events (see figure 5.9). This has an influence on the calibration because gluon jets develop in slightly wider showers than the ones associated to light quark jets. This leads to more particles not being clustered into the jet by the algorithm depending on the jet size parameter and results in a lower response for gluon jets. In addition, the softer particle spectrum in gluon jets results in many signals under detection threshold determining a lower response for these jets compared to quark jets. In figure 5.9, the difference of the response for quark and gluon initiated calorimeter and particle flow jets corrected for eta dependence is presented for the Z boson. The response of gluon initiated jets is slightly smaller compared to quark initiated for both calorimeter and particle flow jets.

5.4.5 Systematics on the Response

As previously discussed in section 5.3.2, to ensure a clean sample of events in which the Z boson is exactly balanced by one jet of the hard process, only event topologies are considered, which have a second jet with a transverse momentum that is small compared to the Z boson and a well balanced leading jet with respect to the azimuthal angle. The tighter the topology cuts are, the cleaner the sample is, but the smaller is the number of signal events at disposal. Therefore, the final

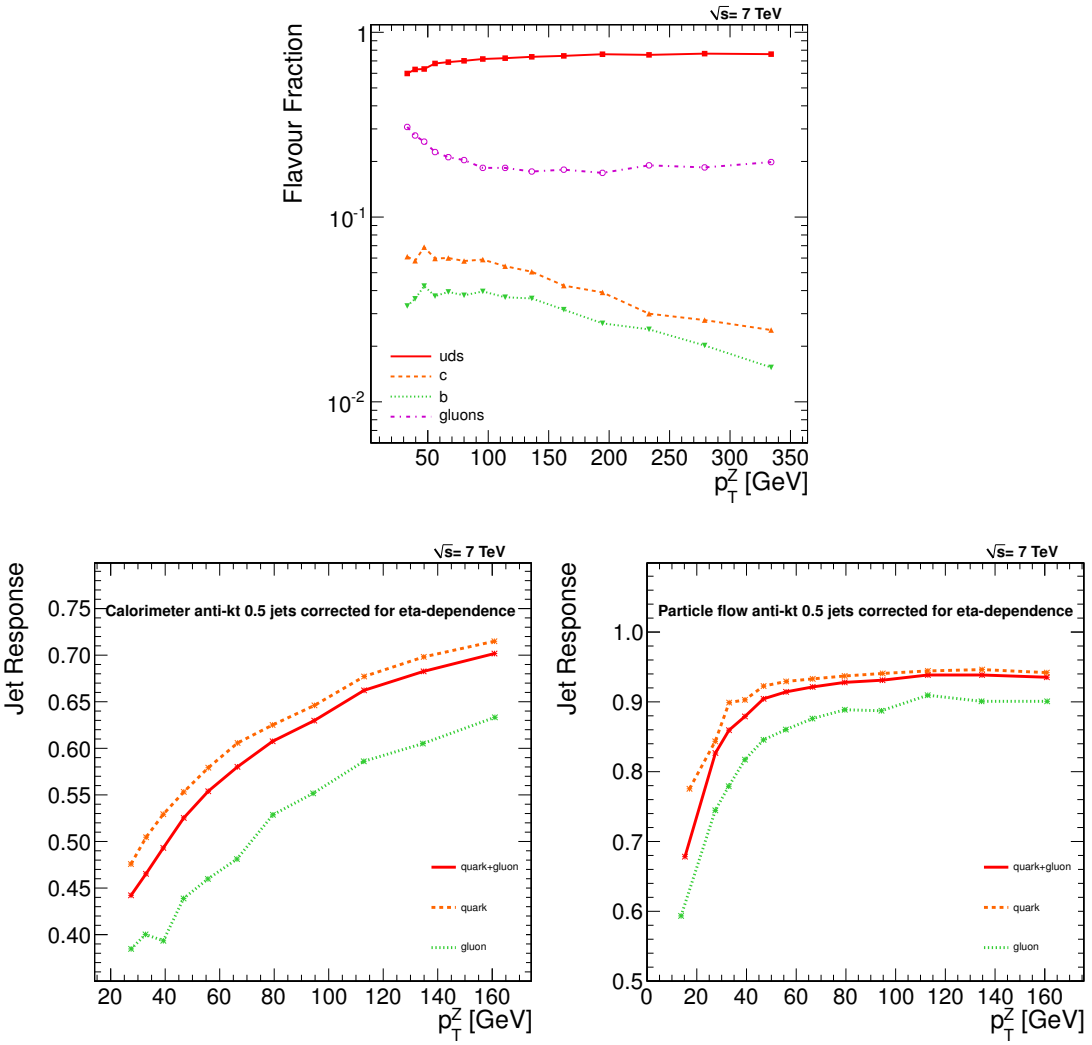

Figure 5.9: Top: Expected flavour composition of the leading jet in the Z($\rightarrow \mu^+\mu^-$)+jet sample at 7 TeV centre of mass energy. Bottom: Response for quark and gluon initiated calorimeter and particle flow jets corrected for η-dependence. The original response curve relative to the proper mixture of quark and gluon jets is shown as reference. The response of gluon initiated jets is slightly smaller compared to quark for both calorimeter and particle flow jets.

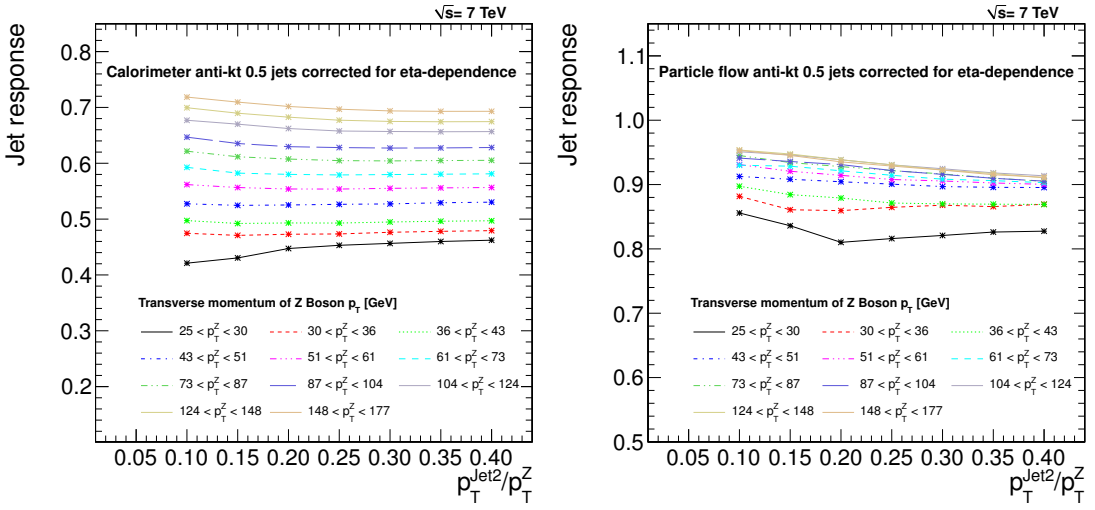

Figure 5.10: Influence of the cut on the second leading jet on the response. All other cuts are fixed to their default values. The different curves represent the variation of the response for the investigated bins of the transverse momentum of the Z boson.

choice of these selection cuts is an optimisation for a clean sample with sufficient statistics, above all, in this initial LHC phase. In the following, the influence of the Z plus one jet selection cuts on the response is discussed. For this, the following cut variations have been applied:

- Fraction of the transverse momentum of the second leading compared to the Z boson:

$$\mathrm{F_{Z,Jet2}} = \frac{p_\mathrm{T}^\mathrm{Jet2}}{p_\mathrm{T}^\mathrm{Z}} < \{0.1, 0.15, 0.2, 0.25, 0.3, 0.35, 0.4\} \tag{5.8}$$

- Deviation from back-to-back orientation in the azimuthal angle of Z boson and the leading jet in p_T:

$$\phi_\mathrm{Z,Jet} = |\Delta\phi(\mathrm{Z}, \mathrm{leadingJet}) - \pi| < \{0.15, 0.2, 0.25, 0.3\} \tag{5.9}$$

Figure 5.10 shows the influence of the variations of the cuts on the second leading jet on the response for all bins of the transverse momentum of the Z boson. All other cuts are fixed to their default values. For all p_T^Z bins, a slight dependence of the response is observable for lower values of the cut on the second leading jet. An extrapolation to $\mathrm{F_{Z,Jet2}} = 0$ may be used to improve the determination of calibration factors.

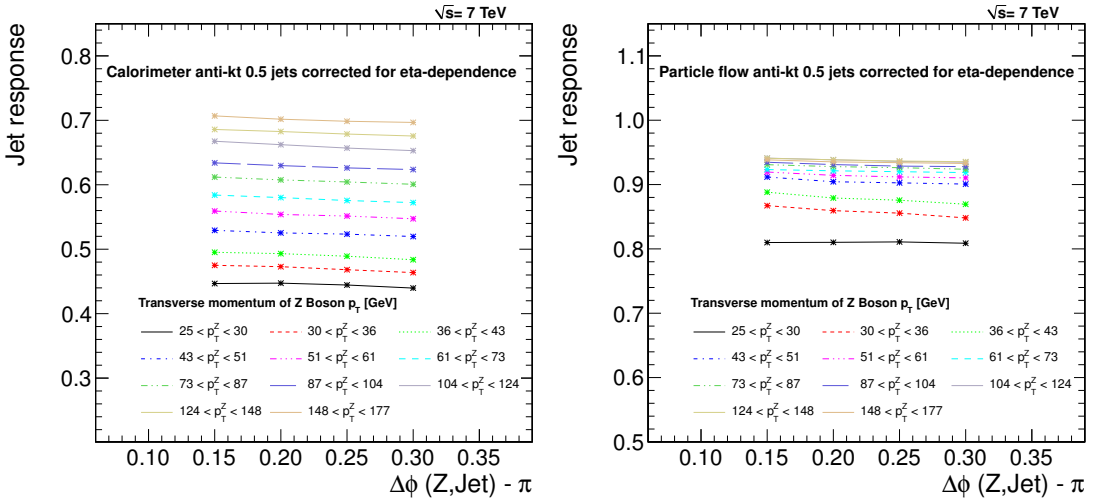

Figure 5.11: Influence of the cut on the deviation from back-to-back orientation in the azimuthal angle of Z boson and the leading jet in p_T on the response. All other cuts are fixed to their default values. The different curves represent the variation of the response for the investigated bins of the transverse momentum of the Z boson.

In figure 5.11 the influence of the cut on the balancing of the leading jet and the Z boson on the response is shown. Again, all other cuts are fixed to their default values and only a slight dependence is visible.

5.5 Calibration Exploiting Balancing

5.5.1 Range of Jet Energy Scale Determination and Luminosity

A study is carried out to understand the expected uncertainty on the jet response and the transverse momentum range in which a jet energy scale determination and correction can be carried out. Figure 5.12 shows the mean of the response versus the mean of the transverse momentum of the Z boson for each bin including the statistical uncertainty for the expected number of events corresponding to an integrated luminosity of $100\,\mathrm{pb}^{-1}$ and $200\,\mathrm{pb}^{-1}$ respectively. Up to 160 GeV, sufficient events per bin are available to deduce calibration factors from Z boson balancing with an integrated luminosity of $200\,\mathrm{pb}^{-1}$. In addition, the statistical uncertainty on the response stays below $\pm 4.5\%$. Even with a reduced number of events corresponding to an integrated luminosity of $100\,\mathrm{pb}^{-1}$, calibration factors from Z boson balancing can be deduced up to a transverse momentum of the Z bo-

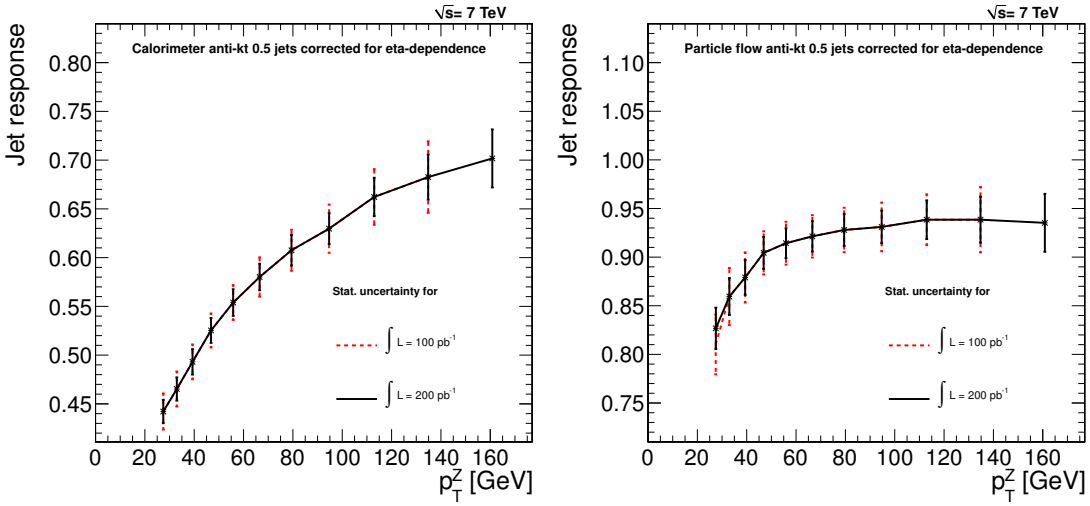

Figure 5.12: Expected uncertainty on the mean jet response for 100 and 200 pb^{-1} in presence of calorimeter jets (left) and particle flow jets (right). The range in which the jet energy scale determination is possible for both calorimeter and particle flow jets reaches 160 GeV and 140 GeV assuming of 200 and 100 pb^{-1} respectively.

son of about 140 GeV. According to the present LHC machine schedule, 100 pb^{-1} of integrated luminosity could already be collected before spring 2011.

5.5.2 Determination of the Correction Factors

Showing the response as a function of the Z boson transverse momentum is the natural choice for the comparisons among different types of jets since the kinematical properties of a di-muon system can be measured very precisely in the CMS detector. Nevertheless, the Level 3 correction factors are functions of the transverse momenta of jets corrected for the η-dependence of the jet response. To determine the functional form of the jet response as a function of jet transverse momentum, for every p_T bin of the Z boson not only the jet response and Z boson p_T histograms are constructed, but also the balancing jet p_T one. Therefore, for every Z boson p_T bin, a one-to-one mapping between the distributions of the boson transverse momentum and the balancing jet one is constructed.

The corresponding distribution is shown in figure 5.13. The error bars represent the expected uncertainty for the luminosity considered.

Correction factors

The correction factors can finally be expressed as the inversion of the response

$$C(p_{\mathrm{T}}^{\mathrm{Jet}}) = \frac{1}{R(p_{\mathrm{T}}^{\mathrm{Jet}})} \tag{5.10}$$

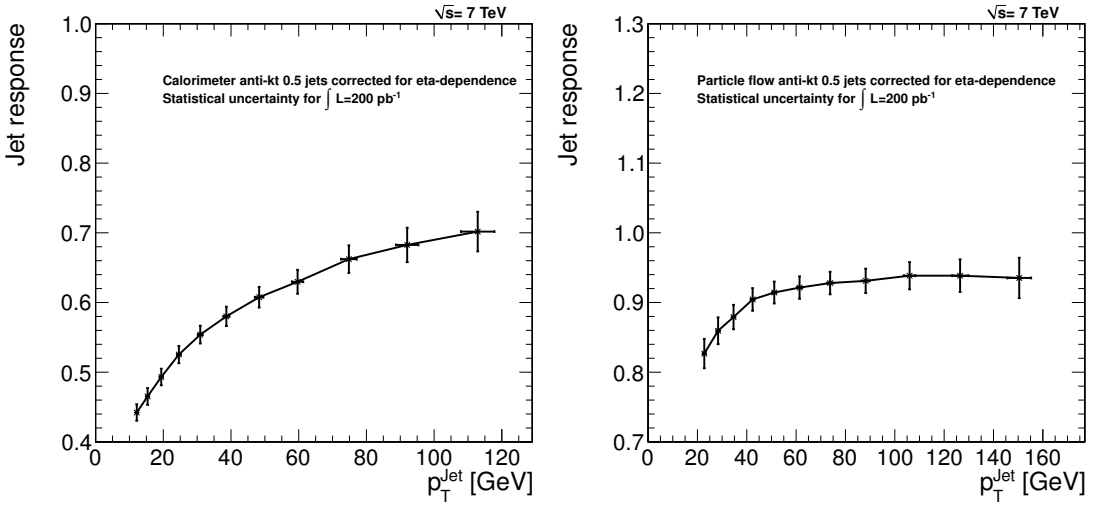

Figure 5.13: Response as a function of the transverse momentum of calorimeter jets, corrected for the η-dependence. The error bars indicate the statistical uncertainty on the mean for a number of events corresponding to an integrated luminosity of $200\,\mathrm{pb}^{-1}$. For the uncertainty on the transverse momentum of the jet, the $RMS/\sqrt{\mathrm{N}}$ of the corresponding distribution is given.

and the uncertainty on this quantity is calculated following the standard error propagation as

$$\Delta C(p_{\mathrm{T}}^{\mathrm{Jet}}) = \frac{\Delta R(p_{\mathrm{T}}^{\mathrm{Jet}})}{R(p_{\mathrm{T}}^{\mathrm{Jet}})^2} \tag{5.11}$$

where $\Delta R(p_{\mathrm{T}}^{\mathrm{Jet}})$ denotes the uncertainty on the response. The trend of the correction factors is fitted, according to the CMS standards, by the function

$$C(p_{\mathrm{T}}^{\mathrm{Jet}}) = a + \frac{b_1}{{p_{\mathrm{T}}^{\mathrm{Jet}}}^{m_1}} + \frac{b_2}{{p_{\mathrm{T}}^{\mathrm{Jet}}}^{m_2}} \tag{5.12}$$

which provides the final correction factors to be applied. Both, the correction factors as well as the fits are shown in figure 5.14 for anti-kt 0.5 calorimeter and particle flow jets.

Closure Test

The validation of the obtained correction functions for all algorithms is performed with a *closure test*. This test is a consistency check which corresponds to the application of the correction factors to the same dataset used for their derivation. The result is presented in figure 5.15. For what concerns calorimeter jets, the deviation from unity is smaller than 1% for the region of $p_{\mathrm{T}}^{\mathrm{Z}} > 50\,\mathrm{GeV}$. The closure for particle flow jets, on the other hand, shows a deviation smaller than

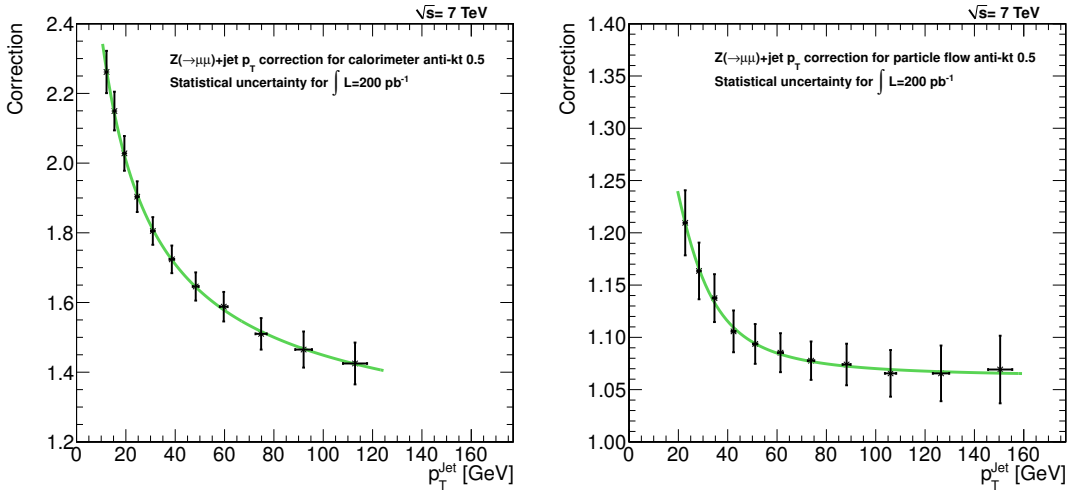

Figure 5.14: Correction factors as a function of the transverse momentum of calorimeter and particle flow jets, corrected for the η-dependence of the response. The algorithm considered is the anti-kt with a size of R=0.5. The error bars represent the statistical uncertainty expected for $200\,\mathrm{pb}^{-1}$ of integrated luminosity.

1% for the whole transverse momentum range investigated. These results demonstrate the successful derivation of the jet energy calibration for all the algorithms investigated.

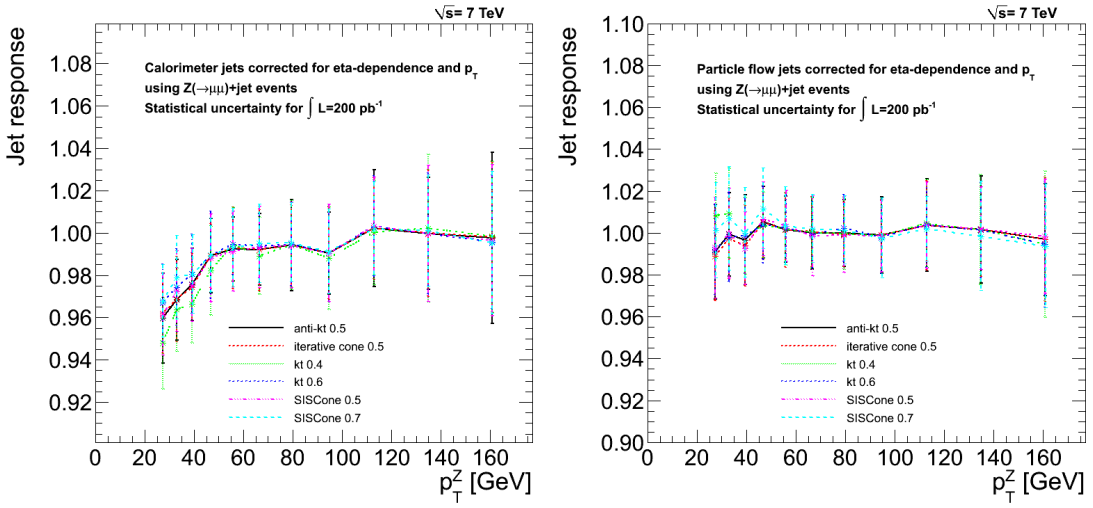

Figure 5.15: Mean of the ratio of the transverse momentum of jets corrected for η-dependence and p_T using the Z boson balancing calibration. The error bars indicate the statistical uncertainty for the expected number of events corresponding to an integrated luminosity of 200 pb^{-1}. The deviation from unity is smaller than 1% for the region of $p_T^Z > 50\,\mathrm{GeV}$ for calorimeter jets and over the whole transverse momentum range for particle flow jets. The behaviour of all algorithms is consistent.

5.6 Jet Energy Scale and Resolution Determination at 7 TeV

Up to now, the LHC did not deliver enough collisions to carry out a complete jet energy calibration (see figure 5.16). Nevertheless, useful checks can be performed

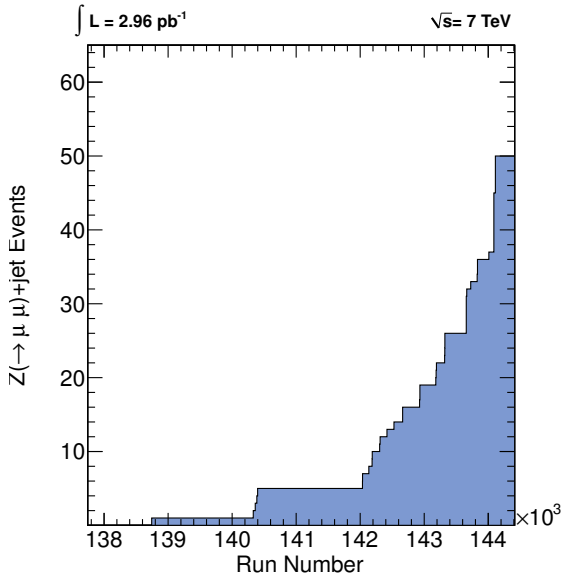

Figure 5.16: Available Z($\rightarrow \mu^+\mu^-$)+jet events as a function of run number until September the 24th. The total collected integrated luminosity is 2.96 pb^{-1} and the available Z($\rightarrow \mu^+\mu^-$)+jet events are 50.

which improve the understanding of the CMS detector through the investigation of the comparison between the simulation and the acquired data.
Moreover, the available Z($\rightarrow \mu^+\mu^-$)+jet events collected can be used to estimate the jet energy scale and jet resolution with respect to the expected values with the aid of the profile likelihood method.

5.6.1 Datasets and Additional Selections

The dataset considered was a subset of the events recorded by CMS in 2010 which contains all the events flagged by the L1 and HLT triggers when at least one muon was detected (for the details see appendix C). Since 2010, such datasets are automatically delivered to the Tier-2s computing centres by CMS and no further action aiming at low level selections is normally required by the analysers.

The anti-k_T 0.5 particle flow jets are chosen as a reference. To exclude objects due to noise in the various detector components, additional quality cuts on these

jets are necessary to treat the detector data. For this purpose, the CMS official recommendations are being used. The loose particle flow *jet-ID* selections, listed in table 5.4, are applied [124].

Table 5.4: The jet-id selections to avoid the presence of fake particle flow jets.

Description	Cut Value
Number of particles in the jet	at least 2
Number of charged particles in the jet	at least 1
Fraction of energy carried by charged hadrons	> 0.0
Fraction of energy carried by neutral hadrons	< 1.0
Fraction of energy carried by electrons and muons	< 1.0
Fraction of energy carried by photons	< 1.0

The effectiveness of the selections was compared between data and Monte Carlo. The simulated sample chosen involved an inclusive $Z \rightarrow \mu\mu$ sample (see appendix C). To equalise the data to this sample, a set of pre-cuts was applied to select events in which at least a di-muon system was present. The muons had to be opposite signs, transverse momentum greater than 15 GeV and $|\eta|$ smaller than 2.3. The effect of cuts is quantified in table 5.5

Table 5.5: The relative effectiveness of the is shown for detector data and Monte Carlo after pre-cuts to equalise the simulated sample and the data. Each efficiency refers to the previous cut. Anti-kt 0.5 jets were considered.

Cut	Data	Monte Carlo
Total events	14,888 (100%)	2,533,960 (100%)
Pre-cuts	1,448 (1.0%)	667,745 (26.4%)
$\lvert\eta^{jet1}\rvert < 1.3$	808 (58.8%)	393,181 (58.9%)
$p_T^Z/p_T^{jet2} < 0.2$	122 (15.1%)	60,141 (15.3%)
$\lvert\Delta\phi(\text{Z,leadingJet})\rvert - \pi < 0.2$	57 (46.7%)	29,138 (48.4%)
$\lvert Z - M_Z\rvert < 20\,\text{GeV}$	50 (87.8%)	26,073 (89.4%)
Total (compared with pre-cuts)	50 (3.3%)	26,073 (3.9%)

5.6.2 Basic object properties

In this section the properties of the basic objects involved in the Z($\rightarrow \mu^+\mu^-$)+jet transverse momentum balancing are investigated after the selection cuts. Figures 5.17, 5.18 and 5.19 show that within the statistical uncertainties, the Monte

Carlo predictions correctly reproduce the acquired data. The Monte Carlo samples used for this study are a slightly more recent version of the one used for the studies described at the beginning of the chapter, see appendix C.

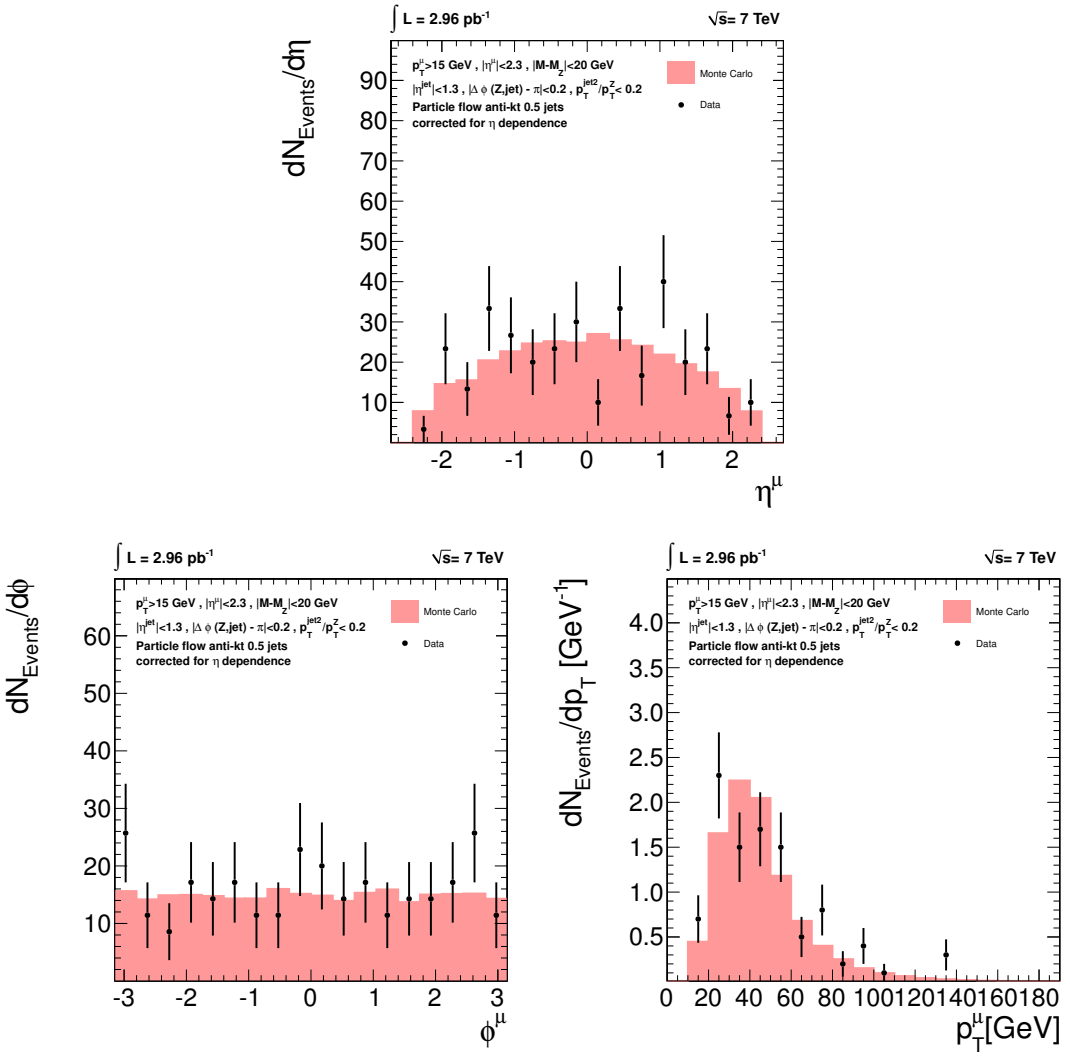

Figure 5.17: The η, ϕ and p_T of the muons after the selections. The Monte Carlo description of data is accurate within the statistical uncertainties.

P_T Balancing

The balancing of the leading jet and the Z boson deserves particular attention and is investigated for every p_T^Z bin. The comparison of the response distributions for Monte Carlo and data (blue ticks) is shown in figure 5.20. The Monte Carlo response distributions are fitted using Gaussian functions:

$$G(A, \mu, \sigma; R) = \frac{A}{\sqrt{2\pi\sigma^2}} e^{-\frac{(R-\mu)^2}{2\sigma^2}} \tag{5.13}$$

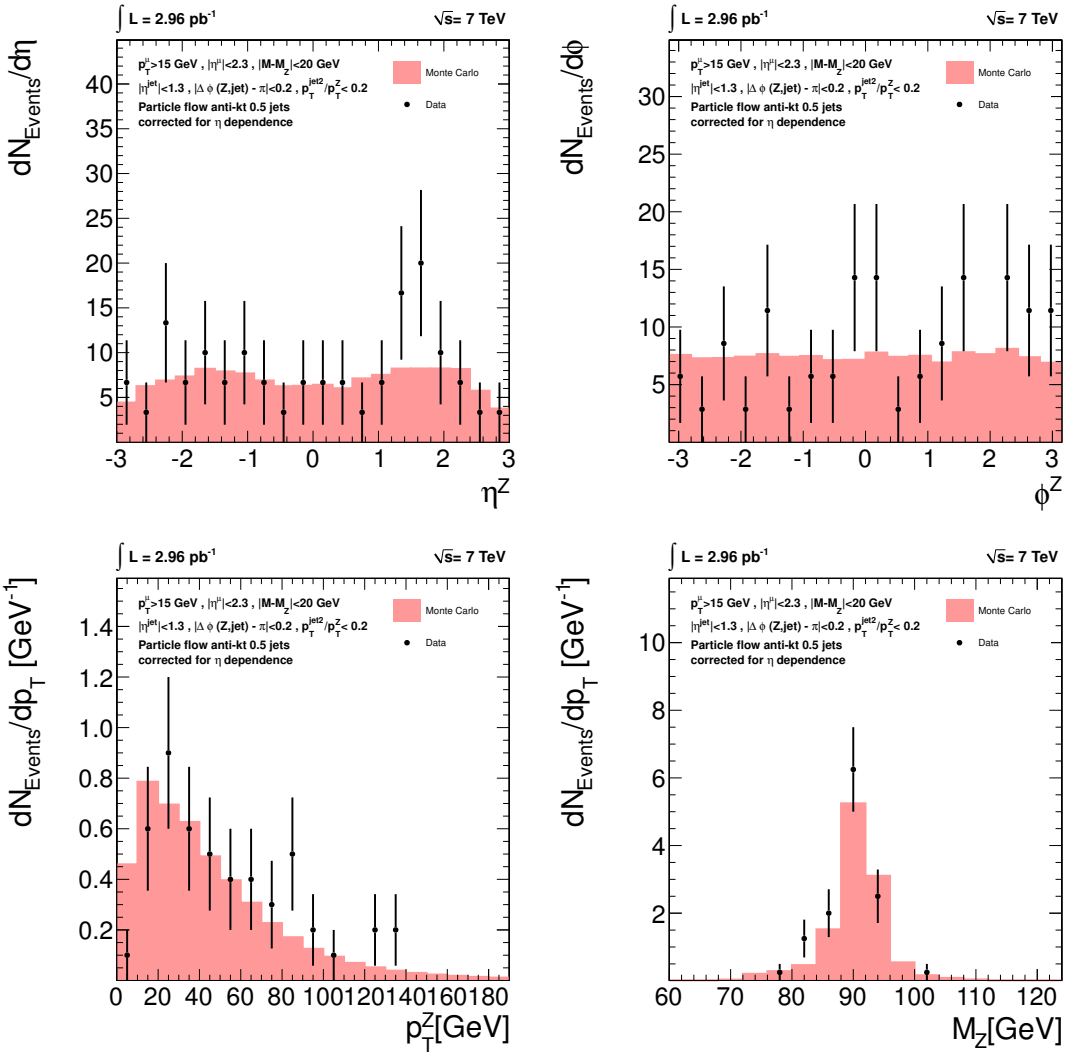

Figure 5.18: The η, ϕ, p_T and mass of the Z boson after the selections.

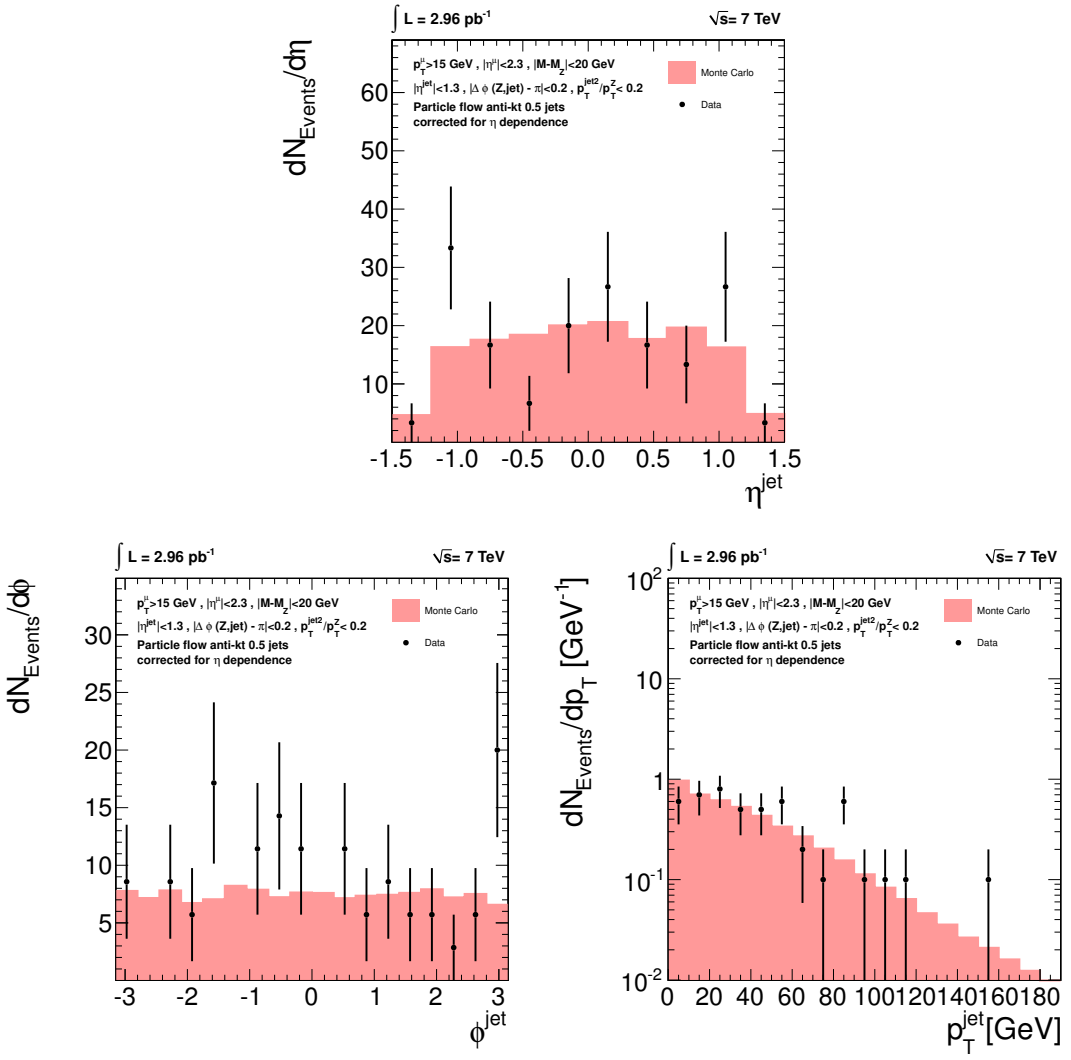

Figure 5.19: The η, ϕ and p_T of the leading jet after the selections.

5.6.3 Jet Energy Scale and Resolution Measurement

Despite the small number of events available, a preliminary quantitative determination of the jet energy scale (figure 5.21) and resolution with respect to the expected Monte Carlo values can be carried out with the profile likelihood method (see section 4.3).

As described in section 5.6.2, for each p_T bin of the Z transverse momentum, the expected response distribution is fitted with a Gaussian function. The mean of the Gaussian is in this context the estimator of the jet response, while the sigma parameter is the estimator of the jet energy resolution. The first step to build a combined likelihood function including the information of all bins consists in a slight modification of the Gaussian parametrisation, namely

$$G(A, \mu, \sigma, s, r; R) = \frac{A}{\sqrt{2\pi\sigma^2 \cdot r^2}} e^{-\frac{(R-s\cdot\mu)^2}{2(\sigma\cdot r)^2}} \tag{5.14}$$

The parameters s and r represent the jet energy scale and jet energy resolution with respect to the Monte Carlo prediction. The values of the parameters A, μ and σ are fixed to the values of the fits, while s and r are unconstrained. For every i-th p_T^Z bin, excluding the very first one, a partial likelihood is built as follows

$$\mathrm{L}_i(s, r) = \prod_{j \in evts\ \ in\ \ bin} G_i(R_j; s, r), \tag{5.15}$$

Therefore, the overall negative log likelihood is built as

$$-\log \mathrm{L}(s, r) = \sum_{i \in bins} -\log \mathrm{L}_i(s, r). \tag{5.16}$$

To obtain the best values of the s and r parameters, the combined negative log-likelihood in equation 5.16 is minimised using Minuit [54], and the asymmetric parameters errors are obtained with the MINOS method. The correlation matrix for the parameters errors is then also calculated. The result of the minimisation is shown as contours in figure 5.22. The values obtained for the two parameters are

$$\begin{aligned} \textit{Jet Energy Scale wrt Monte Carlo} &= 0.93^{+0.04}_{-0.04} \\ \textit{Jet Energy Resolution wrt Monte Carlo} &= 0.98^{+0.12}_{-0.10}. \end{aligned}$$

The correlation coefficient between the errors of the two parameters is zero. Given the small number of events at disposal, the systematic uncertainty due to machine and detector effects is much smaller than the statistical one.

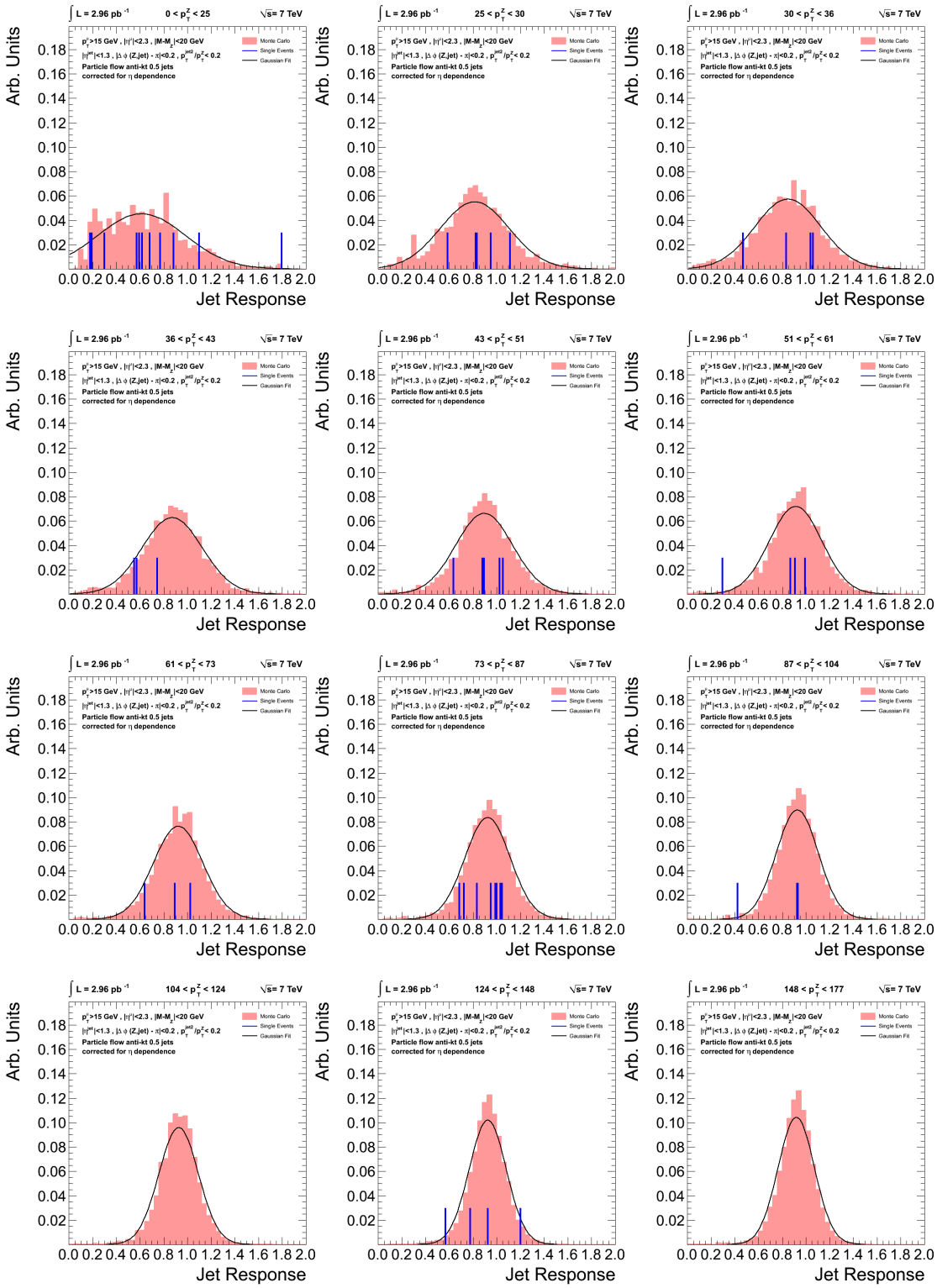

Figure 5.20: The response distribution for every p_T bin in which at least an event was found. Single data events are represented by blue ticks. Except for the first p_T bin, the expected response distribution is well reproduced by a Gaussian function.

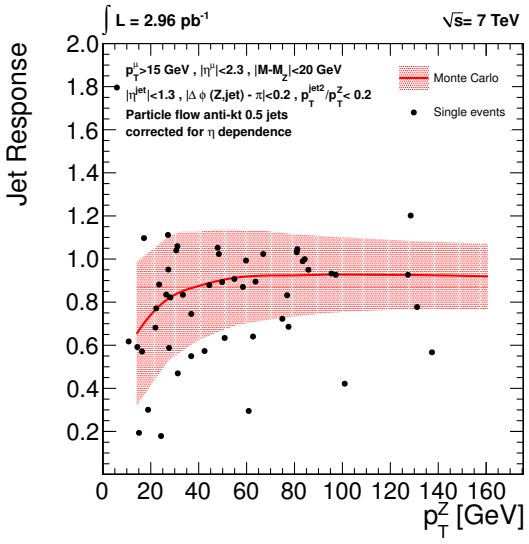

Figure 5.21: The expected jet response as a function of p_T^Z predicted by the Monte Carlo and the jet response of the single Z($\rightarrow \mu^+\mu^-$)+jet events. The band represents the expected 68% confidence interval for the jet response.

The parameter values obtained from the negative log-likelihood minimisation, demonstrate that the Monte Carlo prediction of the jet energy scale does not describe correctly the data, being two standard deviations smaller than the predicted one. Therefore, the Z($\rightarrow \mu^+\mu^-$)+jet data driven method has to be applied to data to derive a reliable jet calibration.

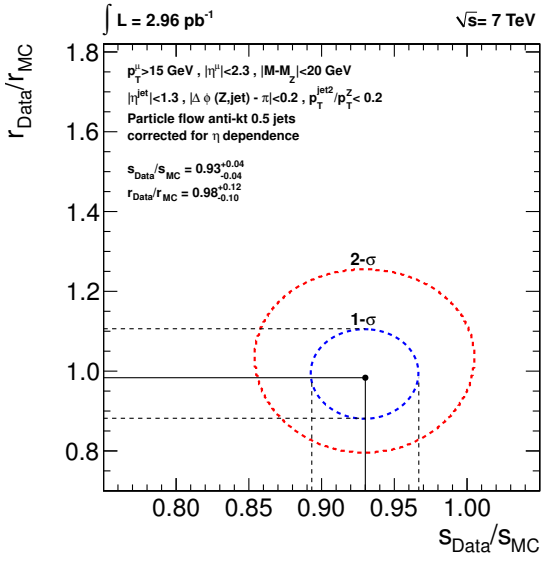

Figure 5.22: Bi-dimensional contours delimiting the for 68% CL and 95% CL regions for the r and s parameters. For the jet energy scale was measured to be $0.93^{+0.04}_{-0.04}$ while the resolution $0.98^{+0.12}_{-0.10}$. While measured resolution is well in agreement with the Monte Carlo prediction, the jet energy scale measured in data is two standard deviations smaller than the predicted one.

5.7 Summary

In this chapter, the factorised approach envisaged for jet energy calibration by the CMS collaboration is described. Three calibration levels are considered as compulsory, the pileup and thresholds subtraction, the removal of the response dependence on η and the correction of the absolute transverse momentum of the jets. This calibration is a fundamental step for all analyses that consider jets in the final state topology. A procedure to derive this last level of correction exploiting the $Z(\rightarrow \mu^+\mu^-)$+jet events is described, and its validity is demonstrated. Two jet types are investigated, calorimeter and particle flow jets, and for each of them seven different jet algorithms are considered. Assuming an integrated luminosity of $200\,\mathrm{pb}^{-1}$ and a centre of mass energy of $\sqrt{s} = 7\,\mathrm{TeV}$, the $Z(\rightarrow \mu^+\mu^-)$+jet calibration strategy could be applied up to a p_T^Z of 160 GeV. The corrected jet response is equal to one within 1% for a p_T^Z greater than 50 GeV for calorimeter jets and for the whole p_T^Z range for particle flow jets.

In addition, a portion of the 2010 data recorded by CMS was analysed to isolate $Z(\rightarrow \mu^+\mu^-)$+jet events. Even though the number of events at disposal is not large enough to allow a calibration, the first measurement of jet energy scale and resolution using the $Z(\rightarrow \mu^+\mu^-)$+jet topology is performed with respect to the Monte Carlo predictions. For the jet energy scale was measured to be $0.93^{+0.04}_{-0.04}$ while the resolution $0.98^{+0.12}_{-0.10}$. While measured resolution is well in agreement with the Monte Carlo prediction, the jet energy scale measured in data is two standard deviations smaller than the predicted one. This proves the need for a data driven calibration in CMS, which can be provided with the $Z(\rightarrow \mu^+\mu^-)$+jet strategy described in this chapter.

Chapter 6

Towards Underlying Event and Pileup Subtraction

As discussed in chapter 5, at a hadron collider, hard events are polluted with many soft particles due to pile-up and underlying event contributions. These particles are an inseparable contribution to the hard scattering in the clustering procedure that leads to the creation of jets.

Correcting for this energy surplus is a necessary step for analyses that involve processes which foresee jets in the final state. Two theoretical publications [125, 126], which consider generator particle jets only, describe a jet area based approach for event-by-event and jet-by-jet underlying event and pile-up subtraction. This approach is suitable for all infrared and collinear safe jet algorithms (see section 3.5.3) and performs the corrections after the jet finding has been carried out, so as to ensure independence of the detector.

This technique is known as the *Jet Area/Median* approach. For the first time, this strategy is investigated with experimental data and compared with full detector simulation, exploiting the collisions recorded by CMS in 2009 [127]. Furthermore, the predictions of several Pythia 6 and Pythia 8 underlying events modellings (see section 3.3.1) are compared to the data. This study allows to improve the understanding of QCD low pt processes in the LHC energy regime.

6.1 The Jet Area/Median method

This method is based on two concepts, namely the measurement of the susceptibility of each jet to diffuse radiation and a parameter free technique to measure the level of contamination due to underlying event and pileup in each event. This contamination will be indicated with the density ρ in the following.

The jet susceptibility to contamination is embodied in its area (see section 3.5.3).

The corrected transverse momentum of a jet will be therefore expressed as

$$p_{T\ jet}^{corr} = p_{T\ jet} - \rho \cdot A_{jet}. \tag{6.1}$$

At high luminosity at LHC, ρ is expected to be of the order of 10-20 GeV per unit area due to pileup [58].

An estimate of ρ must take into account the difficulty of distinguishing the low transverse momentum contribution due to the underlying event and pileup from the large amounts of energy deposited by the final states of the hard scattering. For this reason, a simple mean of the total jet transverse momenta in an event divided by the detector area is not a suitable measure for ρ. Hence, the proposed estimator for ρ is the median of the distribution of the p_T^{jet}/A_{jet} for the ensemble of jets in an event

$$\rho = \underset{j \in \text{jets}}{\text{median}} \left[\left\{ \frac{p_{\text{T}j}}{A_j} \right\} \right] \tag{6.2}$$

as shown schematically in figure 6.1. The usage of the median has the upside of being almost insensible to the contamination of outliers (like hard jets). In addition, unlike other estimators, e.g. truncated mean, no parameters are required for its definition.

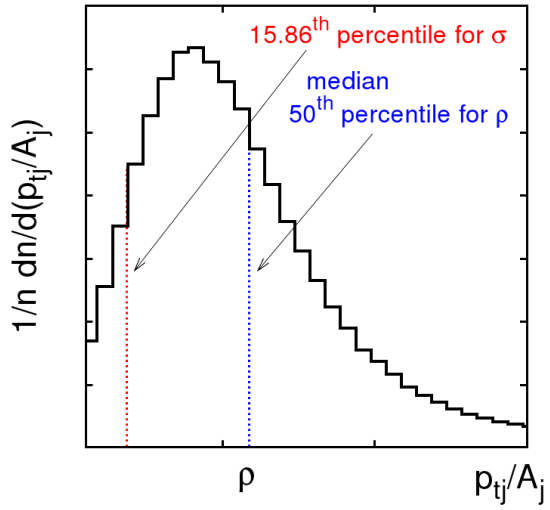

Figure 6.1: Schematic distribution of ρ, with two quantiles shown. One histogram of this type is built per event. Each histogram entry is given by a single jet. Taken from [125].

In the original work describing the Jet Area/Median strategy, generator particle jets are considered. This simplified approach is ideal for a proof of principle study, but unfortunately does not correspond to the environment one has to deal with in a real detector, where effects such as detection thresholds, geometrical constraints,

noise and inefficiencies must be taken into account. To adapt the Jet Area/Median approach to data acquired by CMS, track-jets are chosen. On the one hand, these objects allow to investigate a transverse momentum region much lower than other jet types and do not need a calibration in this low energy regime, given the superior performance of the CMS tracker (section 2.2.1). On the other hand, this choice modifies the original jet area/mean strategy, concentrating the study only on the charged component of underlying event and pileup. The validity of the approach under the effect of this restriction is discussed in section 6.4.

6.2 The Datasets and the Observable

For the study presented in this thesis, the commissioning $\sqrt{s} = 0.9\,\text{TeV}$ data collected by CMS in 2009 were used (see appendix C). Seven different Monte Carlo predictions were compared to data. The generators investigated were Pythia 6, considering six different underlying event tunes (ProQ-20, DW, P0, CW, D6T and Z1), and Pythia 8, considering the default tune. In 2009, the LHC was still in a commissioning phase, and the luminosity reached was relatively low, of the order of $10^{20}\,\text{cm}^{-2}\text{s}^{-1}$. Despite the fact that, with such luminosity, pileup contributions are negligible, an accurate characterisation of the underlying event is performed. In addition, the $\sqrt{s} = 0.9\,\text{TeV}$ events are characterised by a low occupancy (section 3.5.3), i.e. the summed area $\sum_j A_j$ covered by all physical jets divided by the considered detector region A_{tot}, such that large portions of the $\eta - \phi$ plane are covered by jets that are purely made of ghosts.

The ρ' Variable

As it can be inferred from equation 6.2, if the number of ghost jets is larger than the number of physical jets, ρ is zero for the given event. For this reason, the occupancy variable C introduced in section 3.5.3 is used to define the ρ' variable as:

$$\rho' = \underset{j \in \text{physical jets}}{\text{median}} \left[\left\{ \frac{p_{\text{T}j}}{A_j} \right\} \right] \cdot C. \tag{6.3}$$

This modification avoids counting ghost jets as an estimate of the "emptiness" of an event. Nevertheless, at the same time low activity events are mapped to small values of ρ' . To give a concrete example one can consider the one or two jet case: Assuming an average jet area of $A_j \approx 1$ a single entry in the jet p_T results in a median equal to p_T and with an occupancy of $1/(8\pi) \approx 0.04$, ρ' becomes $\approx 0.04 \cdot p_T$. For a two-jet event balanced in p_T one gets similarly $\rho' \approx 0.08 \cdot p_T$ since the median is unchanged but the area occupied by physical jets has been doubled.

The k_T algorithm with a jet size of $R = 0.6$ is chosen as a reference, to avoid biasing the analysis with constant areas like the one produced by the anti-k$_T$ algorithm (figure 3.9). This choice allows not to loose the information about the

geometry of the event, which plays a fundamental role in this analysis.

6.3 Selection and Reconstruction

6.3.1 Event Selection

Triggers The first selection step consists in the evaluation of Level 1 trigger flags to isolate proton-proton collisions from other phenomena that could mimic them, like signal peaks induced by electronic noise in absence of beams or interactions of non colliding beams with residual gas molecules present in the beam pipe [127]. Furthermore, a selection of reliable runs, luminosity blocks and bunch crossings is applied [128]. For more details see appendix C.

Vertices Three more requirements on the vertices [122] of the studied events are specified: In order to suppress fake vertices, the event has to contain exactly one reconstructed vertex. Its position has to be in a 15 cm window along the longitudinal direction, centred around the average of all reconstructed primary vertices in the corresponding run. Finally, there have to be at least three tracks associated to the vertex that have been used to reconstruct this particular vertex.

Detailed efficiencies of the event selection are given in table 6.1.

Table 6.1: Numbers of events satisfying the criteria of the different selection steps together with absolute and relative event fractions.

Selection Criterion	Abs. Event Frac.	Rel. Event Frac.	Total Events
Trigger requirements	100%	100%	453, 409
Good runs selection	56.1%	56.1%	254, 270
1 primary vertex	52.5%	93.7%	238, 248
15 cm vertex z window	52.5%	99.9%	238, 188
≥ 3 tracks fitted to vertex	49.7%	94.7%	225, 447

6.3.2 Track Selection

The strategy for the selection of tracks, follows closely the one presented in [78]. In detail, the following criteria have been applied:

- High purity track quality as defined in [122]
- Transverse momentum $p_T > 0.3\,\mathrm{GeV}$
- Pseudorapidity $|\eta| < 2.3$

- Transverse impact parameter significance $d_{xy}/\sigma_{d_{xy}} < 5$
- Longitudinal impact parameter significance $d_z/\sigma_{d_z} < 5$
- Relative track p_T uncertainty $\sigma_{p_T}/p_T < 5\%$

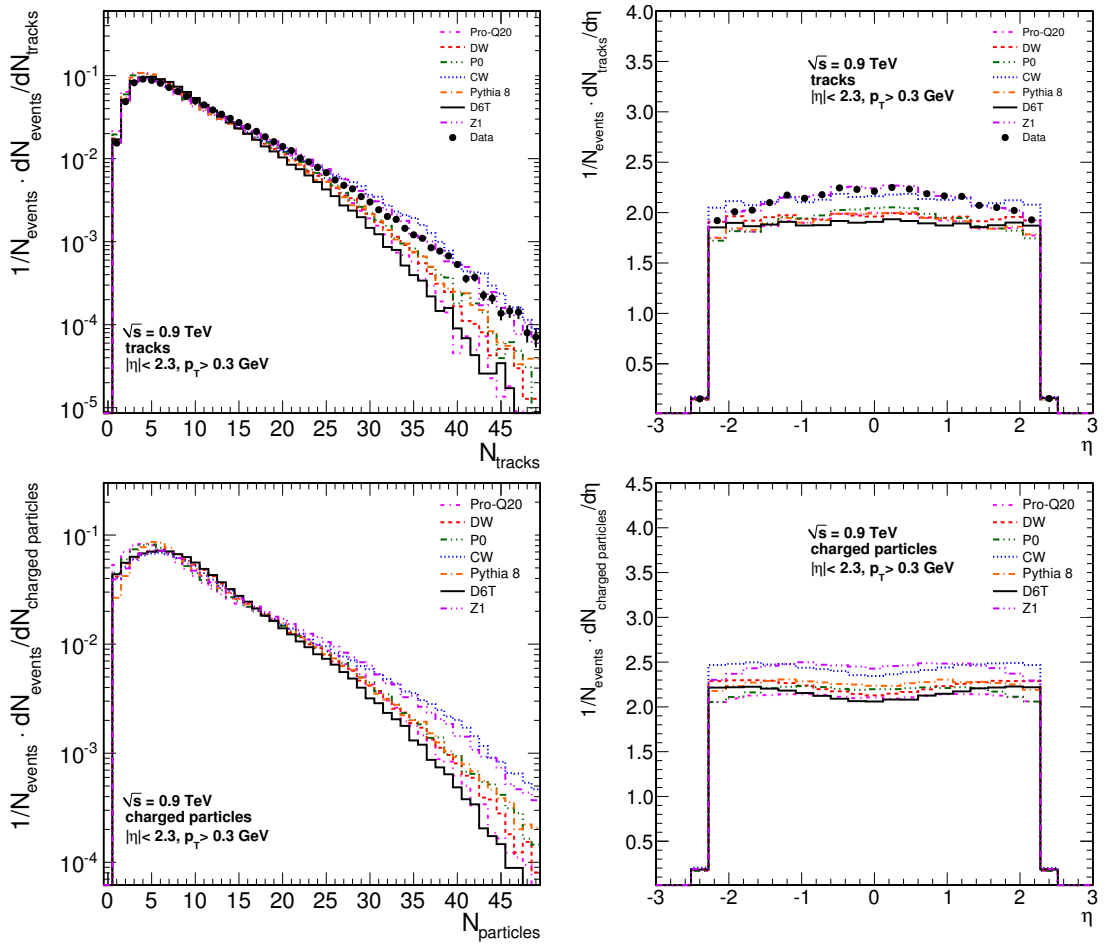

Figure 6.2: Multiplicity normalised to events of reconstructed tracks in data (black circles) and for different Pythia 6 tunes and Pythia 8 default tune. The same quantities (bottom) are shown for charged generator particles.

Figure 6.2 displays the resulting normalised track multiplicity and pseudorapidity distributions for data as well as for the different Monte Carlo predictions. All tunes except Z1 and CW exhibit multiplicities which are too small. The newer tune Z1 describes the data best, as well as the structure of the pseudorapidity distribution.

6.3.3 Charged Generator Particles

To estimate the influence of the detector on a particular observable, it is necessary to compare the prediction as given by a Monte Carlo generator before and after detector simulation including trigger effects (see section 3.4). This procedure also prepares the way for an unfolding procedure to correct the results for detector effects. A generator particle level correspondent to the tracks used in the analysis is a subset of the total ensemble of stable (section 3.3.1) generator particles. The features that a stable generator particle had to satisfy were:

- Charged
- Transverse momentum $p_T > 0.3\,\mathrm{GeV}$
- Pseudorapidity $|\eta| < 2.3$

The transverse momentum minimum threshold and the fact that only charged particles are considered significantly reduces the number of particles entering the clustering process as shown in figure 6.3. Only 60% of all generator particles have

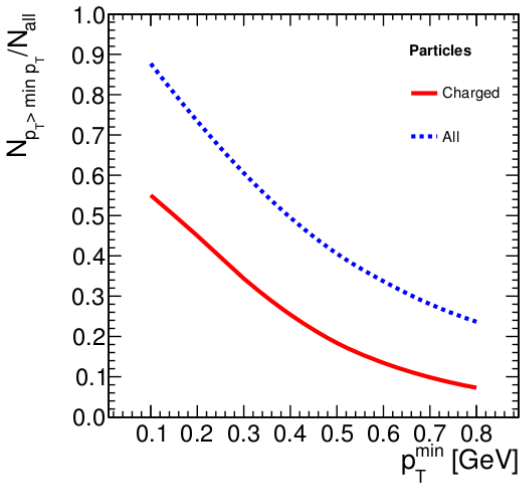

Figure 6.3: Fractions of all and only the charged generator particles in Pythiatune D6T exceeding a minimal p_T.

a transverse momentum greater than 0.3 GeV. Including the charge requirement, only about 35% of all particles remain. The multiplicity and pseudorapidity distributions for the charged generator particles are shown in figure 6.2. They both follow the trends shown by the fully simulated samples.

6.3.4 Jet Definition

The reference jet algorithm used here is the k_T algorithm with a size of $R = 0.6$ (see section 3.5.3). The jets investigated in this analysis foresee two types of inputs: Charged generator particles as defined in section 6.3.3 and tracks which satisfy the selection criteria described above. No further cut on the transverse momenta of the jets is imposed. Due to the selection criteria on the input objects, however, they are implicitly restricted to be larger than 0.3 GeV. To avoid boundary effects due to the tracker geometry in the jet area determination, the absolute pseudorapidity of the jet axis is required to be smaller than 1.8, which has to be compared to $|\eta| < 2.3$ for the clustered objects (see figure 6.4).

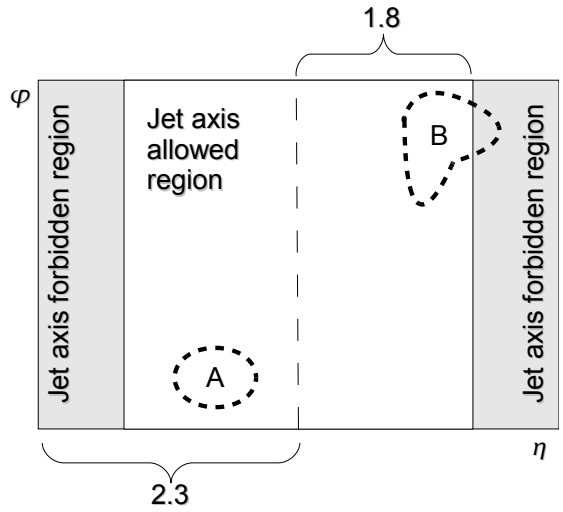

Figure 6.4: Only the jets characterised by an axis with $|\eta|$ <1.8 are considered. This avoids that jets, here indicated by A and B, suffer from boundary effects.

Figure 6.5 shows multiplicity, transverse momentum and jet constituent multiplicity for track-jets for data and Monte Carlo simulation. The best description of these quantities is provided by the Z1 tune of Pythia 6. The jet area distribution and the occupancy are shown at generator charged particle level together with the fully simulated Monte Carlo comparison with data in figure 6.6. Higher average track multiplicities are reflected by higher numbers of track-jets and track-jet constituents, and also by larger occupancies. It is important to observe that the jet area is a purely geometrical quantity: Its distributions are all very similar and reflect the employed jet algorithm and the chosen jet size. The distribution also shows that the areas of k_T jets are not constant, even if the most probable value lies at A=πR^2.

For comparison, the multiplicity, the number of constituents and transverse momentum of charged generator particle jets are shown in figure 6.7. The full

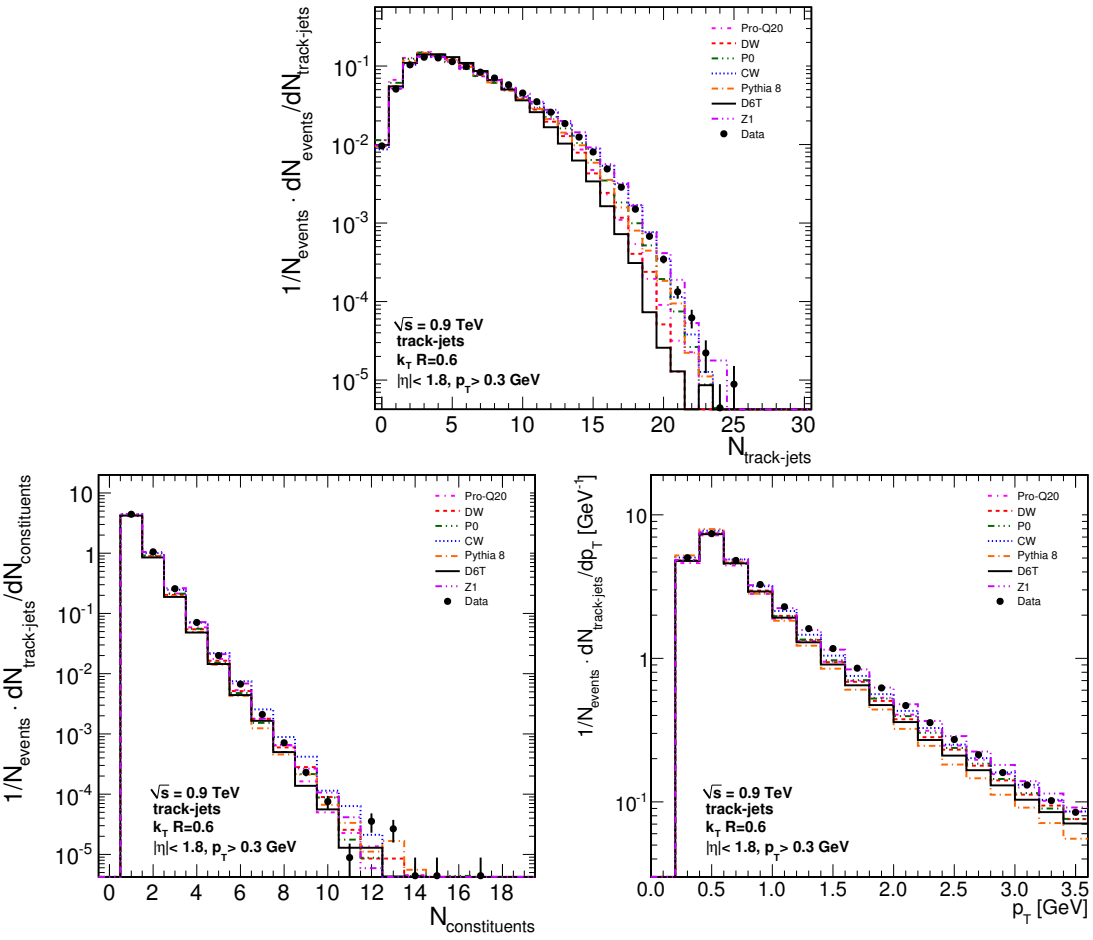

Figure 6.5: Normalised multiplicity (top), jet constituent multiplicity (bottom left) and jet transverse momentum distribution (bottom right) of track-jets in data (black circles) and for different Pythia 6 tunes and Pythia 8 default tune.

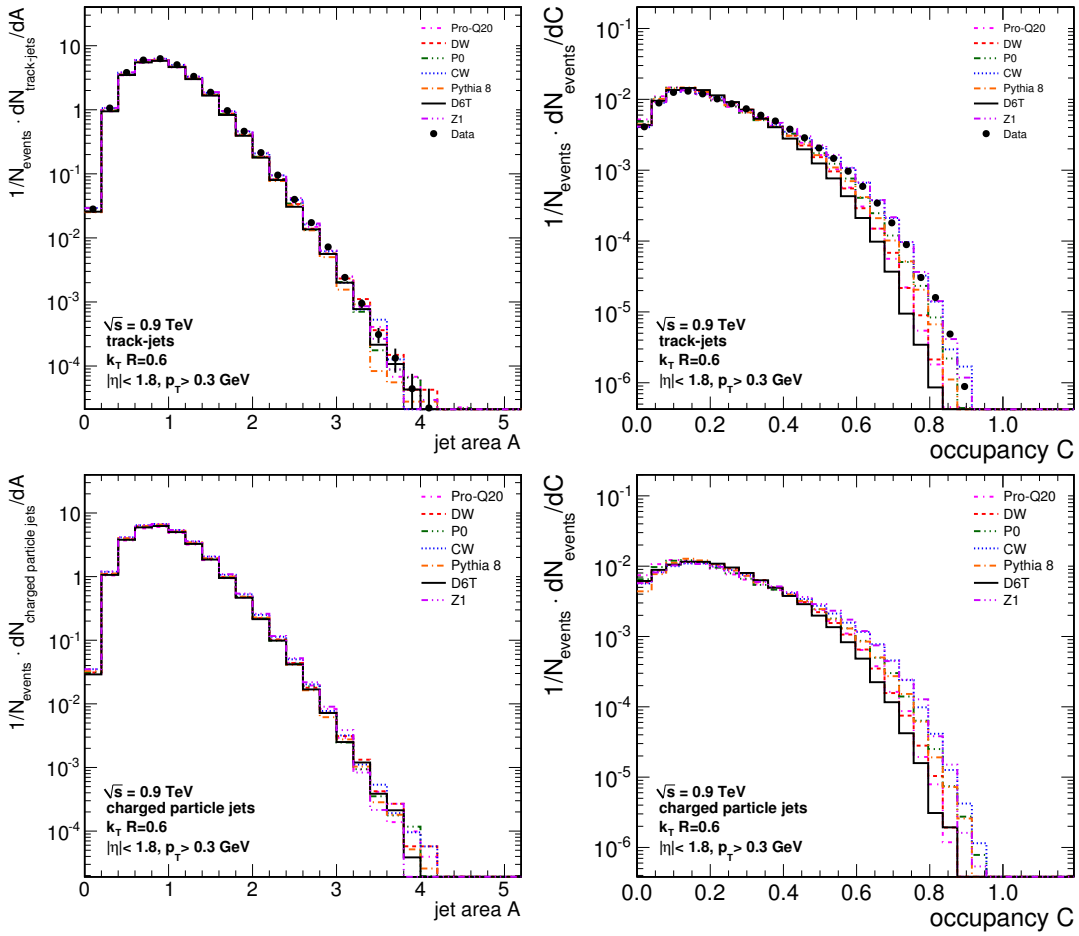

Figure 6.6: Jet area distribution (top left) and area occupancy (top right) of track-jets in data (black circles) normalised to the number of events and for different Pythia 6 tunes and Pythia 8 default tune: higher track multiplicities are reflected by higher numbers of track-jets and also by larger occupancies due to the better area coverage. The same quantities are shown (bottom) at charged generator particle level for comparison.

detector simulation does not alter the hierarchy among the investigated Pythia 6 tunes and Pythia 8 default tune at all.

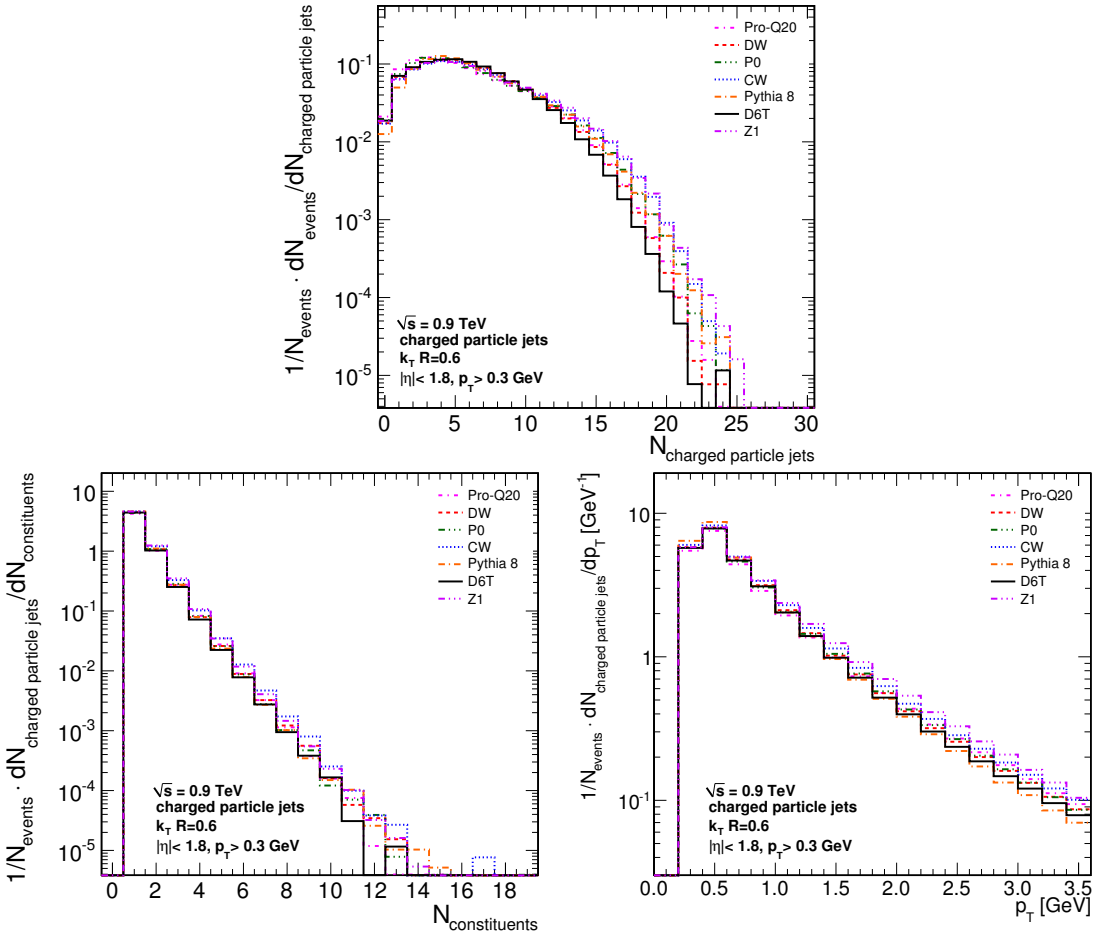

Figure 6.7: Multiplicity (top), jet constituent multiplicity (bottom left) and jet transverse momentum distribution (bottom right) normalised to the number of events for charged generator particles jets for different Pythia 6 tunes and Pythia 8 default tune.

6.4 Sensitivity

The sensitivity of the ρ' variable to the diffuse radiation due to the charged component of the underlying event deserves particular attention since it was not yet described in literature. To perform such an investigation, ρ' calculated with charged generator particle jets is displayed in figure 6.8 left. The right hand side of this figure shows the ρ' distribution relative to the prediction of the Pythia 6 tune Z1. It can be observed that all tunes and Pythia 8 fall below Z1 for $\rho' \geq 0.3\,\text{GeV}$

and only CW overshoot it for values of ρ' above 0.9 GeV. Moreover, the Pythia 8 default tune shows a behaviour similar to D6T. Although the occupied area in $\eta - \phi$ space is much smaller than anticipated in [125] for all particles in dijet events as well as for higher centre-of-mass energies, the adapted variable ρ' is able to differenciate between the diverse tunes and, therefore, different models of the underlying event.

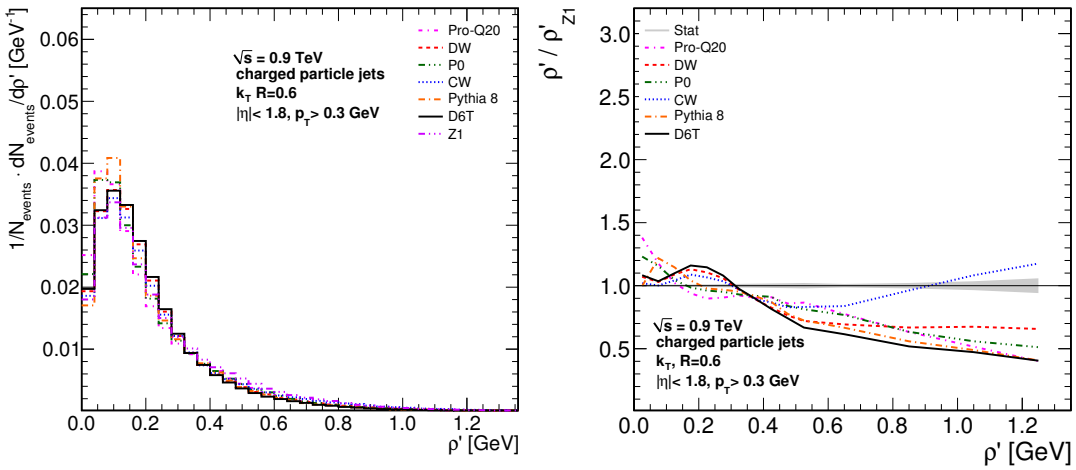

Figure 6.8: Median of jet p_T over area of charged particle jets for the different Pythia 6 tunes and Pythia 8 default tune (left) as well as the ratio of the distributions with respect to the tune Z1 (right). The light-gray shaded band corresponds to the statistical uncertainty calculated with a sample of 300,000 events.

Figure 6.9, shows that the choice of the size 0.6 for the k_T algorithm is safe since it is sufficiently large to avoid the turn-on region of $\langle \rho' \rangle (R)$. This size is officially chosen by the CMS collaboration for the standard reconstruction chain and, therefore, can safely act as standard choice for this analysis.

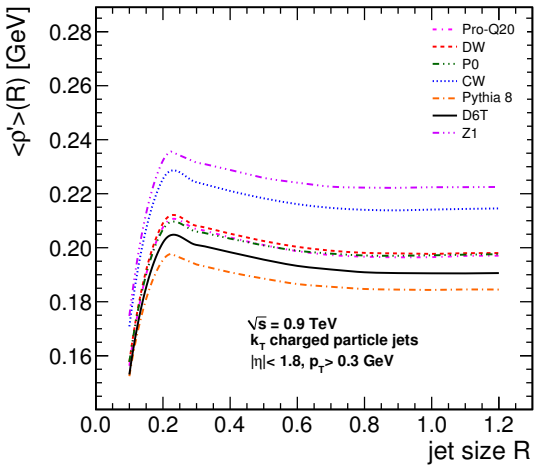

Figure 6.9: Dependence of the mean of the ρ' distribution for the different tunes on the jet size R with charged particle jets.

6.5 Systematic Uncertainties

The systematic uncertainties affecting ρ' were studied in detail to solidify the new Jet Area/Median technique. The following sources of systematic uncertainties in the ratio of fully simulated Monte Carlo predictions divided by the measurement of ρ' are considered:

a) Tracker material budget

b) Tracker alignment

c) Tracker map of non-operational channels

d) Vertex reconstruction

e) Track reconstruction efficiency and fake rate

f) Track selection variations

g) Transverse momentum resolution of track-jets

h) Track-jet response

i) Trigger efficiency bias

Following CMS guidelines, two categories of sources were individuated, namely the ρ' dependent and, within statistical accuracy, ρ' independent effects. The size of each constant effect and its statistical uncertainty are estimated individually. If the systematic uncertainty is compatible with zero within precision, the absolute

values of the individual offset and the statistical uncertainty are added to give a conservative estimate. For the ρ' dependent effects a parameterisation of the functional form is estimated in a fit. The results of this fit represent the quoted uncertainty bin-by-bin. In case of upwards and downwards variations around a central value (sources *a*, *d*, *e*, *f* and *h* of systematic uncertainty), the average of the absolute deviations is taken as the uncertainty estimate. All systematic uncertainties are finally added in quadrature for every bin of the ρ' distribution. The complete set of intermediate figures relevant for the estimation of every systematic effect considered in this study is located in appendix D.

Tracker

The knowledge of the tracker material budget, see section 2.2.1, is precise to about 5%. A comparison of simulations with accordingly varied material shows no significant influence on ρ' . Similarly, two different tracker alignment scenarios, a perfect one and the one reflecting the actual knowledge of the device, are investigated. The impact on the observable is proved to be negligible. Finally, the map of non-operational tracker channels, which varies from run to run, is estimated to yield an effect of the order of 2%.

Vertices

The influence of the vertex reconstruction efficiency is tested using two settings of the minimal separation in z between primary vertex candidates being different from the nominal value. As expected from the very low probability of additional collisions in the same event for the considered data, no large effect on ρ' is visible.

Tracking Efficiency Fake Rate

A detailed understanding of the tracking efficiency and the fake rates is required for this analysis. Any difference between Monte Carlo simulation and data affects the average track multiplicity in the events. According to [129], the uncertainty on the tracking efficiency can be estimated conservatively to be 2% and the uncertainty on the fake rate to be about 0.5%. To reflect this uncertainty, the track content in the events is varied by $\pm 2\%$ by either rejecting tracks in an event with 2% probability or by adding with the same probability an additional track drawn from a pool of separately simulated events. The described procedure yields an effect of the order of at most 6%.

Track Selection Variation

A possible sensitivity of the measurement to the track selection is investigated by varying the corresponding criteria as listed in table 6.2. The only significant effect found is a sensitivity of the order of 6% to the minimal allowed track momentum.

Track-jet Resolution

The finite resolution in transverse momentum of track-jets is also considered. This effect is quantified by artificially smearing the p_T of the track-jets with a Gaussian distribution of 5% width, well above expectations from simulation to obtain a conservative estimate of the systematic uncertainty. No effect on the ρ'

distribution becomes evident.

Track-jets Response

Moreover, the effect of a possible shift of the track-jet response is investigated. An artificial increase or decrease of the p_T of the track-jets compatible with the single track p_T resolution is introduced. The overall effect on ρ' is found to be limited to about 6%.

Trigger Bias

Lastly, the possible bias introduced by differences in trigger efficiencies between the detector and the simulation is studied. The size of this effect is at most 3% over the whole ρ' range.

All investigated systematic effects as well as the applied size estimation method are summarised in table 6.2. The total systematic uncertainty is shown as the dark-grey band in figure 6.11.

Table 6.2: Summary of systematic uncertainties and the applied size estimation method. The first half lists the sources considered to be ρ' independent. The second half represents the ρ' dependent effects where the quoted uncertainty is taken at $\rho' \approx 1.2$ corresponding to the maximally possible deviation.

Systematic Effect	Size
Constant value independent of ρ'	
Tracker material budget: $\pm 5\%$	0.2%
Minimal z separation between multiple vertices: (10 ± 5) cm	0.5%
Maximal track $\lvert\eta\rvert$: 2.3 ± 0.2	0.5%
Significances of track impact parameters: $(5 \pm 1)\sigma$	0.5%
Maximal track p_T uncertainty σ_{p_T}/p_T: $(5 \pm 2)\%$	0.4%
Track-jet p_T resolution: 5%	0.5%
Derived bin-by-bin in ρ' from fit	
Tracker alignment	0.6%
Tracker map of non-operational channels	2.3%
Data - MC track efficiency & fake rate mismatch: $\pm 2\%$	6.0%
Minimal track p_T: (300 ± 30) Mev	5.8%
Track-jet response shift: $\pm 1.7\%$	5.6%
Trigger efficiency bias	3.1%

6.6 Results

The redefined observable ρ' has been demonstrated to be sensitive to the various underlying event tunes when applied to charged particle jets. Hence, the focus is

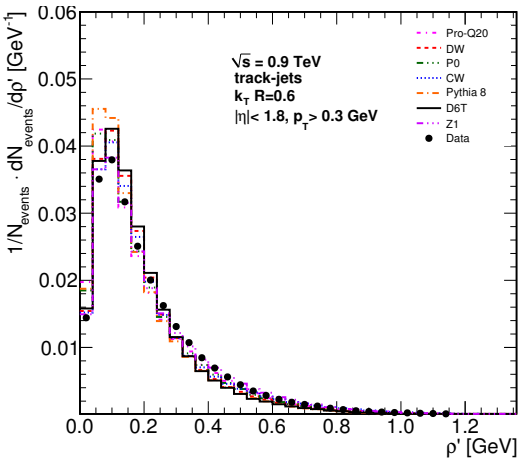

Figure 6.10: Median of ρ' reconstructed from collision data (black circles) and for the different Pythia 6 tunes and Pythia 8 default tune. The distributions are normalised to the number of events.

now on the comparison of generated events after full detector simulation including trigger effects to the CMS collision data.

In figure 6.10, the ρ' distribution for track-jets is presented for data in comparison to the different Pythia 6 tunes and Pythia 8 default tune. All distributions are normalised to the number of events. The curves for the different tunes exhibit a similar behaviour as for charged particle jets.

The ratio of the ρ' distributions with respect to the Pythia 6 Z1 tune is shown in figure 6.11 left to give emphasis to the differences among the curves. Tune DW and CW exhibit the smallest differences when compared to Z1. In comparison to data as presented in figure 6.11 right, none of the Monte Carlo predictions works satisfactorily, although tune Z1 comes closest to the data.

As demonstrated, the Jet Area/Median approach remains sensitive to the underlying event activity even considering charged particles only and at a centre-of-mass energy of 0.9 TeV, which was not foreseen in the original proposition.

With the aid of Monte Carlo, the percentage of energy carried by the neutral components of jets can be derived [91]. Therefore, it is possible to deduce the form of ρ' as if it was calculated also keeping into account the neutrals participating in the underlying event, making a correction possible. In addition to that, the procedure established can be applied to particle flow and calorimeter jets with 7 TeV data, where the occupancy is larger than the one measured at 0.9 TeV.

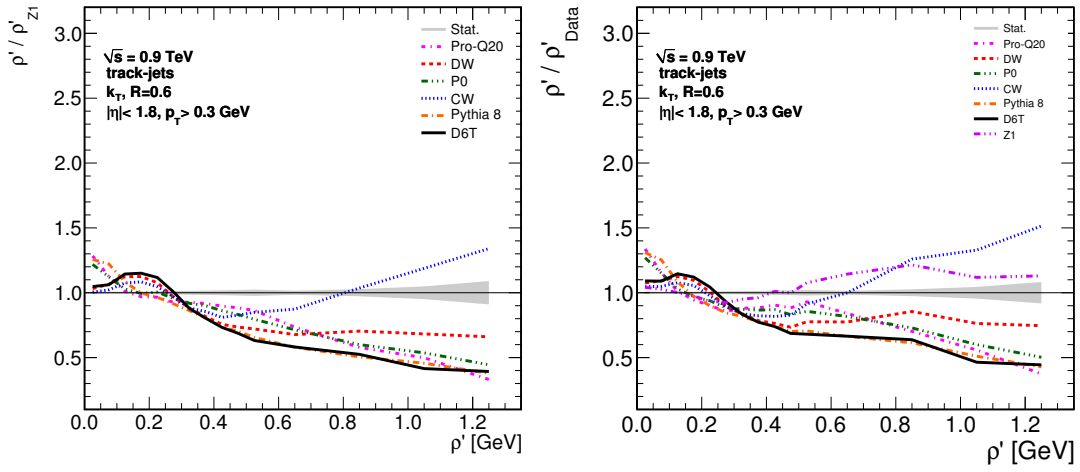

Figure 6.11: Median ρ' for the different Pythia 6 tunes and Pythia 8 default tune relative to Pythia 6 tune Z1 (left) and for all Pythia 6 tunes and Pythia 8 default tune with respect to data (right). The dark-gray shaded band corresponds to the systematic uncertainty and the light-gray shaded band to the total uncertainty. For the comparison to tune Z1 on the left systematic uncertainties are not considered.

6.7 Summary

Recent theory publications [125, 126] proposed the new Jet Area/Median technique for the underlying event and pileup subtraction, which acts on an event-by-event and jet-by-jet basis, exploiting the median ρ of the $p_T^{jet}/Area^{Jet}$ quantity in each event. The first application of this approach to real data is discussed. For this purpose, track-jets were used and collisions at a centre of mass energy of 0.9 TeV recorded by CMS in 2009 are exploited. Since no pileup events are present in the sample, only the underlying event contribution can be studied. To take into account the low occupancy of the events, the ρ' variable is introduced. The different underlying event models of the Monte Carlo can still be distinguished and compared with data. The best description is provided by the Pythia 6 tune Z1. The study provided a deeper understanding of the usage of the jet area quantity with detector data. Moreover, it paved the way for a new technique for the jet energy corrections for extra activity due to pileup and underlying event, which can be applied not only to jets made by tracks but also to other jet types, like calorimeter and particle flow.

Chapter 7

Conclusions and Outlook

A necessary step towards a Higgs discovery is the combination of several decay channels and an accurate statistical treatment of data. Within the scope of this thesis, a framework for analyses modelling, combination and statistical studies, RooStatsCms, was developed and used for the first Higgs analyses combinations to obtain expected exclusion limits and significances for different mass hypotheses. This exercise paves the way for all future combinations in terms of an infrastructure made of well established statistical methods, and guidelines for systematic uncertainty treatment. RooStatsCms is now part of the RooStats component of ROOT, the most frequently used tool for data analysis in high energy physics.

A precise knowledge of the energy of jets is necessary for all LHC analyses dealing with QCD processes either as background or as signal. This can be achieved only with an accurate jet energy calibration procedure.
The calibration proposed in this work focuses on an absolute jet energy scale correction exploiting events where a boosted Z boson decaying into two muons is balanced in transverse momentum by a jet. For the first time, this technique can be applied at a hadron collider given the copious production of Z bosons at the LHC. The estimator chosen for the quality of the balancing was the jet response, defined as the ratio of the transverse momentum of the jet and of the Z boson. The prospects for this calibration technique are investigated for a centre of mass energy of 7 TeV using Monte Carlo simulation. Two types of jets were investigated: Calorimeter jets, composed exclusively of calorimetric energy deposits, and particle flow jets, which involve information coming from all CMS subdetectors. The background for the $Z(\rightarrow \mu^+\mu^-)$+jet events was proved to be negligible. The different behaviour of the response in presence of quark or gluon initiated jets was also studied, and the flavour composition of the jets in the sample was investigated, showing a predominant presence of jets originated by u, d and s quarks. Assuming $200\,\mathrm{pb}^{-1}$ of integrated luminosity, a realistic estimate for the amount of data collected in Spring 2011, a calibration up to a transverse momentum of

160 GeV is derived and studied in detail for four jet algorithms and different jet sizes. The underestimation of the reconstructed energy of the calorimeter jets is improved by the calibration from 60% to 4% in the 20 GeV transverse momentum region and from 25% to less than 1% around 160 GeV. For what concerns particle flow jets, the initial energy underestimation of 20% in the 20 GeV transverse momentum region and 10% around 160 GeV was brought to only 1% over the whole transverse momentum range.
For the first time at a hadron collider, the $Z(\rightarrow \mu^+\mu^-)$+jet events were investigated with $2.96\,\mathrm{pb}^{-1}$ of CMS acquired data for the determination of the jet energy scale. A good agreement in the description of the basic properties of the physics objects involved in the calibration was proved. The transverse momentum balancing of Z bosons and jets was studied using particle flow anti-kt jets with a size parameter of R=0.5. On the one hand, the jet resolution was proved to be $0.98^{+0.12}_{-0.10}$ with respect to the Monte Carlo expectation, therefore perfectly compatible, on the other hand the jet energy scale in the acquired data was $0.93^{+0.04}_{-0.04}$ with respect to the expected value. This disagreement proves the need for a data driven calibration of the absolute jet energy scale, which can be provided by the $Z(\rightarrow \mu^+\mu^-)$+jet technique in the next months. This calibration will then have to be compared with the one based on the balancing of a jet with a photon and a Z decaying in two electrons. Eventually, all these calibrations will be combined to improve their statistical precision.

The Jet Area/Median technique for the jet energy corrections aims at an event-by-event and jet-by-jet subtraction of the pileup and underlying event contributions from the jet energies. This unwanted energy surplus is estimated for each event as the median of the distribution of the jets transverse momenta divided by their area, ρ. In the context of this thesis, this strategy was investigated for the first time with detector data. This investigation was carried out exploiting the 0.9 TeV data acquired by CMS and considering track-jets. It was shown that with a simple modification of the ρ variable, even focussing on charged particles only, with the limitations imposed by a real detector in terms of geometrical acceptance constraints and detection thresholds, the technique provides sensitivity to the different underlying events tunes. The Z1 tune of the Pythia 6 Monte Carlo generator proved to describe the data best.
For the first time, the jet area quantity was investigated with detector data. The study prepares the field for the application of the Jet Area/Median technique with other jet types, like calorimeter and particle flow, at 7 TeV centre of mass energy.

Appendix A

Mathematical Symbols

Pauli matrices:

$$\tau_1 = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix},\ \tau_2 = \begin{pmatrix} 0 & -i \\ i & 0 \end{pmatrix},\ \tau_3 = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix} \tag{A.1}$$

Dirac matrices:

$$\gamma_0 = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix},\ \gamma_i = \begin{pmatrix} 0 & \sigma_i \\ -\sigma_i & 0 \end{pmatrix},\ \gamma_5 = i\gamma_0\gamma_1\gamma_2\gamma_3 = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \tag{A.2}$$

To formulate the fundamental generators of SU(3) the generators can be written as $T^a = \lambda_a/2$, with the Gell-Mann matrices λ_a:

$$\begin{aligned}
\lambda_1 &= \begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}, \quad \lambda_2 = \begin{pmatrix} 0 & -i & 0 \\ i & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}, \quad \lambda_3 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & 0 \end{pmatrix}, \\
\lambda_4 &= \begin{pmatrix} 0 & 0 & 1 \\ 0 & 0 & 0 \\ 1 & 0 & 0 \end{pmatrix}, \quad \lambda_5 = \begin{pmatrix} 0 & 0 & -i \\ 0 & 0 & 0 \\ i & 0 & 0 \end{pmatrix}, \\
\lambda_6 &= \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix}, \quad \lambda_7 = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & -i \\ 0 & i & 0 \end{pmatrix}, \quad \lambda_8 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & -2 \end{pmatrix} \frac{1}{\sqrt{3}}.
\end{aligned} \tag{A.3}$$

The structure constant of the group (f^{abc}) is defined through the commutator relations

$$[T^a, T^b] = if^{abc}T^c. \tag{A.4}$$

Appendix B

RooStatsCms

RooStatsCms [55] is an object oriented statistical framework based on the RooFit technology. Its scope is to allow the modelling, statistical analysis and combination of multiple search channels for new phenomena in high energy physics. It provides a variety of methods described in literature implemented as classes, whose design is oriented to the execution of multiple CPU intensive jobs on batch systems or on the Grid.

B.1 Introduction

The statistical analysis and the combination of measurements has a dominant importance in high energy physics. It is very challenging from the point of view of the tools to be deployed, the communication among the analysis groups and the definition of statistical guidelines. In previous occasions, such as the LEP [22] and Tevatron [110] Electroweak Working Groups already devoted huge efforts in this direction. At the LHC, early results will require the combined analysis of different search channels and eventually the combination of results obtained by different experiments. There will definitely be a need to build complex models, i.e. parametrisations, to describe the experimental distributions, to quantify a possible signal excess in the data or to set a limit on the signal size in the absence of such an excess. In addition, a quantitative statistical treatment will require extensive studies based on toy Monte Carlo pseudo-experiments, and should consider different statistical methods. The combination of analyses require a reliable and practical transfer of data and models across working group and experiment boundaries. Previous attempts to achieve these goals were built upon dedicated code for each analysis, and a very tedious and often error-prone transfer of the obtained results into the combination procedures. In order to perform the statistical treatment of combination of analyses multiple methods are available. The choice of the method to use depends often on the context of the analysis and on

the interpretation of the data by the experimenter. When multiple methods are applicable, a comparison of their results is useful or is even required. It is therefore important to be able to easily switch between methods without too much effort. This was so far not possible as the implementations of different approaches were not unified in a single package and a comparison would require the user to learn how to use a number of packages. It is the lack of a general, easy-to-use tool that drove the decision to develop RooStatsCms (RSC) for analysis modelling, statistical studies and combination of results. A first release of the RSC package has been provided in February 2008. It relies on the ROOT 3.1 extension RooFit 3.1.1, from which it inherits the ability to efficiently describe the analysis model, thereby separating code and descriptive data, and easily perform toy Monte Carlo pseudo-experiments. A selection of different methods for statistical analysis and combination procedures is also included.

B.2 Framework and Software Environment

RooStatsCms is entirely written in C++ and relies on ROOT. To reach a maximum flexibility and exploit all the recent technologies, the RooFit toolkit was chosen as the basis of RooStatsCms. RooStatsCms is integrated in the official CMS Software Framework 3.2, in the `PhysicsTools/RooStatsCms` package, starting from the 3.1.X series.

B.3 Analyses Modelling

In the analysis of a physics process the description of its signal and background components, together with correlations and constraints affecting the parameters, is a critical step. RooStatsCms provides the possibility to easily model the analyses and their combinations through the description of their signal and background shapes, yields, parameters, systematic uncertainties and correlations via analysis configuration files, called *datacards*. The goal of the modelling component of RSC is to parse the datacard and generate from it a model according to the RooFit standards. There are a few classes devoted to this functionality, but the user really needs to deal only with one of them, the `RscCombinedModel` (figure B.1).

The approach described above has mainly three advantages:

1. The factorisation of the analysis description and statistical treatment in two well defined steps.

2. A common base to describe the outcomes of the studies by the analysis groups in a human readable format.

3. A straightforward and documented sharing of the results.

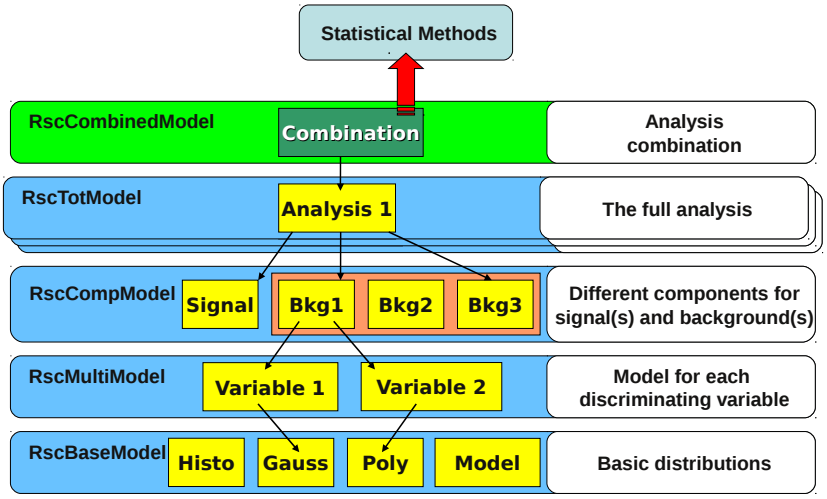

Figure B.1: The RSC model factory classes. At the time, the absence of a low-level object factory in RooFit, drove the development of a hierarchical structure. The user may interact directly only with the RscCombinedModel to build the model to be plugged into the classes implementing the statistical methods.

A datacard is an ASCII file in the ".ini format", therefore presenting key-value pairs organised in sections. This format was preferred to the eXtenisble Markup Language (XML [130]) because of its simplicity and high readability. The parsing and processing of the datacard is achieved through an extension of the RooFit `RooStreamParser` utility class. This class is already rather advanced. Beyond reading strings and numeric parameters from configuration files, it implements the interpretation of conditional statements, file inclusions and comments. In presence of a complicated combination, the user can take advantage in RSC from these features specifying one single model per datacard file and then import all of them in a "combination card". An example of combination datacard for a number counting experiment is:

```
1 # Combination card:
2 #
3 # Models:
4 #
5 # 1) H --> ZZ --> 2mu 2e
6 # 2) H --> ZZ --> 2mu 2e
7 # 3) H --> ZZ --> 4e
8
```

```
9  # Constraints:
10 # The constraint on a variable called "var" must be espressed in a variable
11 # called var_constraint. The syntax for the constraints of different shape are
12 # possible:
13 # - Gaussian:
14 #   example: var_constraint= "Gaussian, 10, 0.3"
15 #   This line generates a gaussian constraint whose mean is 10 and the sigma is
16 #   the 30%. If the mean is 0, the sigma is read as an absolute value.
17
18
19 ############################
20 # The combined model
21 ############################
22 // Here we specify the names of the models built down in the card that we want
23 // to be combined
24 [hzz4l]
25     model = combined
26     components = hzz_4mu, hzz_4e, hzz_2mu2e
27
28 ############################
29 # H -> ZZ -> 4mu
30 ############################
31 [hzz_4mu]
32     variables = x
33     x = 0 L(0 - 1)
34
35 [hzz_4mu_sig]
36     hzz_4mu_sig_yield = 62.78 L(0 - 200)
37
38 [hzz_4mu_sig_x]
39     model = yieldonly
40
41 [hzz_4mu_bkg]
42
43     yield_factors_number = 2
44     yield_factor_1 = scale
45     scale = 1 C L (0 - 3)
46     scale_constraint = Gaussian,1,0.041
47     yield_factor_2 = bkg_4mu
```

```
48      bkg_4mu = 19.93 C
49
50  [hzz_4mu_bkg_x]
51      model = yieldonly
52
53  ##########################
54  # H -> ZZ -> 2mu 2e
55  ##########################
56  [hzz_2mu2e]
57      variables = x
58      x = 0 L(0 - 1)
59
60  [hzz_2mu2e_sig]
61      model = yieldonly
62      hzz_2mu2e_sig_yield = 109.30 L(0 - 200)
63  [hzz_2mu2e_sig_x]
64      model = yieldonly
65
66  [hzz_2mu2e_bkg]
67      yield_factors_number = 2
68      yield_factor_1 = scale
69      scale = 1 C L (0 - 3)
70      scale_constraint = Gaussian,1,0.041
71      yield_factor_2 = bkg_2mu2e
72      bkg_2mu2e = 48.6 C
73
74  [hzz_2mu2e_bkg_x]
75      model = yieldonly
76
77  ##########################
78  # H -> ZZ -> 4e
79  ##########################
80  # Here you can see an example about how a Yield can be set to↩
        be composed of
81  # different factors, so to be able to fit for parameters like↩
        : lumi, xsections..
82  # E.g. Yield = lumi*xsec*eff
83  [hzz_4e]
84      variables = x
85      x = 0 L(0 - 1)
86
87  [hzz_4e_sig]
88      hzz_4e_sig_yield = 38.20 L(0 - 200)
89
```

```
90  [hzz_4e_sig_x]
91      model = yieldonly
92
93  [hzz_4e_bkg]
94      yield_factors_number = 2
95      yield_factor_1 = scale
96      scale = 1 C L (0 - 3)
97      scale_constraint = Gaussian,1,0.041
98      yield_factor_2 = bkg_4e
99      bkg_4e = 17.29 C
100
101 [hzz_4e_bkg_x]
102     model = yieldonly
103
104 ############################
105 # The correlations
106 ############################
107 # The correlations are expressed in blocks. Each blocks ↩
        contains variables.
108 # Only 3 or 2 variables can be grouped in a block.
109 # At first the names of the variables are listed, then the ↩
        values of the
110 # correlations coefficients.
111 # The correlation coefficient 1 represent the correlation ↩
        between the variables
112 # 1-2 and so on, as listed below:
113 #   - corr between var 1 and 2 = correlation_value1
114 #   - corr between var 1 and 3 = correlation_value2
115 #   - corr between var 2 and 3 = correlation_varue3
116 [constraints_block_1]
117 correlation_variable1 = hzz_2mu2e_bkg_yield
118 correlation_variable2 = hzz_4mu_bkg_yield
119 correlation_variable3 = hzz_4e_bkg_yield
120
121 correlation_value1 = 0.99 C # Correlation 1,2
122 correlation_value2 = 0.99 C # Correlation 1,3
123 correlation_value3 = 0.99 C # Correlation 2,3
```

Such a datacard gives rise to a rather complex model, illustrated in the diagram in figure B.2

If the analysis acquires an even higher complexity, like in the case of the $H \to \gamma\gamma$ analysis in CMS [131] which contains several categories of events, a diagram becomes difficult to read. For that purpose, the `model_html` tool can create a small website can to enable the navigation through the model (figure B.3).

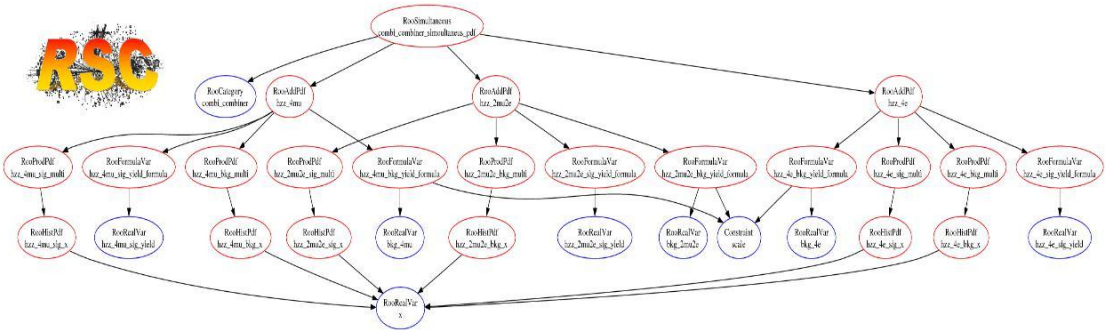

Figure B.2: Models of high complexity can be defined by relative simple datacards. Moreover, their diagrams can be visualised as a diagram.

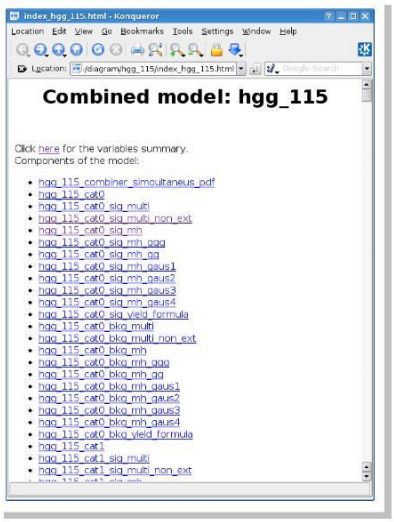

Figure B.3: A small website can be created to inspect the created model. This functionality is particularly useful in presence of very complex models, like the ones that present many categories of events, e.g. for photons in the final state, which are selected according to their quality. With a few clicks, the values of variables, the functional form of signals and backgrounds can be inspected.

Every analysis model can be described as a function of one or many observables, e.g. invariant mass, output of a neural network or topological information regarding the decay products. For each of these variables a description of the signal and background case is to be given, where both signal and background can be divided in multiple components, e.g. multiple background sources. To each signal and background components, a shape and a yield can be assigned. For what concerns the shape, a list of models is present and for those shapes which are not easily parametrisable, a `TH1` histogram in a ROOT file can also be specified. The yields can be expressed as a product of an arbitrary number of single factors, for example

$$Yield = \int \mathrm{L} \cdot \sigma \cdot BR \cdot \epsilon, \tag{B.1}$$

where L is the integrated luminosity, σ a production cross section, BR a decay branching ratio and ϵ is the detection efficiency. All the parameters present in the datacard can be specified as constants or defined in a certain range. In addition to that, exploiting the RSC implementation of the constraints, the user can directly specify the parameter affected by a Gaussian or a log-normal systematic uncertainty. In the former case, correlations can be specified among the parameters via the input of a correlation matrix. In a combination some parameters might need to be the same throughout many analyses, e.g. the luminosity or a background rate. This feature is achieved in the modelling through a "same name, same object" mechanism. Indeed every parameter is represented in memory as a `RooRealVar` or, in presence of systematic uncertainties, as a derived object, the `Constraint` object and the `RscCombinedModel` merges all variables with the same name via an association to the same memory location.

All the features above described are now part of the standard ROOT release. For example, it is now possible to describe the systematic uncertainties via RooFit probability distribution objects, and not parameters. In addition, RooFit implements a low-level object factory able to interpret input strings and, exploiting reflection mechanisms, to allocate the described object: The `RooFactoryWSTool`. On the top of that, a class in RooStats exploits this new feature to allow the users to describe their models through datacards, the `HLFactory` [132].

B.4 Implementation of Statistical Methods

Each statistical method in RooStatsCms is implemented in three classes types and this structure is reflected in the code by three abstract classes (figure B.4):

- The *statistical method* where all the time consuming operations such as Monte Carlo toy experiments or fits are performed.
- The *statistical result* where the results of the computations are collected.

- The *statistical plot* which provides a graphical representation of the statistical result.

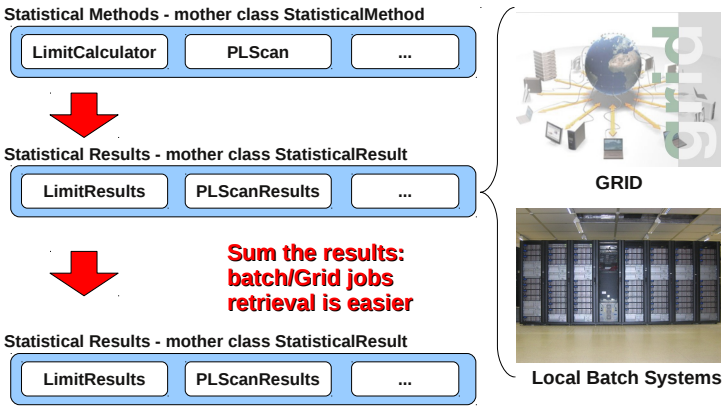

Figure B.4: The RSC classes structure, excluding the modelling components. A Frequentist calculation typically requires almost no input/output but is computationally very intensive. The design eases the submission of jobs to the grid or to a batch system and the recollection of the results.

In many cases, e.g. frequentist approaches, the CPU time needed for the calculations can be considerable. An interesting feature of the statistical result classes is that their objects can be "summed up". Such a feature is very useful to accumulate statistics when combining the outputs of many processes. Indeed, the classes factorisation described above combined with the persistence of the RSC objects, eases the submission of jobs to a batch system or to the Grid and the recollection of the results, allowing to carry out such calculations at in reasonable timescales.

The statistical plot classes play a fundamental role in a statistical analysis, providing a graphical representation of the results in the form of self explanatory plots. The objects of these classes can directly be drawn onto a `TCanvas` via a `draw` method and, when needed, all the graphics primitives like `TGraph`, `TH1F`, `TLegend` that were used to produce the plot can be saved separately in a ROOT file for a further manipulation.

B.5 Graphics Routines

This category of classes is devoted to the mere production of plots. There are two types of plots covered: those that summarise information collected during

the running of the statistical classes, and plots allowing a graphical display of the physics results obtained. In this second category, if on the one hand, RooStatsCms does not provide any user-specific graphics routines, however, during the past decade, the LEP and Tevatron collaborations established a sort of standard to display the results of (combined) searches for new signals [22], for example the so called "green and yellow" plots (figure B.5). This kind of plots are now well accepted in the community, and for this reason utility routines are provided to produce them.

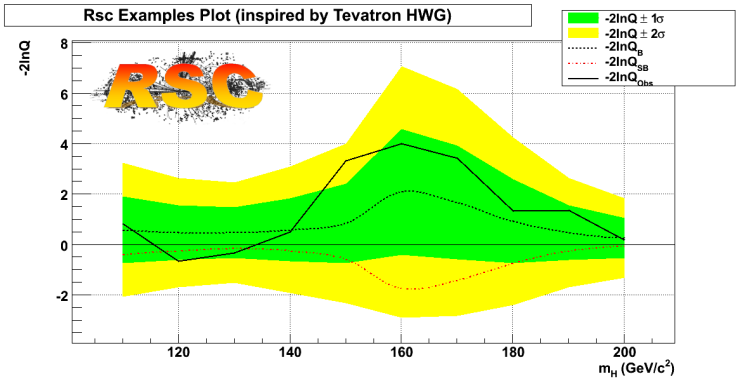

Figure B.5: The green and yellow plot that shows the expected values of the $-2lnQ$ distributions, together with the one and two sigma band associated to the background only one. The plot represents status of the Higgs searches in August 2008 and has here only an illustrative role. For the most recent perspective please visit the Tevatron Electroweak Working Group page [110]

Appendix C

Datasets Used

This appendix summarises the Database Book-keeping System (DBS) [133] entries of the datasets used for this thesis.

Chapter 5

Table C.1: Datasets for Z($\rightarrow \mu^+\mu^-$)+jet Monte Carlo studies

$\hat{p_T}$ GeV	DBS String
0 to 15	/ZmumuJet_Pt0to15/Summer09-MC_31X_V3_7TeV-v1/GEN-SIM-RECO
15 to 20	/ZmumuJet_Pt15to20/Summer09-MC_31X_V3_7TeV-v1/GEN-SIM-RECO
20 to 30	/ZmumuJet_Pt20to30/Summer09-MC_31X_V3_7TeV-v1/GEN-SIM-RECO
30 to 50	/ZmumuJet_Pt30to50/Summer09-MC_31X_V3_7TeV-v1/GEN-SIM-RECO
50 to 80	/ZmumuJet_Pt50to80/Summer09-MC_31X_V3_7TeV-v1/GEN-SIM-RECO
80 to 120	/ZmumuJet_Pt80to120/Summer09-MC_31X_V3_7TeV-v1/GEN-SIM-RECO
120 to 170	/ZmumuJet_Pt120to170/Summer09-MC_31X_V3_7TeV-v1/GEN-SIM-RECO
170 to 230	/ZmumuJet_Pt170to230/Summer09-MC_31X_V3_7TeV-v1/GEN-SIM-RECO
230 to 300	/ZmumuJet_Pt230to300/Summer09-MC_31X_V3_7TeV-v1/GEN-SIM-RECO
greater than 300	/ZmumuJet_Pt300toInf/Summer09-MC_31X_V3_7TeV-v1/GEN-SIM-RECO

Table C.2: Datasets for Z($\rightarrow \mu^+\mu^-$)+jet background estimation Monte Carlo studies.

Description	DBS String
$Z \rightarrow \tau\tau$	/Ztautau_M20_CTEQ66-powheg/Spring10-START3X_V26_AODSIM-v2/AODSIM
$W \rightarrow \mu\nu$	/Wmunu/Spring10-START3X_V26_S09-v1/AODSIM

Table C.3: Datasets for Z($\rightarrow \mu^+\mu^-$)+jet data Monte Carlo comparisons

$\hat{p_T}$ GeV	DBS String
0 to 15	/ZmumuJet_Pt0to15/Summer10-START36_V9_S09-v1/GEN-SIM-RECO
15 to 20	/ZmumuJet_Pt15to20/Summer10-START36_V9_S09-v1/GEN-SIM-RECO
20 to 30	/ZmumuJet_Pt20to30/Summer10-START36_V9_S09-v1/GEN-SIM-RECO
30 to 50	/ZmumuJet_Pt30to50/Summer10-START36_V9_S09-v1/GEN-SIM-RECO
50 to 80	/ZmumuJet_Pt50to80/Summer10-START36_V9_S09-v1/GEN-SIM-RECO
80 to 120	/ZmumuJet_Pt80to120/Summer10-START36_V9_S09-v1/GEN-SIM-RECO
120 to 170	/ZmumuJet_Pt120to170/Summer10-START36_V9_S09-v1/GEN-SIM-RECO
170 to 230	/ZmumuJet_Pt170to230/Summer10-START36_V9_S09-v1/GEN-SIM-RECO
230 to 300	/ZmumuJet_Pt230to300/Summer10-START36_V9_S09-v1/GEN-SIM-RECO
Muon Stream	/Mu/Run2010A-PromptReco-v4/RECO

Table C.4: Dataset for the investigation of Z($\rightarrow \mu^+\mu^-$)+jet cuts efficiencies.

Description	DBS String
$Z \rightarrow \mu\mu$	/Zmumu/Summer10-START36_V9_S09-v1/GEN-SIM-RECO

Chapter 6

Table C.5: Data samples used for the analysis. The Pythia 8 dataset was produced privately and can be looked up in the DBS instance *cms_dbs_ph_analysis_01*.

Data Sample	Events
`/MinimumBias/BeamCommissioning09-Dec19thReReco_336p3_v2/RECO`	19,681,382
`/MinBias/Summer09-STARTUP3X_V8K_900GeV-v1/GEN-SIM-RECO`	10,951,200
`/MinBias/Summer09-STARTUP3X_V8K_900GeV_P0-v2/GEN-SIM-RECO`	2,195,680
`/MinBias/Summer09-STARTUP3X_V8K_900GeV_DW-v1/GEN-SIM-RECO`	2,048,000
`/MinBias/Summer09-STARTUP3X_V8K_900GeV_ProQ20-v1/GEN-SIM-RECO`	2,278,400
`/MinBias/Summer09-STARTUP3X_V8K_900GeV_P8-priv/GEN-SIM-RECO`	310,000
`/MinBiasCW900A/Summer09-STARTUP3X_V8K_900GeV-v1/GEN-SIM-RECO`	2,167,605
`/MinBias/Summer09-MC_3XY_V9B_900GeV-v2/GEN-SIM-RECO`	10,985,000
`Z1 Tune not yet published in DBS`	300,000

Table C.6: List of runs, luminosity blocks and bunch crossings that have been used for this analysis (taken from [128]). Moreover, the HLT trigger "Physics Bit" was required selecting events when the detector was fully operational.

Run	Luminosity Block	Bunch Crossings
124020	12-94	51, 151, 2824
124022	60-69	51, 151, 2824
124023	41-96	51, 151, 2824
124024	2-83	51, 151, 2824
124027	24-39	51, 151, 2824
124030	1-31	51, 151, 2824
124230	26-68	51, 151, 232, 1024, 1123, 1933, 2014, 2824, 2905

Appendix D

Additional Figures

Figures for Chapter 6

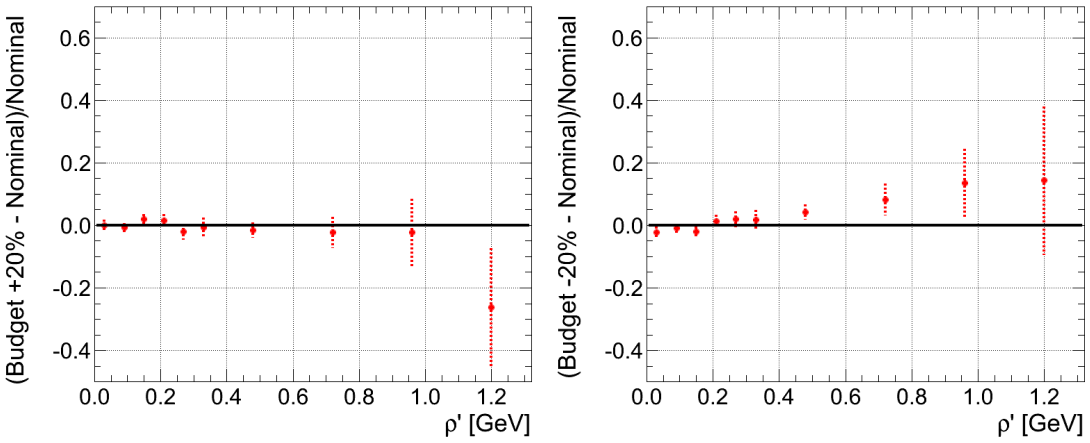

Figure D.1: Systematic effect of increasing or decreasing the material budget by 20%. A conversion factor of 0.25 has to be applied.

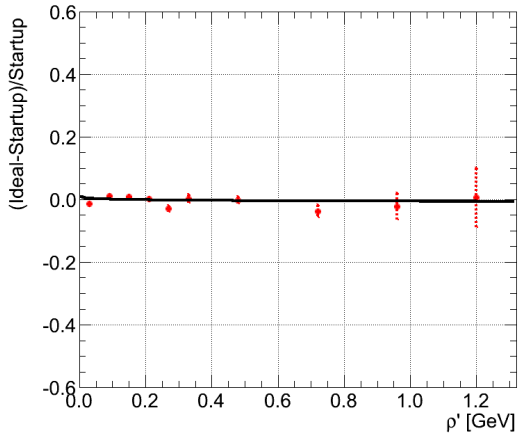

Figure D.2: Systematic effect introduced by the uncertainty in the knowledge of the tracker alignment.

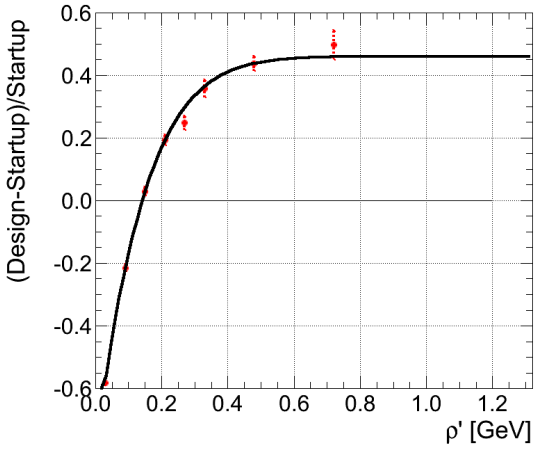

Figure D.3: Systematic effect from non-operational tracker channels. A conversion factor of 0.05 has to be applied.

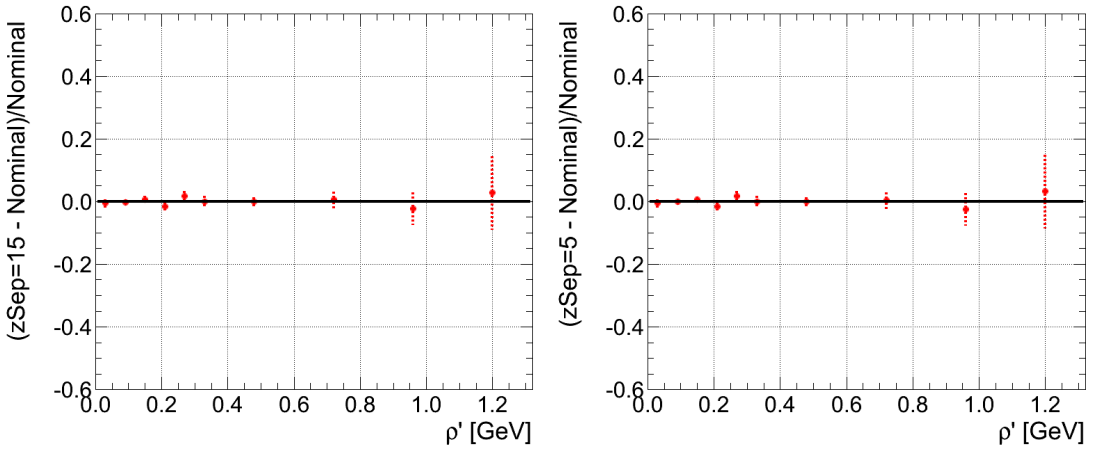

Figure D.4: Systematic effect from varying the minimal vertex separation in z.

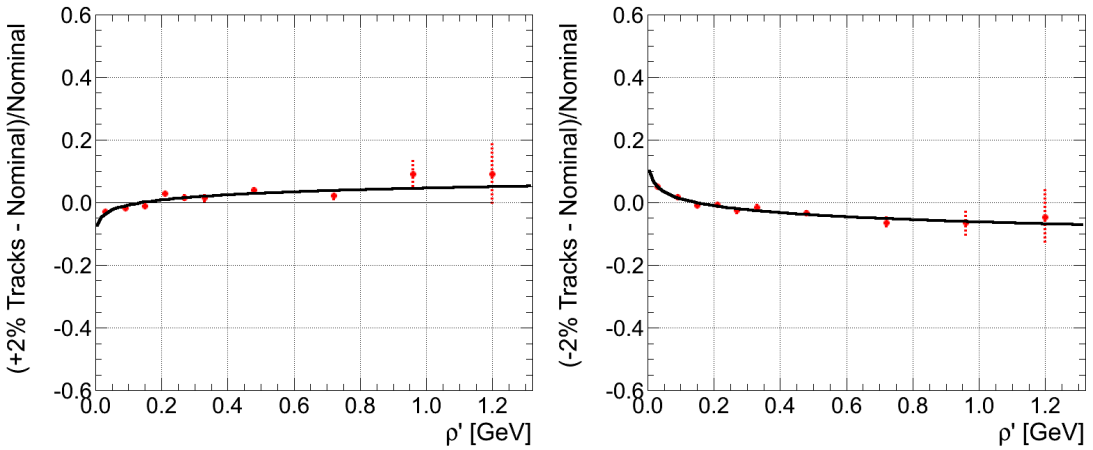

Figure D.5: Systematic effect of artificially adding and removing tracks.

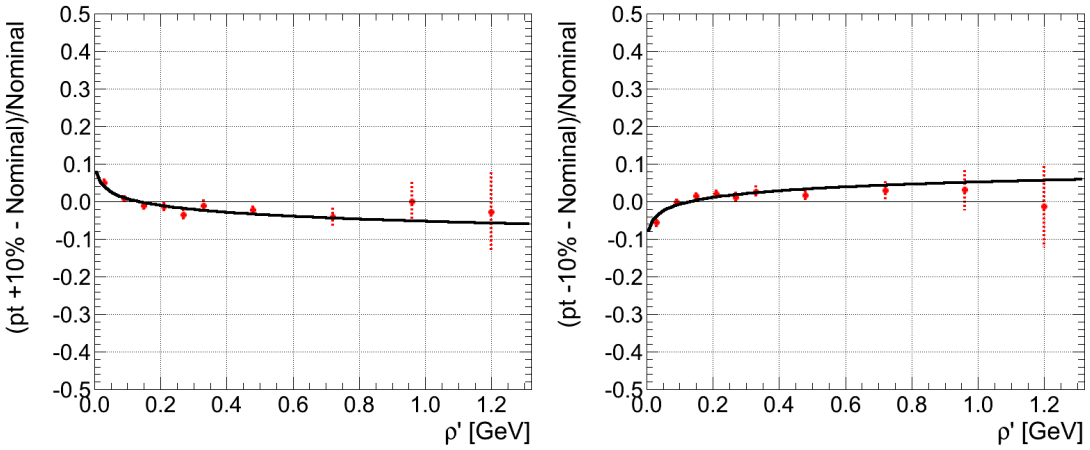

Figure D.6: A systematic uncertainty of a few percent is introduced by the 10% variation of the transverse momentum cut on the jet components.

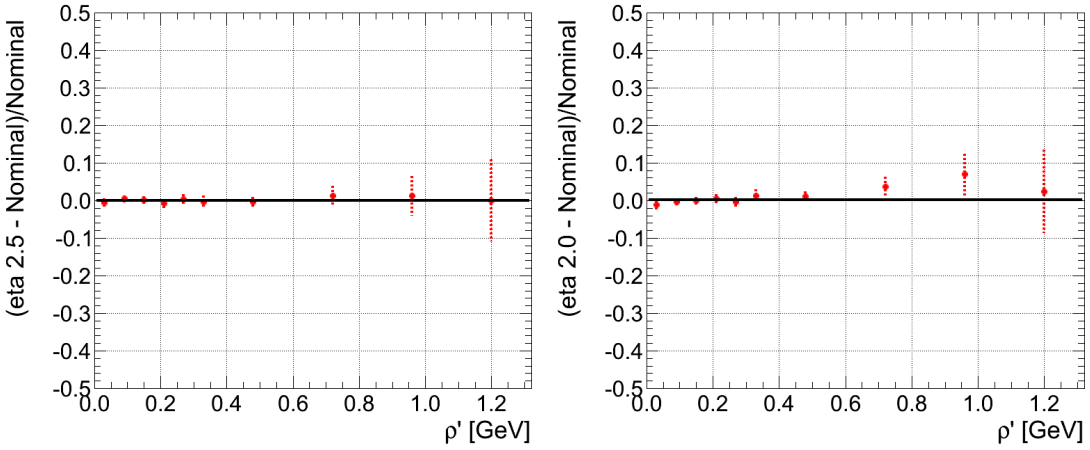

Figure D.7: A systematic uncertainty compatible with zero is introduced by the variation of the pseudo-rapidity cut on the jet components to 2.0 and 2.5 respectively.

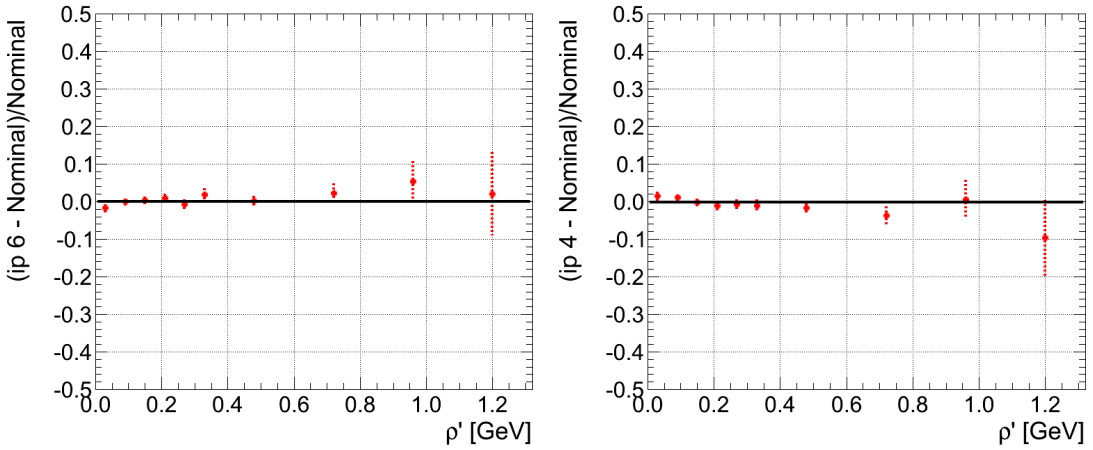

Figure D.8: A systematic uncertainty compatible with zero is introduced by the variation of the longitudinal and transverse impact parameter cuts on the jet components to 3 and 7 respectively.

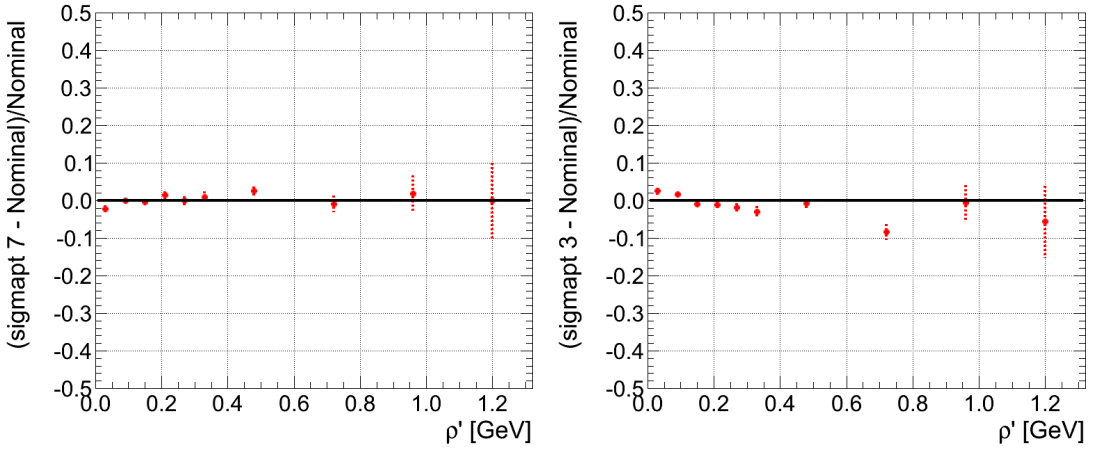

Figure D.9: A negligible systematic uncertainty is introduced by the variation of the cut on σ_{p_T}/p_T on the jet components.

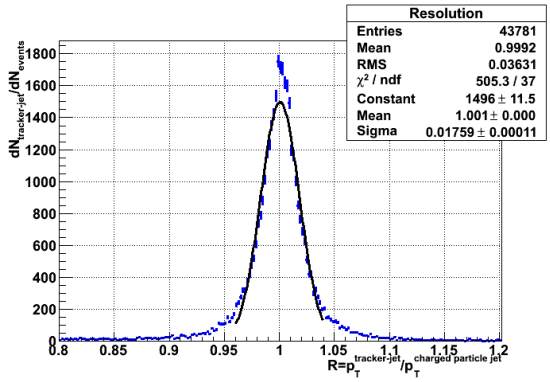

Figure D.10: Transverse momentum resolution of track-jets.

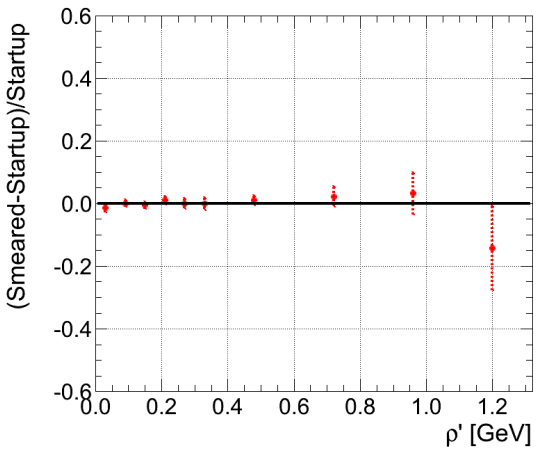

Figure D.11: Systematic effect on ρ' due to artificial smearing of the transverse momentum of track-jets.

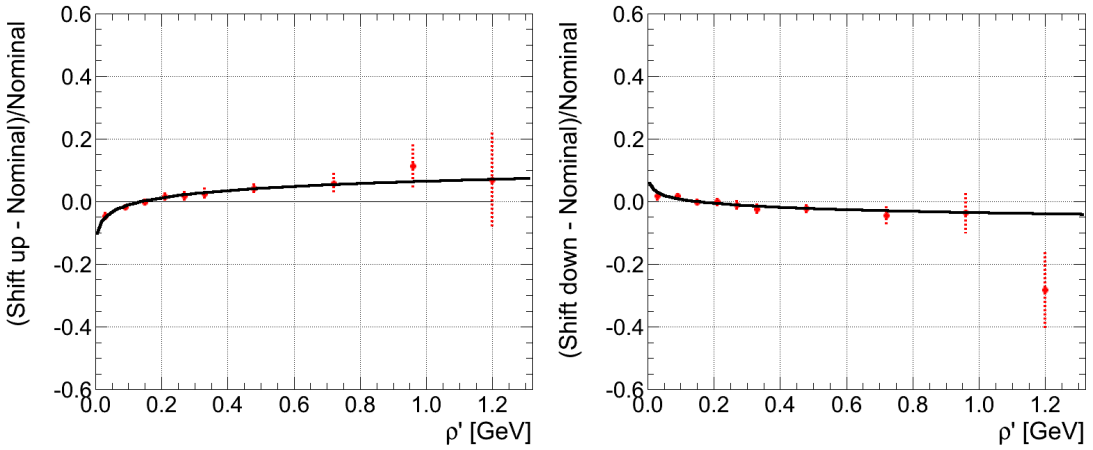

Figure D.12: Effect of the track-jet response uncertainty on ρ'.

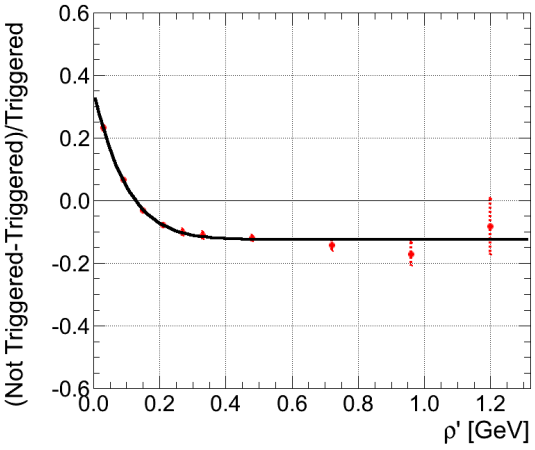

Figure D.13: Effect of trigger efficiency bias between data and trigger simulation. A conversion factor of 0.25 has to be applied.

List of Figures

List of Tables

Bibliography

[1] D. J. Griffiths, *Introduction to elementary particles*. John Wiley and Sons, Inc., 1987.

[2] C. Rubbia, "Experimental observation of the intermediate vector bosons W^+, W^- and Z^0," *Rev. Mod. Phys.*, vol. 57, pp. 699–722, Jul 1985.

[3] The Super-Kamiokande Collaboration, "Evidence for oscillation of atmospheric neutrinos," *PRL*, vol. 81, pp. 1562–1567, 1998.

[4] R. Mann, *An Introduction to Particle Physics and the Standard Model*. CRC Press, 2010.

[5] M. E. Peskin and D. V. Schroeder, *An Introduction to Quantum Field Theory (Frontiers in Physics)*. Perseus Books, 2008.

[6] J. Goldstone, "Field theories with superconductor solutions," *Nuovo Cimento*, vol. 19, pp. 154–164, Aug 1960.

[7] C. Amsler, M. Doser, M. Antonelli, D. Asner, K. Babu, H. Baer, H. Band, R. Barnett, E. Bergren, and J. Beringer, "Review of Particle Physics," *Physics Letters B*, vol. 667, pp. 1–6, 9 2008.

[8] G. C. Branco, L. Lavoura, and J. P. Silva, *CP Violation*. Clarendon Press, 1999.

[9] A. D. Martin and F. Halzen, *Quarks and Leptons*. John Wiley and Sons, Inc., 1984.

[10] P. A. M. Dirac, "The quantum theory of the emission and absorption of radiation," *Proceedings of the Royal Society of London. Series A, Containing Papers of a Mathematical and Physical Character*, vol. 114, no. 767, pp. 243–265, 1927.

[11] The D0 Collaboration, "Determination of the Strong Coupling Constant from the Inclusive Jet Cross Section in $p\bar{p}$ Collisions at $\sqrt{s} = 1.96$ TeV," vol. D0 Note 5979-CONF, 2009.

[12] M. Breidenbach, J. I. Friedman, H. W. Kendall, E. D. Bloom, D. H. Coward, H. DeStaebler, J. Drees, L. W. Mo, and R. E. Taylor, "Observed behavior of highly inelastic electron-proton scattering," *Phys. Rev. Lett.*, vol. 23, pp. 935–939, Oct 1969.

[13] "The ZEUS Experiment." `http://www-zeus.desy.de/`.

[14] "The H1 Experiment at HERA." `http://www-h1.desy.de`.

[15] "The Coordinated Theoretical-Experimental Project on QCD." `http://www.phys.psu.edu/~cteq/`.

[16] "MRS/MRST/MSTW Parton Distributions." `http://durpdg.dur.ac.uk/hepdata/mrs.html`.

[17] The ALEPH, DELPHI, L3, OPAL, SLD Collaborations, the LEP Electroweak Working Group, the SLD Electroweak and Heavy Flavour Groups, "Precision Electroweak Measurements on the Z Resonance," *Phys. Rep.*, vol. 427, p. 302, Sep 2005.

[18] CMS Collaboration, "Measurement of the W and Z inclusive production cross sections at $\sqrt{s}$=7 TeV with the CMS experiment at the LHC," *CMS Physics Analysis Summary*, vol. CMS-PAS-EWK-10-002, 2010.

[19] The CMS Collaboration, "Electromagnetic calorimeter calibration with 7 TeV data," *CMS Physics Analysis Summary*, vol. CMS-PAS-EGM-10-003, 2010.

[20] M. Della Negra, L. Foá, A. Hervé, A. D. Roeck, P. Sphicas, L. Silvestris, A. Yagil, and D. Acosta, *CMS Physics: Technical Design Report*. Technical Design Report CMS, Geneva: CERN, 2006.

[21] The CMS Collaboration, "Commissioning of b-jet identification with pp collisions at $\sqrt{s} = 7$ TeV," *CMS Physics Analysis Summary*, vol. CMS-PAS-BTV-10-001, 2010.

[22] "The LEP Electroweak Working Group." `http://lepewwg.web.cern.ch`.

[23] *Infinitely CERN*. Editions Suzanne Hurter, 2005.

[24] T. S. Pettersson and P. Lefèvre, "The Large Hadron Collider: conceptual design," Tech. Rep. CERN-AC-95-05 LHC, CERN, Geneva.

[25] "Large Electron-positron Collider - LEP." `http://public.web.cern.ch/public/en/research/lep-en.html`.

[26] J. D. Jackson, *Classical Electrodynamics*. Wiley, August 1998.

[27] R. Bailey, B. Balhan, C. Bovet, B. Goddard, N. Hilleret, J. M. Jimenez, R. Jung, M. Placidi, M. Tavlet, and G. Von Holtey, "Synchrotron Radiation Effects at LEP," p. 3, Jun 1998.

[28] "Big Science: The LHC in Pictures." `http://bigscience.web.cern.ch`.

[29] "CMS Luminosity - Public Results." `https://twiki.cern.ch/twiki/bin/view/CMSPublic/LumiPublicResults2010`.

[30] CMS Collaboration, "The Compact Muon Solenoid Technical Proposal," Tech. Rep. CERN/LHCC 94-38 , Geneva, 1994.

[31] "ATLAS: technical proposal for a general-purpose pp experiment at the Large Hadron Collider at CERN," Tech. Rep. CERN/LHCC 94-43, Geneva, 1994.

[32] LHCb Collaboration, "LHCb Technical Proposal," Tech. Rep. CERN/LHCC 98-004, 1998.

[33] ALICE Collaboration, "ALICE Technical Proposal," Tech. Rep. CERN/LHCC 2001-021, 2001.

[34] CMS Collaboration, *The Tracker System Project Technical Design Report*. No. CERN/LHCC 2000-016, 2000.

[35] G. Knoll, *Radiation detection and measurement*. Wiley, 1989.

[36] CMS Collaboration, "The CMS experiment at the cern LHC," *Journal of Instrumentation*, vol. 3, 2008.

[37] L. Masetti, F. Hartmann, S. Y. Shah, and R. Stringer, "The CMS Tracker Detector Control System," *Nucl. Instrum. Methods Phys. Res., A*, vol. 604, no. 1-2, pp. 281–283, 2009.

[38] CMS Collaboration, *The Electromagnetic Calorimeter Technical Design Report*. No. CERN/LHCC 97-033.

[39] A. Benaglia *et al.*, "Intercalibration at 2006 ECAL testbeam with the single crystal technique," *CMS Detector Note*, vol. CMS DN-2007/001, 2007.

[40] A. Benaglia *et al.*, "Intercalibration at 2006 ECAL testbeam with the S25 technique," *CMS Detector Note*, vol. CMS DN-2007/002, 2007.

[41] CMS Collaboration, *The Hadron Calorimeter Project Technical Design Report*. No. CERN/LHCC 97-032.

[42] CMS Collaboration, *The Muon Project Technical Design Report*. No. CERN/LHCC 97-032.

[43] CMS Collaboration, *The Trigger and Data Acquisition Project, Volume I: The Level-1 Trigger Technical Design Report*. No. CERN/LHCC 2000-038 .

[44] CMS Collaboration, *The Trigger and Data Acquisition Project, Volume II: Data Acquisition and High Level Trigger Technical Design Report*. No. CERN/LHCC 2002-026 .

[45] "CMSSW CVS Repository." `http://cmssw.cvs.cern.ch`.

[46] R. Brun and F. Rademakers, "ROOT - An Object Oriented Data Analysis Framework," *Proceedings AIHENP'96 Workshop*, no. DOE-ER-40389-69, pp. 81–86, 1997. `http://root.cern.ch/`.

[47] W. Verkerke and D. Kirkby, "The RooFit toolkit for data modeling," Tech. Rep. physics/0306116, SLAC, Stanford, CA, June 2003. `http://root.cern.ch/drupal/content/roofit`.

[48] L. Moneta, K. Belasco, K. Cranmer, A. Lazzaro, D. Piparo, G. Schott, W. Verkerke, and M. Wolf, "The RooStats Project," *Proceedings Of Science*, To be Published. ArXiv Preprint 1009.1003.

[49] B. Stroustrup, *The C++ Programming Language, First Edition*. Addison-Wesley, 1986.

[50] "The PAW History Seen by the CNLs," Apr 2001.

[51] D. Piparo and G. Quast, *Diving into ROOT*. Particle Physics Institute: Karlsruhe Institute of Technology (KIT).

[52] "The BaBar Collaboration." `http://www.slac.stanford.edu`.

[53] G. Lepage, "A new algorithm for adaptative multidimensional integration," *J. Comput. Phys.*, vol. 27, pp. 192–203, 1978.

[54] F. James, "Determining the statistical significance of experimental results," *CERN Report*, vol. 6th CERN School of Computing, pp. 182–219, March 1981.

[55] D. Piparo, G. Quast, and G. Schott, "RooStatsCms a tool for analysis modelling, combination and statistical studies," *J. Phys. Conf. Ser.*, vol. 219 032034, 2010.

[56] K. Cranmer, "Statistics for the LHC: Progress, challenges and future," *Proceedings of the PHYSTAT LHC Workshop*, vol. 47, June 2007.

[57] M. Lutz, *Programming Python, Third Edition*. O'Reilly Media, August 2006.

[58] S. Mrenna, T. Sjostrand, and P. Skands, "Pythia 6.4 physics and manual," *JHEP*, vol. 605, 2006.

[59] G. Corcella, I. G. Knowles, G. Marchesini, S. Moretti, K. Odagiri, P. Richardson, M. H. Seymour, and B. R. Webber, "HERWIG 6: an event generator for Hadron Emission Reactions With Interfering Gluons (including supersymmetric processes)," *J. High Energy Phys.*, vol. 01, p. 93, Nov 2000.

[60] M. L. Mangano, M. Moretti, F. Piccinini, R. Pittau, and A. Polosa, "ALPGEN, a generator for hard multiparton processes in hadronic collisions," *J. High Energy Phys.*, vol. 07, p. 35, Jun 2002.

[61] J. Alwall, P. Demin, S. de Visscher, R. Frederix, M. Herquet, F. Maltoni, T. Plehn, D. L. Rainwater, and T. Stelzer, "MadGraph/MadEvent v4: the new web generation," *Journal of High Energy Physics*, vol. 2007, no. 9, p. 28, 2007.

[62] T. Gleisberg, S. Hoeche, F. Krauss, M. Schoenherr, S. Schumann, F. Siegert, and J. Winter, "Event generation with SHERPA 1.1," *J. High Energy Phys.*, vol. 02, p. 007, Dec 2008.

[63] R. K. Ellis, W. J. Stirling, and B. R. Webber, *QCD and Collider Physics*. Cambridge Monographs on Particle Physics, Nuclear Physics and Cosmology, Cambridge University Press, December 2003.

[64] "FNAL: Fermi NAtional Laboratory." `http://www.fnal.gov/`.

[65] "CDF Collaboration." `http://www-cdf.fnal.gov/`.

[66] CDF Collaboration, "Charged jet evolution and the underlying event in $p\bar{p}$ collisions at 1.8 TeV," *Phys. Rev.*, vol. D65, p. 092002, 2002.

[67] CDF Collaboration, "The underlying event in hard interactions at the Tevatron $\bar{p}p$ collider," *Phys. Rev.*, vol. D70, p. 072002, 2004.

[68] F. Rick, "Physics at the Tevatron," *Acta Phys. Polon.*, vol. B39, p. 2611, 2008.

[69] R. Field, "Studying the underlying event at CDF and the LHC," in *Proceedings of the First International Workshop on Multiple Partonic Interactions at the LHC MPI'08, October 27-31, 2008* (P. Bartalini and L. Fanó, eds.), (Perugia, Italy), Oct. 2009.

[70] J. Pumplin *et al.*, "New generation of parton distributions with uncertainties from global QCD analysis," *JHEP*, vol. 07, p. 012, 2002.

[71] A. Moraes, C. Buttar, and I. Dawson, "Prediction for minimum bias and the underlying event at LHC energies," *Eur. Phys. J.*, vol. C50, p. 435, 2007.

[72] A. Buckley, H. Hoeth, H. Lacker, H. Schulz, and J. E. von Seggern, "Systematic event generator tuning for the LHC," *Eur. Phys. J.*, vol. C65, p. 331, 2010.

[73] P. Z. Skands, "The Perugia Tunes," 2009.

[74] H. L. Lai *et al.*, "Global QCD analysis of parton structure of the nucleon: CTEQ5 parton distributions," *Eur. Phys. J.*, vol. C12, p. 375, 2000.

[75] P. Z. Skands and D. Wicke, "Non-perturbative QCD effects and the top mass at the Tevatron," *Eur. Phys. J.*, vol. C52, p. 133, 2007.

[76] CMS Collaboration, "Transverse momentum and pseudorapidity distributions of charged hadrons in pp collisions at $\sqrt{s} = 0.9$ and 2.36 TeV," *J. High Energy Phys.*, vol. 02, p. 33, Feb 2010.

[77] CMS Collaboration, "Transverse-momentum and pseudorapidity distributions of charged hadrons in *pp* collisions at $\sqrt{s} = 7$ TeV," *Phys. Rev. Lett.*, vol. 105, p. 26, May 2010.

[78] CMS Collaboration, "Measurement of the underlying event activity in proton-proton collisions at 900 gev," *CMS Physics Analysis Summary*, vol. CMS-PAS-QCD-10-001, 2010.

[79] R. Field, "Soft QCD at the LHC: Findings and Surprises," 2010. ISMD10 Conference - Antwerp.

[80] M. Heinrich, *The Influence of Hadronisation and Underlying Event on the Inclusive Jet Cross-Section at the LHC*. 2007.

[81] S. Agostinelli *et al.*, "G4 a simulation toolkit," *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, vol. 506, no. 3, pp. 250–303, 2003.

[82] T. Speer, W. Adam, R. Frohwirth, A. Strandlie, T. Todorov, and M. Winkler, "Track Reconstruction in the CMS Tracker," Tech. Rep. CMS-CR-2005-014 , CERN, Geneva, Jul 2005.

[83] G. P. Salam, "Towards jetography," Tech. Rep. arXiv:0906.1833, Jun 2009. Comments: 88 pages, 27 figures, an extended version of lectures given at the CTEQ/MCNET school, Debrecen, Hungary, August 2008.

[84] CMS Collaboration, "Measurement of the Inclusive Jet Cross Section in pp Collisions at 7 TeV," *CMS Physics Analysis Summary*, vol. CMS-PAS-QCD-10-011, 2010.

[85] M. Cacciari, "Fastjet: a code for fast kt clustering, and more," Tech. Rep. hep-ph/0607071 LPTHE-P-2006-04, Paris 11. Lab. Phys. Theor. Hautes Energ., Orsay, Jul 2006.

[86] G. P. Salam and G. Soyez, "A practical seedless infrared-safe cone jet algorithm," *J. High Energy Phys.*, vol. 05, p. 086. 42 p, Apr 2007.

[87] S. D. Ellis and D. E. Soper, "Successive combination jet algorithm for hadron collisions," *Phys. Rev. D*, vol. 48, pp. 3160–3166. 15 p, Apr 1993.

[88] Y. L. Dokshitzer, G. Leder, S. Moretti, and B. R. Webber, "Better jet clustering algorithms," *J. High Energy Phys.*, vol. 08, no. hep-ph/9707323. CAVENDISH-HEP-97-06, p. 001, 1997.

[89] M. Cacciari, G. P. Salam, and G. Soyez, "The anti-k_t jet clustering algorithm," *J. High Energy Phys.*, vol. 04, p. 063. 12 p, Feb 2008.

[90] M. Cacciari, G. P. Salam, and G. Soyez, "The catchment area of jets," *Journal of High Energy Physics*, vol. 2008, no. 04, p. 005, 2008.

[91] CMS Collaboration, "Commissioning of TrackJets in pp Collisions at 7 TeV," *CMS Physics Analysis Summary*, vol. CMS-PAS-JME-10-006, 2010.

[92] L. Apanasevich, A. Bhatti, F. Chlebana, C. Dragoiu, R. Harris, M. Jha, K. Kousouris, P. Kurt, Z. Qi, F. Ratnikov, P. Schieferdecker, A. Smoron, H. Topakli, N. Varelas, and M. Zielinski, "Performance of Jet Algorithms in CMS," no. CMS AN-2008/001, 2008.

[93] CMS Collaboration, "The Jet Plus Tracks Algorithm for Calorimeter Jet Energy Corrections in CMS," *CMS Physics Analysis Summary*, vol. CMS-PAS-JME-09-002, 2009.

[94] CMS Collaboration, "Commissioning of the particle-flow event reconstruction with the first LHC collisions recorded in the CMS detector," *CMS Physics Analysis Summary*, vol. CMS-PAS-PFT-10-001, 2010.

[95] C. Eck, J. Knobloch, L. Robertson, I. Bird, K. Bos, N. Brook, D. Dellmann, I. Fisk, D. Foster, B. Gibbard, C. Grandi, F. Grey, J. Harvey, A. Heiss, F. Hemmer, S. Jarp, R. Jones, D. Kelsey, M. Lamanna, H. Marten, P. Mato-Vila, F. Ould-Saada, B. Panzer-Steindel, L. Perini, Y. Schutz, U. Schwickerath, J. Shiers, and T. Wenaus, *LHC computing Grid: Technical Design Report*. Technical Design Report LCG, Geneva: CERN, Jun 2005.

[96] "GridKa." `http://grid.fzk.de`.

[97] "Centro Nazionale per la Ricerca e Sviluppo nelle Tecnologie Informatiche e Telematiche - CNAF." `http://www.cnaf.infn.it`.

[98] "German National Analysis Facility - NAF." `http://naf.desy.de/`.

[99] G. Bayatyan, M. Della Negra, L. Foá, A. Hervé, and A. Petrilli, *CMS computing: Technical Design Report*. Technical Design Report CMS, Geneva: CERN, 2005. Submitted on 31 May 2005.

[100] A. Scheurer and the German CMS Community, "German contributions to the CMS computing infrastructure," *Journal of Physics: Conference Series*, vol. 219, no. 6, p. 062064, 2010.

[101] F. James, *Statistical Methods in Experimental Physics*. World Scientific Publishing Co Pte Ltd, January 2007.

[102] W.J. Metzger, *Statistical Methods in Data Analysis*. Katholieke Universiteit Nijmegen, 2002.

[103] P. Laplace, "Théorie analytique des probabilités," *Courcier Imprimeur, Paris*, 1812.

[104] R. Kass and L. Wasserman, "The selection of prior distributions by formal rules," *Journal of the American Statistical Association*, vol. 91, pp. 1343–1370, September 1996.

[105] L. Demortier, S. Jain, and H. B. Prosper, "Reference priors for high energy physics," Tech. Rep. arXiv:1002.1111, Feb 2010.

[106] V. Bartsch and G. Quast, "Expected signal observability at future experiments," vol. CERN-CMS-NOTE-2005-004, Feb 2005.

[107] A. Read, "Modified frequentist analysis of search results (the CLs method)," vol. CERN-OPEN-2000-205, 2000.

[108] M. Wolf, "Statistical Combination of Decay Channels in Searches for the Higgs Boson at the LHC," *EKP Note*, vol. IEKP-KA/2009-9, July 2010.

[109] "Online documentation of the HypotestInverter class of RooStats." `http://root.cern.ch/root/html/RooStats__HypoTestInverter.html`.

[110] "The Tevatron Electroweak Working Group." `http://tevewwg.fnal.gov`.

[111] CMS Collaboration, "Search for the Standard Model Higgs Boson Produced in Vector Boson Fusion and Decaying into a τ Pair in CMS with 1 fb^{-1}," *CMS Physics Analysis Summary*, vol. CMS-PAS-HIG-08-008, 2008.

[112] CMS Collaboration, "Search Strategy for a Standard Model Higgs Boson Decaying to Two W Bosons in the Fully Leptonic Final State," *CMS Physics Analysis Summary*, vol. CMS-PAS-HIG-08-006, 2008.

[113] S. Baffioni *et al.*, "Projected exclusion limits on the SM Higgs boson cross sections obtained by combining the $H \to WW$ and ZZ decay channels," *CMS Analysis Note*, vol. AN-2009/020, 2009.

[114] CMS Collaboration, "Plans for jet energy corrections at CMS," *CMS Physics Analysis Summary*, vol. CMS-PAS-JME-07-002, 2008.

[115] The CMS collaboration, "Offset Energy Correction for Cone Jets," *CMS Physics Analysis Summary*, vol. CMS-PAS-JME-09-003, Jun 2009.

[116] The CMS Collaboration, "Determination of the Relative Jet Energy Scale at CMS from Dijet Balance," *CMS Physics Analysis Summary*, vol. CMS-PAS-JME-08-003, May 2009.

[117] CMS Collaboration, "Jet Performance in pp Collisions at 7 TeV," *CMS Physics Analysis Summary*, vol. CMS-PAS-JME-10-003, 2010.

[118] CMS Collaboration, "Jet energy calibration with photon+jet events," *CMS Physics Analysis Summary*, vol. CMS-PAS-JME-09-004, Jul 2009.

[119] CMS Collaboration, "Calibration of the absolute jet energy scale with $Z \to \mu^+\mu^-$ + jet events at CMS," *CMS Physics Analysis Summary*, vol. CMS-PAS-JME-09-009, July 2009.

[120] C. Ochando, "Search for the Higgs boson in the $ZH \to \nu\nu bb$ channel," FERMILAB-THESIS-2008-78.

[121] CMS Collaboration, "Determination of the jet energy scale using $Z \to e^+e^-$ + jet p_t balance and a procedure for combining data driven corrections," *CMS Physics Analysis Summary*, vol. CMS-PAS-JME-09-005, July 2009.

[122] CMS Collaboration, "Tracking and vertexing results from first collisions," *CMS Physics Analysis Summary*, vol. CMS-PAS-TRK-10-001, 2010.

[123] C. Buttar *et al.*, "Standard model handles and candles working group: Tools and jets summary report," Tech. Rep. arXiv:0803.0678, Mar 2008.

[124] N. Saoulidou, "Particle flow jet identification criteria," *CMS Analysis Note*, vol. AN-2010/003.

[125] M. Cacciari, G. P. Salam, and S. Sapeta, "On the characterisation of the underlying event," Dec 2009. Comments: 40 pages, 17 figures.

[126] M. Cacciari and G. P. Salam, "Pileup subtraction using jet areas," *Phys. Lett. B*, vol. 659, pp. 119–126. 10 p, Jul 2007. Comments: 10 pages, 6 figures.

[127] The CMS Collaboration, "Measurement of the Underlying Event Activity with the Jet Area/Median Approach at 0.9 TeV," *CMS Physics Analysis Summary*, vol. CMS-PAS-QCD-10-005 and AN 2010/103, 2010.

[128] CMS Collaboration, "The underlying event in proton - proton collisions at 900 gev," *CMS Analysis Note*, vol. AN-2010/018, 2010.

[129] The CMS Collaboration, "Transverse-momentum and pseudorapidity distributions of charged hadrons in pp collisions at $\sqrt{s} = 0.9$ and 2.36 TeV," *Journal of High Energy Physics*, vol. 2010, pp. 1–35, 2010. 10.1007/JHEP02(2010)041.

[130] "The eXtensible Markup Language." `http://www.w3.org/XML`.

[131] M. Pieri, S. Bhattacharya, I. Fisk, J. Letts, V. Litvin, and J. G. Branson, "Inclusive Search for the Higgs Boson in the $H \to \gamma\gamma$ Channel," *CMS-NOTE*, Jun 2006.

[132] "Online documentation of the HighLevelFactory class of RooStats." `http://root.cern.ch/root/html/RooStats__HLFactory.html`.

[133] V. Kuznetsov, D. Riley, A. Afaq, V. Sekhri, Y. Guo, and L. Lueking, "The CMS DBS query language," *Journal of Physics: Conference Series*, vol. 219, no. 4, p. 042043, 2010.