Università degli Studi di Torino

facoltà di scienze matematiche, fisiche e naturali
Corso di Laurea Magistrale in Fisica

Study of the b-tagging algorithms of the experiment CMS at LHC

Tesi di Laurea Magistrale in Fisica

Relatore:
Prof. Ernesto Migliore
Controrelatore:
Prof. Elena Botta

Candidato:
Alberto Traverso

Anno Accademico 2011–2012

Contents

List of Figures

List of Tables

Ringraziamenti

Contrariamente a quanto si possa pensare questa é la parte piú difficile da scrivere poiché non ci sono formule nei cui termini nascondere le emozioni e forse perché una formula perfetta per dire grazie non é ancora stata plasmata. Penso che il primo ringraziamento sia doveroso nei confronti dei miei genitori, per i sacrifici che hanno fatto per permettermi di studiare, per aver investito tempo e denaro pur sapendo che non li avrei ripagati con niente di materiale e immediato: spero che il futuro lavorativo possa portarci soddisfazioni con cui ripagarvi. Non dimentico lo zio, la nonna, il nonno (purtroppo assente fisicamente per motivi di salute: questa tesi dedicata a te, che ti infonda coraggio contro la malattia) per l' esperienza e l'educazione che mi hanno donato. Ti devo citare tra i primi, Maddalena, per avermi aiutato a scrivere, in questa tesi e nella mia vita, quell' appendice che non compare scritta, ma che é stata fondamentale per me: quella fatta di sorrisi, incoraggiamenti, suggerimenti, infinita dolcezza. Mi hai permesso di riempire l'appendice della mia vita con figure che sono i tuoi sorrisi, formule che sono gli occhi con cui mi guardi: non c' é risultato sperimentale che possa competere con la felicitá di quando sono accanto a te e per tutte le conquiste che abbiamo ottenuto nella vita e che continueremo ad ottenere per costruire il futuro. Un grazie a tutti gli amici, di Alessandria e non, per essere una presenza costante, la cui presenza é piú solida degli algoritmi di b-tagging: grazie in particolare a Fabio per avermi trascinato nel mondo del volontariato permettendomi di aiutare ad accrescermi aiutando gli altri; un augurio a Lorenzo, per l'onore che avrá di studiare a Berlino e l' onere di rappresentare gli italiani all'estero. Un grazie a Pietro e Vincenzo per avere condiviso un anno di vita da casalinghi. Un ulteriore grazie a Vincenzo per le lunghe discussioni sulla Scienza, e sulla ricerca in Italia: gli auguro che il suo lavoro a Candiolo ora e nel futuro lo possa portare a rendere piú ricca di speranze la vita di chi soffre di cancro. A livello accademico il ringraziamento va a tutto il gruppo CMS Torino: per avermi accettato al loro interno, per aver fatto in modo che i miei errori diventassero punti su cui costruire e non da eliminare, per avermi fatto capire cosa vuole dire fare Ricerca. Un ringraziamento particolare ad

Ernesto, per essere stato una presenza costante e per i consigli di fisica e non. Un grazie a Marco Musich per aver saputo accettare le mie stupidaggini, sia in ambito di programmazione che fuori e per l' infinita disponibilitá dimostrata. Grazie a Stefano C. con cui ho condiviso l'ufficio e discussioni su Z+b. Un grazie generale al Corso di Studi in Fisica per aver rafforzato la consapevolezza che l'Universitá Italiana, sebbene maltrattata da chi invece dovrebbe tutelarla, sia tra le migliori in Europa: ricordiamoci che essere italiani vuole dire essere orgogliosi delle nostre istituzioni accademiche e anche minimamente sono fiero di aver contributo in un gruppo di ricerca italiano collaborando con l' INFN. Qualunque sia il futuro, grazie del passato e di quello che é ora.

e per tutti i ragazzi e le ragazze
che difendono un libro, un libro vero
cosí belli a gridare nelle piazze
perché stanno uccidendo il pensiero .....
perché le idee sono come farfalle
che non puoi togliergli le ali
perché le idee sono come le stelle
che non le spengono i temporali
perché le idee sono voci di madre
che credevano di avere perso
e sono come il sorriso di Dio
in questo sputo di universo
Roberto Vecchioni, Chiamami ancora amore.

Introduction

The main motivation, when the LHC (Large Hadron Collider) was built, was to investigate the hidden sectors of the Standard Model of particle physics. The main goal was to study and understand the electroweak symmetry breaking governed by the Higgs mechanism by searching for the intermediate boson or supersymmetric partners explaining the mechanism. These sectors of physics are available only at a high energy scale and therefore the LHC was designed to provide proton-proton collisions at a center of mass energy of 14 TeV. The collider started operating at $\sqrt{s} = 7$ TeV in 2011 and $\sqrt{s} = 8$ TeV in 2012. On 4 July 2012 the two LHC experiments, ATLAS and CMS, presented their results for the search of a Standard Model Higgs boson: both experiments observe a new particle in the mass region about 125-126 GeV consistent with the Higgs boson with a statistical significance of 5 standard deviation, based on the data collected in 2011 and 2012. The next step for the LHC experiments will be to determine the precise nature of this particle and its role for our understanding of the universe. The work presented in this thesis has been carried out within the Torino CMS group. It is devoted to the study of the algorithms used by the experiment to select heavy quarks (b-tagging algorithms) produced in pp collisions. In Chapter 1 there is a short introduction on the main features of the Standard Model of fundamental interactions with a section dedicated to searches of the Higgs boson. Chapter 2 contains a description of the characteristics of the CMS experiment, such as the experimental setup and the properties of the detectors. In Chapter 3 the reconstruction of the jets and the b-tagging algorithms of the CMS are explored, while Chapter 4 contains the study of the impact of the alignment of the tracker detector on b-tagging algorithms. In Chapter 5 there is an example of use of b-tagging algorithm in a electroweak analysis: the production of a Z boson in association with heavy quarks, a process which represents a reducible background in the Higgs boson search.

Chapter 1

The standard model of elementary particles

The standard model of elementary particles (SM) describes the matter as composed by twelve elementary particles called *fermions*, having half-integer spin. They can be divided into two groups, *leptons* and *quarks* as shown in Figure 1.1. Quarks do not exist as free states, but only as constituents of *hadrons* and they are subject to both strong and electroweak interactions; leptons instead interact only trough the electromagnetic force.

Fermions	1^{st} fam.	2^{nd} fam.	3^{rd} fam.	Charge	Interactions
Quarks	u	c	t	$+2/3$	All
	d	s	b	$-1/3$	
Leptons	e	μ	τ	-1	Weak, E.M.
	ν_e	ν_μ	ν_τ	0	Weak

Figure 1.1: Table of particles in the SM

The interactions between particles are described in terms of exchanging *bosons*, integer-spin particles which are carriers of the fundamental interactions. The main features of bosons and interactions are summarized in Figure 1.2. The *gravitational* force is not taken into account in the standard model because it is not relevant at the typical mass and distance scale of particle physics.

The SM describes these interactions in terms of two *gauge* theories: the Quantum Chromo-Dynamics (QCD) and the theory of electroweak interaction (GWS model), which unifies the electroweak and weak interactions. In all the following descriptions, as usual, in high energy physics, we work in natural unit with $\hbar$=c=1.

	Electromagnetic	Weak	Strong
Quantum	Photon (γ)	$W^{\pm}$, Z	Gluons
Mass [GeV/c^2]	0	80, 90	0
Coupling constant	$\alpha(Q^2 = 0) \approx \frac{1}{137}$	$\frac{G_F}{(\hbar c)^3} \approx 1.2 \cdot 10^{-5}$ GeV^{-2}	$\alpha_s(m_Z) \approx 0.1$
Range [cm]	∞	10^{-16}	10^{-13}

Figure 1.2: Basic properties of fundamental interactions described by the standard model.

1.1 The Electroweak interactions

The gauge theory for weak interactions is conceived as an extension of the theory of electromagnetic interactions, the Quantum Electro-Dynamics (QED), which is based on the group U(1), associated to the conserved quantum number Q (*electric charge*). The condition of local invariance applied on this group leads to the existence of a massless vector, the *photon*. Fermions can be divided in doublets of negative-chirality (*'left-handed'*) particles and singlets of positive-chirality (*'right-handed'*) particles in the massless limit, as shown in Figure 1.3. The unification of electroweak and electromagnetic

	I_3	Y	Q
$\begin{pmatrix} u_L \\ d_L \end{pmatrix}$	$\begin{pmatrix} 1/2 \\ -1/2 \end{pmatrix}$	$\begin{pmatrix} 1/3 \\ 1/3 \end{pmatrix}$	$\begin{pmatrix} 2/3 \\ -1/3 \end{pmatrix}$
u_R, d_R	0, 0	$4/3$, $-2/3$	$2/3$, $-1/3$
$\begin{pmatrix} \nu_{\ell,L} \\ \ell_L \end{pmatrix}$	$\begin{pmatrix} \frac{1}{2} \\ -\frac{1}{2} \end{pmatrix}$	$\begin{pmatrix} -1 \\ -1 \end{pmatrix}$	$\begin{pmatrix} 0 \\ -1 \end{pmatrix}$
ℓ_R	0	-2	-1

Figure 1.3: Definitions of fermions by their chirality in the standard model.

interactions is achieved extending the gauge symmetry to the group $SU(2)_L \times U(1)_Y$. The generators of $SU(2)_L$ are the three components of the *weak isospin* operator $t^a = \frac{1}{2}\tau^a$ where τ^a are the Pauli matrices. The generator of $U(1)_Y$ is the *weak hypercharge* Y operator. An important relation between

quantities is:

$$Q = I_3 + \frac{Y}{2}$$

where I_3 is the third component of the weak isospin, an eigenvalue of t^3. The requirement of local gauge invariance with respect to the $SU(2)_L \times U(1)_Y$ introduces four massless vector fields, $W_\mu^{1,2,3}$ and B_μ which couple to fermions with two different coupling constants g and $g^{'}$. $W_\mu^{1,2,3}$ and B_μ do not represent physical fields, which are given instead by linear combinations of the four fields: the charged bosons W^+ and W^- correspond to

$$W_\mu^\pm = \sqrt{\frac{1}{2}}(W_\mu^1 \mp iW_\mu^2)$$

while the neutral bosons γ and Z correspond to

$$A_\mu = B_\mu \cos\theta_W + W_\mu^3 \sin\theta_W$$

$$Z_\mu = -B_\mu \sin\theta_W + W_\mu^3 \cos\theta_W$$

obtained by mixing the neutral fields W_μ^3 and B_μ with a rotation defined by the *Weinberg angle* θ_W. A_μ can be identified with the photon field. The equivalence between the coupling constants is:

$$g \sin\theta_W = g^{'} \cos\theta_W = e$$

which represents the electroweak unification. The GWS model thus predicts the existence of two charged gauge fields, which only couple to left-handed fermions and two neutral gauge fields which interact with both left and right-handed components.

1.2 The Higgs Mechanism

In order to reproduce the phenomenology of the weak interactions, both fermion and gauge boson fields must acquire mass. In the electroweak Lagrangian a mass term for the gauge bosons would violate gauge invariance which ensures the theory can be renormalized. Masses are introduced using the *Higgs mechanism* which allows fermions and $W^\pm, Z$ bosons to be massive, keeping the photon massless. This mechanism is obtained introducing a doublet of complex scalar fields:

$$\Phi = \begin{pmatrix} \Phi^+ \\ \Phi^0 \end{pmatrix} = \frac{1}{\sqrt{2}} \begin{pmatrix} \Phi^1 + i\Phi^2 \\ \Phi^3 + i\Phi^4 \end{pmatrix}$$

which is introduced in the electroweak Lagrangian with the term:

$$\mathcal{L}_{EWSB} = (D^{\mu}\Phi)^{+}(D_{\mu}\Phi) - V(\Phi^{+}\Phi)$$

where D_{μ} is the covariant derivative.
The potential in the form:

$$V(\Phi^{+}\Phi) = -\mu^2\Phi^{+}\Phi - \lambda(\Phi^{+}\Phi)^2$$

with $\mu^2 < 0$ and $\lambda > 0$ has a minimum for

$$\Phi^{+}\Phi = \frac{1}{2}((\Phi_1)^2 + (\Phi_2)^2 + (\Phi_3)^2 + (\Phi_4)^2) = -\frac{\mu^2}{2\lambda} = \frac{v^2}{2}$$

The minimum is not for a single value of Φ but for a manifold of non-zero values. The choice corresponds to the ground state is arbitrary and the chosen point is not invariant under rotations in the (Φ^{+}, Φ) plane: this is referred to as *spontaneous symmetry breaking*. If one chooses to fix the ground state on the Φ^0 axis, the vacuum expectation value of the Φ field is:

$$\Phi = \frac{1}{\sqrt{2}}\begin{pmatrix}0\\v\end{pmatrix}$$

where

$$v^2 = -\frac{\mu^2}{\lambda}$$

The Φ field can be rewritten in a generic gauge, in terms of its vacuum expectation value:

$$\Phi = \frac{1}{\sqrt{2}}\exp\frac{i\Phi^a t_a}{v}\begin{pmatrix}0\\H+v\end{pmatrix}, a = 1,2,3$$

where the tree fields Φ^a and the fourth $\Phi^4 = H + v$ are called *Goldstone fields*. Being scalar and massless they introduce four new degrees of freedom. The electroweak Lagrangian, results then from the sum of three terms:

$$\mathcal{L}_{EWSB} = \mathcal{L}_H + \mathcal{L}_{HW} + \mathcal{L}_{HZ}$$

where the three terms can be written as follows, using the approximation $V = \mu^2 H^2 + \text{const}$:

$$\mathcal{L}_H = \frac{1}{2}\partial_a H \partial^a H + \mu^2 H^2$$

$$\mathcal{L}_{HW} = \frac{1}{4}v^2 g^2 W_{\alpha}^{-} W^{+\alpha} + \frac{1}{2}v g^2 H W_{\alpha}^{-} W^{+\alpha} = (m_W)^2 W_{\alpha}^{-} W^{+\alpha} + g_{HW} H W_{\alpha}^{-} W^{+\alpha}$$

$$\mathcal{L}_{HZ} = \frac{1}{8}v^2(g^2+g'^2)Z_\alpha Z^\alpha + \frac{1}{4}v(g^2+g'^2)HZ_\alpha Z^\alpha = \frac{1}{2}(m_Z)^2 Z_\alpha Z^\alpha + \frac{1}{2}g_{HZ}HZ_\alpha Z^\alpha$$

Each of the three gauge bosons has acquired mass and an additional degree of freedom, corresponding to the longitudinal polarization. Only the H scalar field is still present and has acquired mass itself: that is the Higgs field. In a nutshell we can summarize the Higgs mechanism: it was used to give mass to the weak bosons without explicitly breaking the gauge invariance, which implies preserving the renormalisability of the theory. When a symmetry is *spontaneously* broken it is not properly eliminated, but hidden by the choice of the ground state: the minimum of the Higgs field is still invariant under the U(1) group. The electromagnetic symmetry is not broken and in fact the photon does not couple to the Higgs boson and remains massless. The masses of the vector bosons $W^\pm$ and Z are related to the parameter v and to the electroweak coupling constants:

$$\frac{m_W}{m_Z} = \frac{g}{\sqrt{g^2+g'^2}} = \cos\theta_W$$

The electroweak symmetry breaking energy scale can be determined from the relation between the v parameter and the Fermi constant G_F:

$$v = (\frac{1}{\sqrt{2}G_F})^{\frac{1}{2}}$$

The Higgs mechanism is also used to generate the masses of the fermions, by introducing in the Lagrangian a term called *Yukawa term*, which represents the interaction between the Higgs and the fermion fields. The mass of a fermion (apart from neutrinos) and its coupling constant to the Higgs boson are:

$$m_f = \frac{G_{Hf}}{\sqrt{2}}v$$

$$g_{Hf} = \frac{G_{Hf}}{\sqrt{2}} = \frac{m_f}{v}$$

Being G_{Hf} free parameters, the mass of the fermions cannot be predicted by the theory.

The Higgs boson mass (m_H) is still an unknown free parameter of the SM. It has not been observed experimentally and it depends on the parameter v and λ, but while the former can be estimated by the relation with G_F, the latter is characteristic of the field Φ and cannot be determined other than measuring the Higgs mass itself.

1.2.1 Theoretical Constraints

Theoretical constraints on the Higgs mass can be found imposing the energy scale Λ up to which the SM is valid, before the perturbation theory breaks down. The upper limit is obtained requiring that the running quartic coupling of Higgs potential λ remains finite up to the scale Λ. A lower limit is found requiring that λ remains positive after the inclusion of radiative corrections at least up to a scale Λ. These theoretical constraints are summarized in the Figure 1.4.

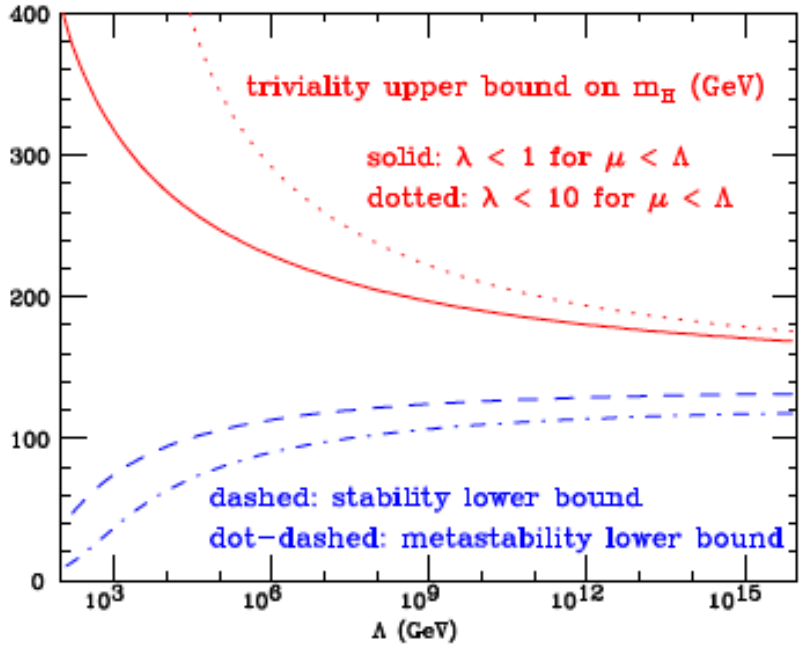

Figure 1.4: Theoretical constraints for the Higgs boson mass

If the validity of the SM is assumed up to the Plank scale ($\Lambda = 10^{19}$ GeV), the allowed Higgs mass range is between 130 and 190 GeV, while for $\Lambda = 1$ TeV the Higgs mass can be as high as to 700 GeV.

1.2.2 Experimental Constraints before year 2011

Until the 2011 bound on the Higgs mass were provided by measurements at LEP, SLC, and Tevatron. Searches at LEP-II set a lower limit with 95 % confidence level (CL) of 114.4 GeV and results from the Tevatron experiments excluded the mass range from 158 to 175 GeV. Moreover, constraints on

the Higgs boson mass were extracted indirectly from the measurements of other electroweak observables, which have a logarithmic dependence on m_H through the radiative corrections. All electroweak measurements performed by the four LEP experiments were combined together and fitted, assuming the Higgs mass as a free parameter and assuming the SM as the correct theory. The result is summarized in Figure 1.5 where the $\Delta\chi^2$, defined as $\Delta\chi^2 = \chi^2 - \chi^2_{min}$, is plotted as a function of m_H. The black solid curve is the result of the fit, while the blue band represents the theoretical uncertainty due to higher order corrections, while the yellow areas are the exclusion by measurements, underlying that these results are model-dependent.

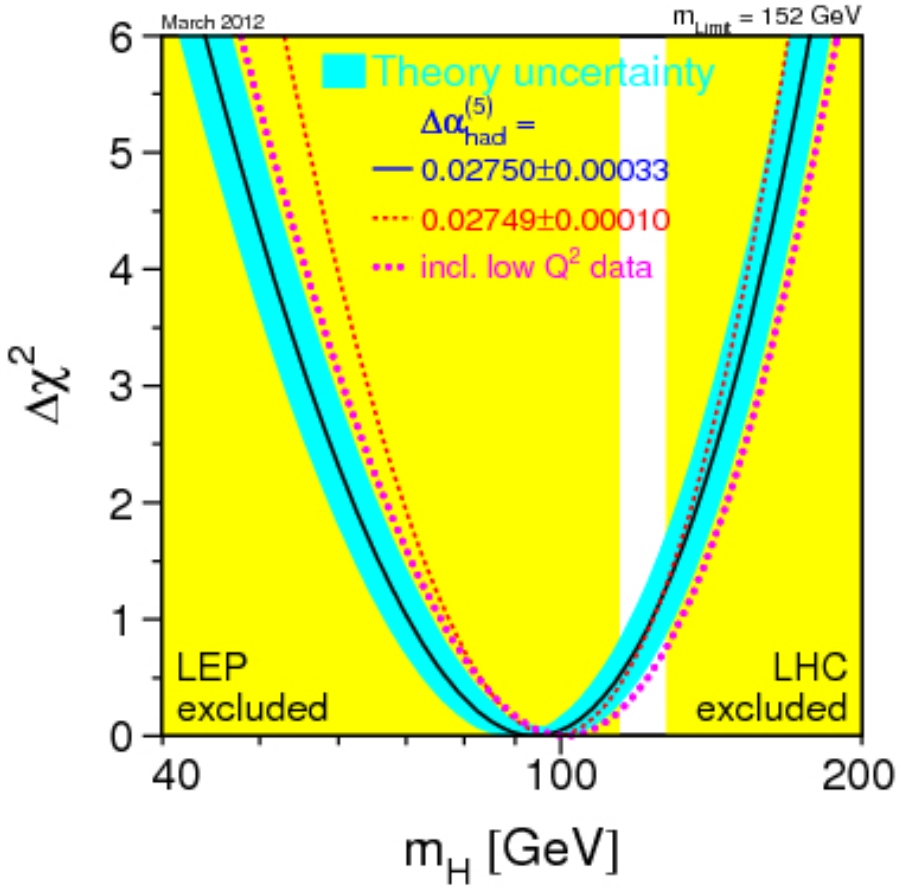

Figure 1.5: The preferred Higgs boson mass from the SM consistency test.

1.3 Higgs boson searches at LHC

The main processes which contribute to the Higgs production in a proton-proton collision at the energy scale reached by LHC are shown in Figure 1.6. The fundamental interactions occur between quarks and gluons: gluon-gluon fusion is the dominant process, while only at very high masses the vector boson fusion (VBF) dominates. Cross sections are typically of the order of few picobarns.

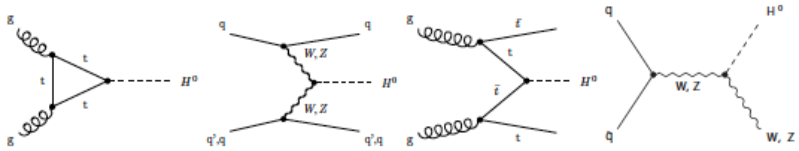

Figure 1.6: Processes for Higgs boson production in a proton-proton collision.

The WW scattering has a cross section 10 times smaller than gluon-gluon fusion, but it has a clear experimental signature such as the presence of two spectator jets, the associate production instead (third diagram in Figure 1.6) has very small cross section except for very low m_H. Once produced, the Higgs can decay in different ways shown like branching ratios as a function of the Higgs mass in Figure 1.7.

The main dacay channels can be divided into three groups according to the different Higgs mass:

- low mass Higgs ($m_H < 130$ GeV): the dominating channel is $H \to b\bar{b}$, even if this is difficult to observe at the LHC because of the huge QCD background. The most clear channel is $H \to \gamma\gamma$ which has a very clear signature, but requires excellent π^0 rejection;

- intermediate mass Higgs ($130 < m_H < 180$ GeV): the branching ratio is high but purely hadronic final states are not accessible. The Higgs decay in four leptons is the golden channel because it has a clean final state and suffers only partially from irreducible background;

- high mass Higgs ($m_H > 180$ GeV): the cross section becomes small and the Higgs width becomes very broad so that the reconstruction of a mass peak becomes difficult.

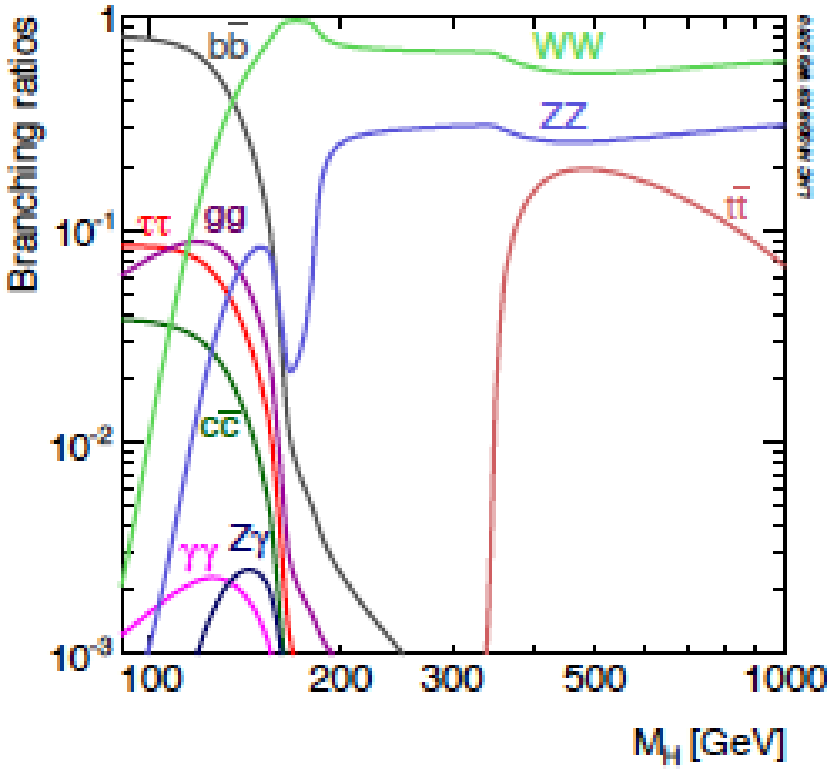

Figure 1.7: Branching ratios for Higgs decay as a function of the Higgs mass.

1.3.1 Recent results

In Figure 1.8 is shown the state of the Higgs searches in July 2011 (before the first results from LHC experiments): 95% confidence level upper limits on the Higgs boson cross section production as a function of m_H are shown. The solid (dashed) line represents the observed (expected) ratio between the excluded cross section and the SM one. The green and yellow bands indicate, respectively the 68% and 95% probability regions for a fluctuation of the expected limit. The regions where the line is below unity represent the mass ranges where the SM Higgs boson is excluded at 95% CL.[1]

In Figure 1.9 is shown the situation up to December 2011. Analyses by the ATLAS and CMS collaborations had further narrowed the range of masses available for the Standard Model Higgs particle to 116-127 GeV.

[1]Evidence for a Higgs boson of 125 GeV has been reported from CMS at the beginning of July 2012.

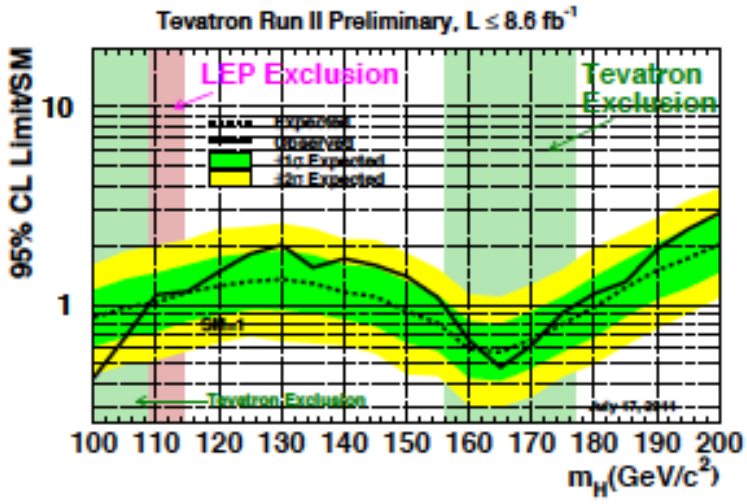

Figure 1.8: 95% CL upper limits on the Higgs boson cross section production.

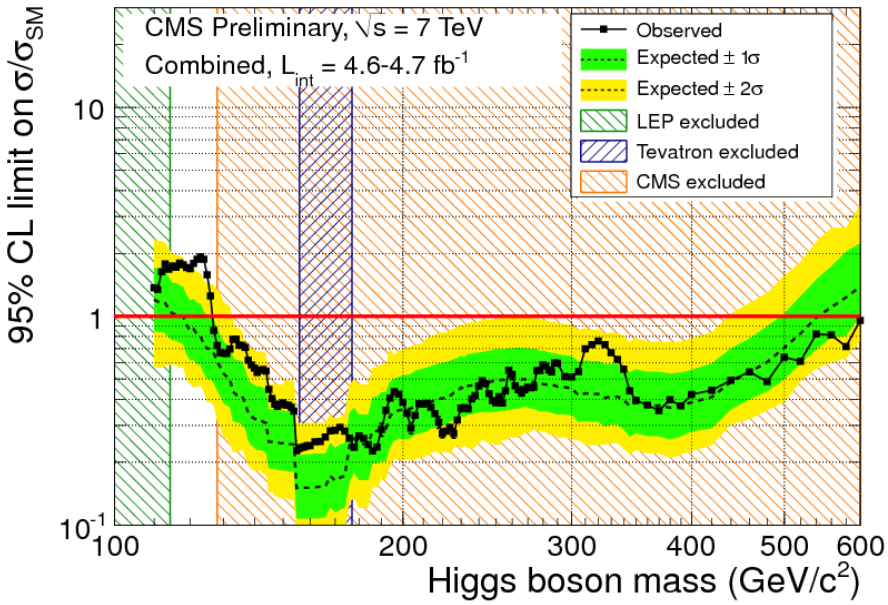

Figure 1.9: 95% CL exclusion limit on the mass of the standard model Higgs boson. The analysis is based on proton proton data collected by CMS in 2010 and 2011.

Chapter 2

The Large Hadron Collider

The Large Hadron Collider (LHC) accelerator was built to explore the hidden sectors of the standard model and investigate possible new physics processes. The main features of the machine are:

- hadron collider: the fundamental constituents entering in the scattering are partons carrying a fraction x of the four momentum of the particles in the beam. The design center-of-mass energy of LHC for proton-proton collisions is $\sqrt{s} = 14$ TeV.[1] With respect to an electron-positron machine it is easier to accelerate protons to high energy since the energy loss for synchrotron radiation is much lower than electrons;

- a high luminosity collider: the event rate of a process is proportional to the cross section by a factor $\mathcal{L}$ called luminosity which represents the numbers of collisions per unit time and cross-sectional area of the beams depending on the collider parameters.

To compensate for the low cross sections of processes of interest, LHC must have a very high luminosity: the nominal very short bunch crossing interval (25 ns) and the high number of bunches accelerated (2808 per beam) will allow to reach the peak luminosity 10^{34} cm $^{-2}$ s $^{-1}$. The design parameters of the machine are summarized in the Figure 2.1. The Figure 2.2 instead summarizes the cross sections and the production rates at a luminosity of 2 10^{33} cm $^{-2}$ s $^{-1}$. The total proton-proton cross section at 7 TeV is approximately 110 mb. This total can be broken down in contributions from:

- inelastic (60 mb);

- single diffractive (12 mb);

[1] LHC started operating at $\sqrt{s} = 7$ TeV (2010-2011) and $\sqrt{s} = 8$ TeV (2012). From 2015 LHC is going to run at $\sqrt{s} = 14$ TeV.

- elastic (40 mb);

The major part of the inelastic collisions are processes with a low p_T exchange, the so called *minimum bias* events.

Parameters	p-p	Pb-Pb
Circumference (km)	26.659	
Centre of mass energy (TeV)	14	1148
Dipole magnetic field (T)	8.3	
Number of particles per bunch	2808	608
Bunch length (mm)	53	75
Bunch crossing rate (MHz)	40.08	0.0006
Design Luminosity ($cm^{-2}s^{-1}$)	10^{34}	2×10^{27}
Beam radius at interaction point (μm)	15	

Figure 2.1: Design parameters for LHC machine

Process	σ	Events/sec	Events/year
$W \rightarrow e\nu$	20 nb	15	10^8
$Z \rightarrow ee$	2 nb	1.5	10^7
$t\bar{t}$	1 nb	0.8	10^7
$b\bar{b}$	0.8 mb	10^5	10^{12}
$\tilde{g}\tilde{g}$ (m= 1 TeV)	1 pb	0.0001	10^4
H (m = 0.8 TeV)	1 pb	0.001	10^4
H (m=0.2 TeV)	20 pb	0.01	10^5

Figure 2.2: Cross sections and production rates for nominal luminosity at LHC.

Figures 2.3 and 2.4 show the total integrated luminosity delivered by LHC and recorded by CMS in 2010 and 2011.

2.1 The Compact Muon Solenoid experiment

The two main multi-purpose experiments designed specifically to detect and discover new physics are the A Toroidal LHC ApparatuS (ATLAS) and

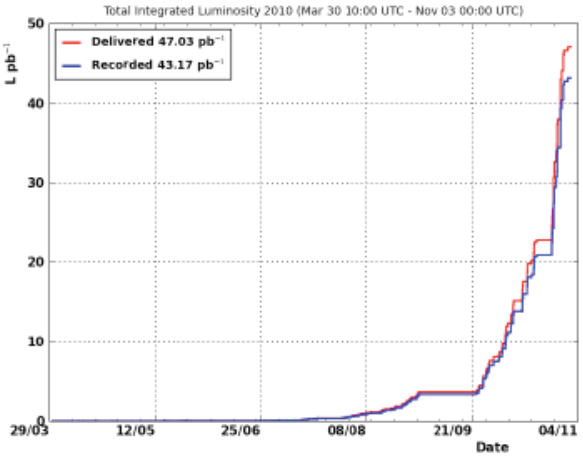

Figure 2.3: Integrated luminosity delivered by LHC (in red) and recorded by CMS (in blue) in proton proton collisions during 2010.

the Compact Muon Solenoid (CMS) experiments. The main difference between the two experiments is in the configuration of the magnetic field: ATLAS uses a toroidal field, while CMS a solenoidal field.[2]

The most important requirements for CMS detector are:

- muons: good identification and momentum resolution up to $|\eta| < 2.5$ and capability to the determine the charge of muons up to 1 TeV;
- charged particles: good momentum resolution and reconstruction efficiency in the tracking system together with an efficient triggering and offline tagging;

[2]CMS uses a system of cartesian coordinates centered in the interaction point and with the z axis tangent to the beam line. The x axis is chosen to be in the orbit plane and pointing towards the center of the ring, and the y axis is perpendicular to the plane of the orbit and pointing upwards. The direction of the z axis is referred as *longitudinal*, while the $x - y$ plane is called *transverse*. The *rapidity* is defined as:

$$y = \frac{1}{2} \log \frac{E + p_z}{E - p_z}$$

For an ultra-relativistic particle y can be good approximated to the *pseudo rapidity*

$$\eta = -\log(\tan \frac{\theta}{2})$$

where θ is the angle between the particle momentum $\mathbf{p}$ and the z axis.

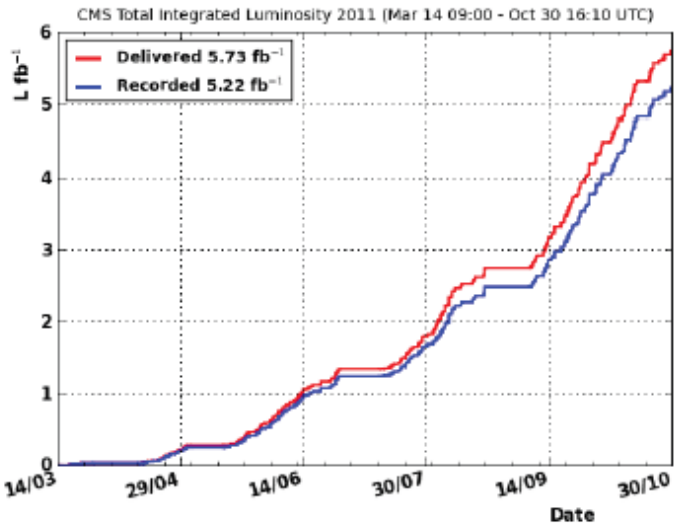

Figure 2.4: Integrated luminosity delivered by LHC (in red) and recorded by CMS (in blue) in proton proton collisions during 2011.

- electrons and photons: good energy resolution (both di-photon and di-electron mass resolution), good π^0 rejection;
- good missing energy and dijet mass resolution using large hadron calorimeters with a large hermetic geometric coverage ($|\eta| < 5$) with lateral segmentation.

A global view of the experiment is shown in Figure 2.5.

The detector structure consists of a cylindrical barrel closed by two endcap disks. The overall length is 21.6 m, the diameter is 14.6 m and the total weight is about 12 500 tons. The thickness of the detector in radiation lengths is greater than 25 X_0 for the electromagnetic calorimeter and 11 λ_1 for the hadronic calorimeter. The choice of a solenoidal magnetic field led to a very compact design for the CMS system, with a strong improvement in the detection and energy measurement of electron and photons. The core of the detectors is shown in Figure 2.6.

First of all there is the the big magnet, a 13 m long super-conductive solenoid cooled with liquid helium, which can provide a 4 T magnetic field. The magnet coil with a diameter of 5.9 m contains the Tracker and the electromagnetic and hadronic calorimeters. The iron return yoke of the magnet hosts the muon spectrometer composed by 4 stations of drift tube detec-

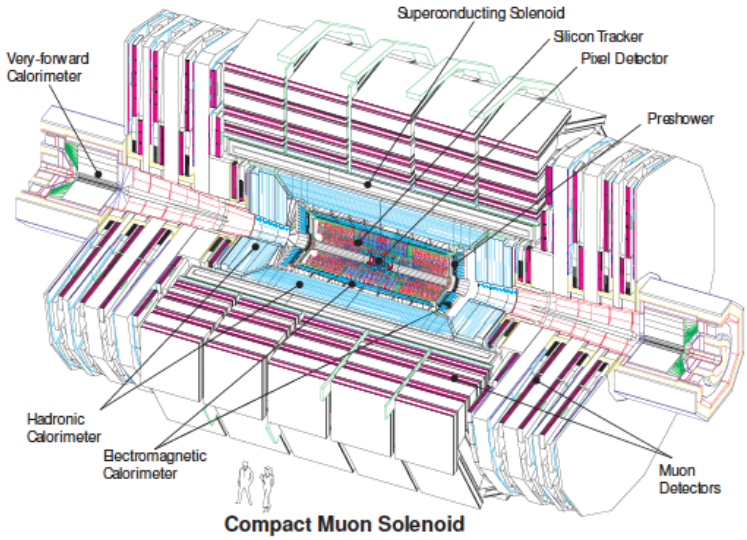

Figure 2.5: CMS experiment layout.

tors (DT) in the barrel and 4 stations of cathode strip chambers (CSC) in the endcap. Both barrels and endcaps are also equipped with resistive plate chambers (RPC). The average coverage in pseudo rapidity is up to $|\eta| = 2.4$. The hadronic calorimeter (HCAL) is a brass/scintillator sampling calorimeter: the barrel and endcap parts are complemented by a forward calorimeter. Inside HCAL, the electromagnetic calorimeter (ECAL) is installed: it is an homogeneous calorimeter made of lead-tungstate scintillating crystals with a coverage of $|\eta| < 3.0$. The tracking detector was built under the requirement of a precise vertex reconstruction and b-tagging with very high track multiplicity: this was achieved with a very fine segmentation. CMS decided in the barrel to use 10 layers of silicon microstrip detectors and three layers of silicon pixel detectors close to the interaction region in order to improve the measurement of the impact parameter of charged particle tracks. The tracking system allows charged particle tracks reconstruction up to 14 measurement points and a coverage of $|\eta| < 2.5$.

2.1.1 The pixel tracker

The pixel Tracker, a view is shown in Figure 2.7, consists of three 53.3 cm long barrel layers (BPIX) and two endcap disks (FPIX) on each side of the barrel section. The innermost layer of the BPIX is at radius of 4.4 cm, the second 7.3 cm and the third layer 10.2 cm. The layers are composed

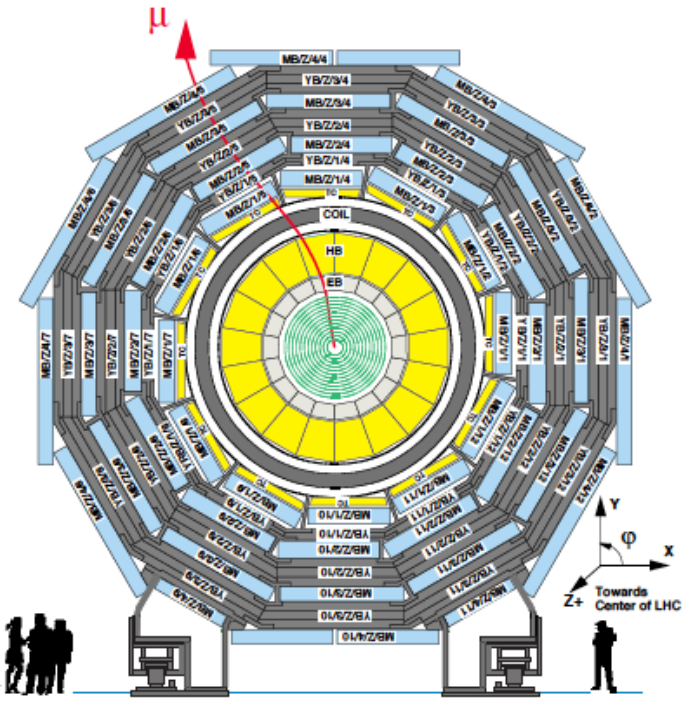

Figure 2.6: Transverse view of the CMS experiment.

by modules[3] placed on carbon fiber supports (called *ladders*), hosting eight modules, consisting of 258 μm thick segmented silicon sensors with integrated read-out. The BPIX region is composed by 672 full modules and 96 half modules with 66560 pixels per module for the full modules and 33280 for the half modules. The FPIX endcap disks are placed at $z = \pm 35.5$ cm and $z = \pm 48.5$ cm, split into half disks, each including 12 trapezoidal blades forming a turbine-like geometry. The endcap disks are made of 672 plaquettes (270 μm thick), for a total of 17971200 pixels. In localizing secondary decay vertices, both transverse and longitudinal coordinates are important and a nearly square pixel shape (100×150 μm^2) is adopted and an analog charge readout is implemented.

2.1.2 The strip tracker

Outside the pixel detector, the Tracker is composed of silicon microstrip detectors, as shown in Figure 2.8.

[3] A module is the mechanical structure on which the sensors are directly mounted.

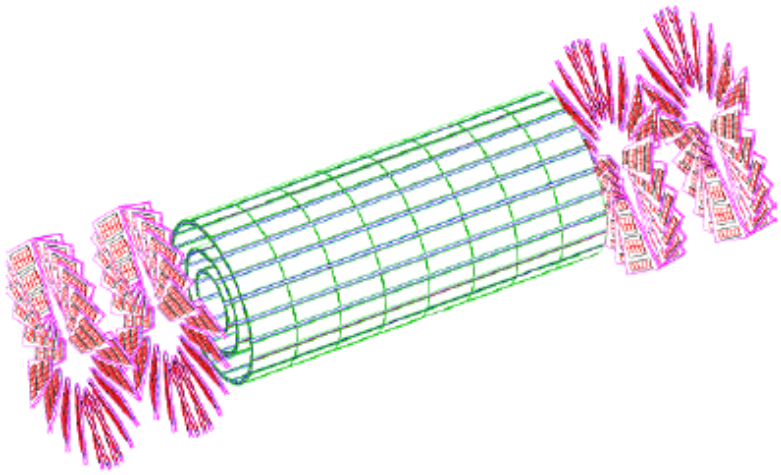

Figure 2.7: View of the CMS pixel tracker.

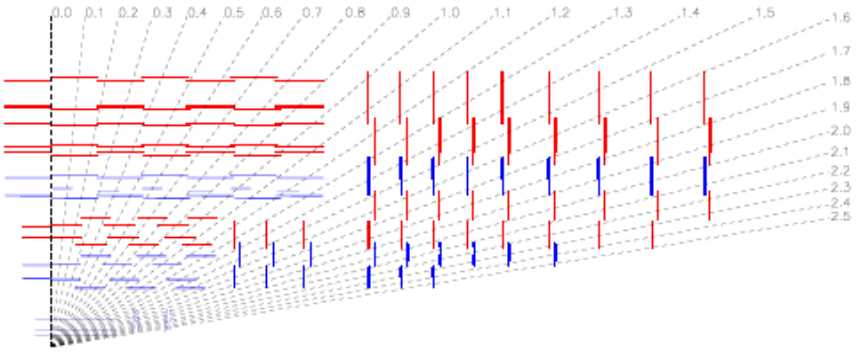

Figure 2.8: Quarter view (rz) of the CMS strip tracker, blue modules are double-sided, while red are one sided.

The barrel region is divided into two parts: the Tracker Inner Barrel (TIB) and the Tracker Outer Barrel (TOB). The first is composed by four layers of silicon sensors with a thickness of 320 μm and strip pitches varying from 80 to 120 μm. The TOB is made of six layers, using thicker silicon sensors (500μm): the strip pitch varies from 120 to 180 μm. The end cap region ($|\eta| > 1.6$) is covered by the Tracker Inner Disks (TID) and the Tracker End Cap (TEC). The three disks of the TID fill the gap between the TIB and the TEC. The TEC while has nine disks extending into the region in $|z|$ from 120 to 280 cm. The first two layers of the TIB and the TOB, the first two rings of the TID and of the TEC are made by modules are glued back-to-back to

have a bidimensional point of measure.

2.1.3 The electromagnetic calorimeter

The Electromagnetic Calorimeter (ECAL) is made of 74 848 lead-tungstate crystals: this material is characterized by a high density (8.28 g/cm ^{3}), giving it a very compact form allowing to be placed inside the magnetic coil. It also has a very small Moliere radius (2.2 cm) which gives ECAL the possibility to contain very well the expansion of the electromagnetic shower. These crystals are also characterized by a very short scintillation decay time which allows to be used at a crossing rate of 40 MHz. ECAL barrel covers the central rapidity region ($|\eta| < 1.48$) and two ECAL end caps up to $|\eta| = 3$. The barrel crystals have a front face area of 2.2 $\times$ 2.2 cm^2, 23 cm length and they are positioned at $r = 1.29$ m. Crystals in the end caps have 2.47 $\times$ 2.47 cm^2 front face, 22 cm length positioned at $z = 3.17$ m. For trigger purpose the ECAL crystals are grouped together into 68 *trigger towers*, in the endcaps. A *pre-shower* device with higher granularity is used to distinguish between showers started from neutral pions and photons or charged pions and electrons. Silicon Avalanche Photodiodes (APDs) and Vacuum Phototriodes (VPTs) are used to collect the scintillation light in the barrel and in the endcaps respectively.

2.1.4 The Hadronic Calorimeter

The hadronic calorimeter (HCAL) is built to identify hadrons and neutrinos by measuring the energy and direction of jets and missing transverse energy in the event. The HCAL angular coverage must include the very forward region, because the identification of forward jets is very important for CMS. The HCAL is divided into four regions which provide a full angular coverage up to $|\eta| = 5$. The *barrel hadronic calorimeter* (HB) surrounds the electromagnetic calorimeter and covers up to $|\eta| = 1.3$, while two *endcap hadron calorimeters* (HE) cover up to $|\eta| = 3.0$. To satisfy the hermeticity requirements there are two *forward hadronic calorimeters* (HF) up to $|\eta| = 5.0$, in the surround of the beam pipe at $z = 11$ m. At the end there is an array of scintillators outside the magnet called *outer hadronic calorimeter* (HO) to improve the central shower containment. The HB and HE are sampling calorimeters with active plastic scintillators interleaved with brass plates, while the readout is composed of wavelength-shifting fibres. The energy resolution (with E in GeV) is:

- $\frac{\sigma_E}{E} = \frac{65\%}{\sqrt{E}}$ in the barrel;

- $\frac{\sigma_E}{E} = \frac{85\%}{\sqrt{E}}$ in the endcaps;
- $\frac{\sigma_E}{E} = \frac{100\%}{\sqrt{E}}$ in the forward part;

2.1.5 The muon system

The muon spectrometer is placed outside the magnet and it has fundamentally two goals:

- muon momentum measurement in combination with the silicon tracker;
- muon trigger as high p_T muons are a signature for many physics processes.

Three different detection technologies are used: the barrel muon system uses *drift tube chambers* (DTs), the endcaps use *cathode strip chamber* (CSCs) and both are complemented with *resistive plate chambers* (RPCs) for triggering purposes.

Drift Tubes Chambers

Each chamber is made up of twelve layers of DTs, placed into three independent subunits called *super-layers*. Two of them measure the transverse coordinate, and the remaining one is orthogonal to measure the longitudinal coordinate. They are made of parallel aluminum plates, the anodes are 50 μm diameter steel wires placed between the cathodes. The internal volume is filled with a gas mixture of Ar (85%) and CO_2 at atmospheric pressure. The single hit position resolution is about 260 μm at nominal voltage values with an angular resolution of 1.8 mrad.

Cathode Strip Chambers

The CSCs are multi-wire proportional chambers with one cathode plane segmented in strips. The orthogonal orientation allows to have the determination of coordinates from a single plane. Each chamber is formed by six trapezoidal layers: wires have a resolution on the measured position of 0.5 cm, while the strips of 50 μm.

Resistive Plate Chambers

The RPC are installed both in endcaps and in barrels: they have a poor spatial resolution, but a very fast time resolution (3 ns), which makes them well suited for being used for trigger purposes. The RPCs used by CMS are

double gap RPCs, made of four bakelite planes formed two gaps filled with a $C_2H_2F_4$ and C_4H_{10} gas mixture operating in avalanche mode.

2.2 The Trigger System

The expected event rate at the nominal luminosity is 10^9 Hz, the limit imposed in storing and processing data is 100 Hz, so the goal of the CMS trigger system is to reduce the data by a factor 10^7. CMS adopts a multi-level trigger design using a little part of available data, so high-level triggers can use more refined algorithms. The trigger is divided into two physical levels: *Level-1 Trigger (L1)* and the *High Level Trigger* (HLT), The L1 is basically hardware and uses only segmented data from calorimeters and muon detectors. The structure is summarized in Figure 2.9.

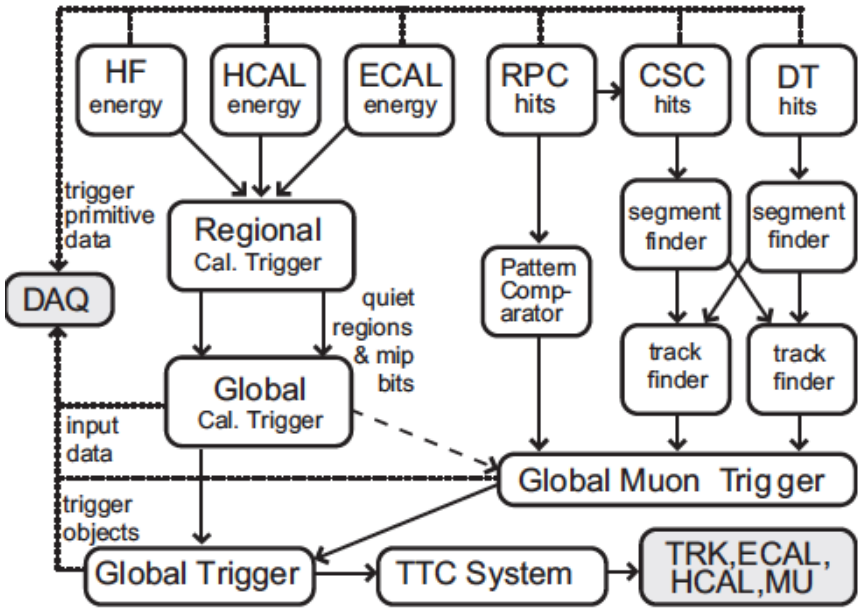

Figure 2.9: L1 trigger scheme

The HLT is basically software implemented in a computer farm organized into growing complexity logic levels and it processes all the events accepted by the L1 trigger. The basic idea is to reconstruct those part of each physics object that can be used for selection while minimizing the overall CPU usage. As a convention the term Level-2 Trigger is used to refer to algorithms and requirements of the first selection step in the HLT process. In contrast, Level-3 Trigger refers to a further selection step that includes the reconstruction of

full tracks in the tracker. To minimize the CPU time required by the HLT, a feature of the algorithms is to reconstruct the information in the CMS detector only partially, in fact in many cases the decision on whether an event should be accepted by the HLT involves the reconstruction of quantities in only a limited region of the detector.

(Unprescaled) Object	Trigger Threshold (GeV) @ 6E33
Single Muon	40
Single Isolated muon	24
Double muon	(17, 8) [13, 8 for parked data]
Single Electron	80
Single Isolated Electron	27
Double Electron	(17, 8)
Single Photon	150
Double Photon	(36, 22)
Muon + Ele x-trigger	(17, 8), (5, 5, 8), (8, 8, 8)
Single PFJet	320
QuadJet	80 [50 for parked data]
Six Jet	(6 x 45), (4 x 60, 2 x 20)
MET	120
HT	650

Figure 2.10: HLT trigger scheme in 2012.

Chapter 3

Reconstruction of the physics objects in CMS

3.1 Jets reconstruction

Since gluons and quarks could not exist into free state, they fragment into hadrons resulting into a jet of particles depositing energy in the detectors. There is a wide range of algorithms for jet reconstruction which could take also into account biases coming from detectors effects (jet energy calibration). Cone-based algorithms generally define a metric $\Delta R = \sqrt{(\eta_1 - \eta_2)^2 + (\phi_1 - \phi_2)^2}$ based on the pseudorapidity and the azimuthal angle, while k_T based algorithms use a E_T metric. The input objects of jets clustering can be of different types: traditionally, jet have been built by taking the sum of the energies deposited in the hadronic and electromagnetic calorimeters. The latest jets clustering algorithms use all the informations available from all detectors and subdetectors to measure the energy and the momentum of each particle which leaves a signal: this approach, extensively used in the CMS experiment, is called *particle flow*. There are two possibilities to add four momenta when performing a clustering process: the first one is adding the jet constituents as four vectors (*massive jets*), the second is adding the sum of the transverse energy E_T of the constituents (*non massive jets*).
The schemes for two different algorithms are briefly described below:

- **Seedless infrared safe cone algorithm**: it uses any input object, that is not only those above a certain E_T threshold, to look for the most stable cone. To find the stables cones the four-momenta of the input objects into a cone are added together and the direction of the cone is compared to the summed four-momentum of the input objects

enclosed in this cone. A cone is considered stable when the direction of the cone agrees with the direction of the three momenta of the input objects. Then it is added to a list of *candidate jets*. All these candidates are ordered in descending transverse momentum and the ones below a minimum threshold are discarded. Then the highest p_T candidate jet is selected and the highest p_T candidate jet that shares objects with the first candidate jet is looked for. If no matching is found the candidate jet is removed from the list and added to the jets list. If the overlapping fraction is above a certain threshold (f), the two candidate jets are merged instead. This process is repeated iteratively until no candidate jet remains.

- **Inclusive k_T algorithm**: for each input object i and for each pair of input objects i,j two quantities (*distances*) are calculated:

$$d_i = E_{T,i}^2 R^2$$

$$d_{ij} = min(E_{T,i}^2, E_{T,j}^2 R_{ij}^2)$$

where R is an adimensional parameter and R_{ij} is the angular distance defined before. The algorithm checks the smallest value among d_i and d_{ij}: if it is d_i the first object is removed from the list of candidates and inserted into the list of jets, while if it is d_{ij} the two candidate jets are merged. The procedure is repeated until all jets are found.

3.2 The tagging of the flavor of the jets

Hadrons containing b and c quarks B-hadrons and D-hadrons in the following, have a long lifetime ($c\tau = 500$ μm) and carry away a significant fraction of the energy of the parent parton. Their decay produces tracks significantly displaced with respect to the primary vertex as they originated from the so called *secondary vertex*. Another hint for the presence of a B or a D-hadron could be the presence of a soft lepton coming from the semileptonic decay. The algorithms which attempt to reconstruct these properties in terms of heavy quark identification are called b-tagging algorithms. The main idea of an algorithm is to associate a floating point number to a jet, where conventionally large values are associated to a jet originating from a heavy hadron while small values are associated to jets originating from light partons. We can classify these algorithms according to the set of variables used to tag the hadron.

3.2.1 Impact parameter based algorithms

The association of tracks to the jets is based on a ΔR criterion: tracks with $\Delta R < 0.5$ with respect to the jet axis are considered associated with the jet. The impact parameter of a track is then defined as the distance of the track trajectory to the primary vertex. The point of closest approach to the jet axis is determined and from this point, taking the tangent to the track, the impact parameter is obtained as the distance between the primary vertex and the tangent, as illustrated in Figure 3.1.

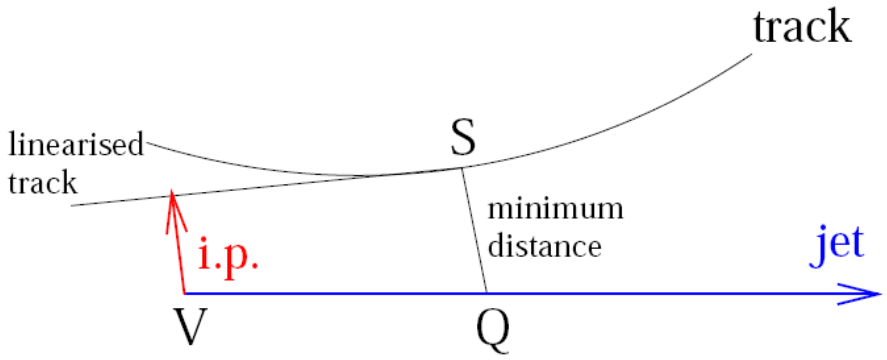

Figure 3.1: Sketch of the extrapolation of the impact parameter in 3D.

The impact parameter is positive if the angle between the impact parameter and the jet direction is smaller than 90°: for a B-hadron it is expected to be positive because the flight path of a B-hadron is in the same direction of the jet. Negative values could come from a bad jet or track reconstruction. Because of the impact parameter is an experimentally measured quantity, the ratio between the impact parameter and its experimental error, called *significance* of the impact parameter, $S = \frac{IP}{\sigma_{IP}}$ it is used in the b-tagging algorithms. In Figure 3.2 a sketch for an event with a jet is shown.

In CMS there are two algorithms based on the impact parameter significance of the tracks: the track counting (TC) and the jet probability (JP) algorithms:

- **Track counting based algorithm**: the discriminant for the track counting algorithm is the 3D impact parameter significance. The algorithm sorts the tracks according to descending 3D significance and it considers the significance of the n-th track. When choosing $n = 2$ the algorithm is called TCHE (track counting high efficiency), while when choosing $n = 3$ it is called TCHP (track counting high purity). Jets

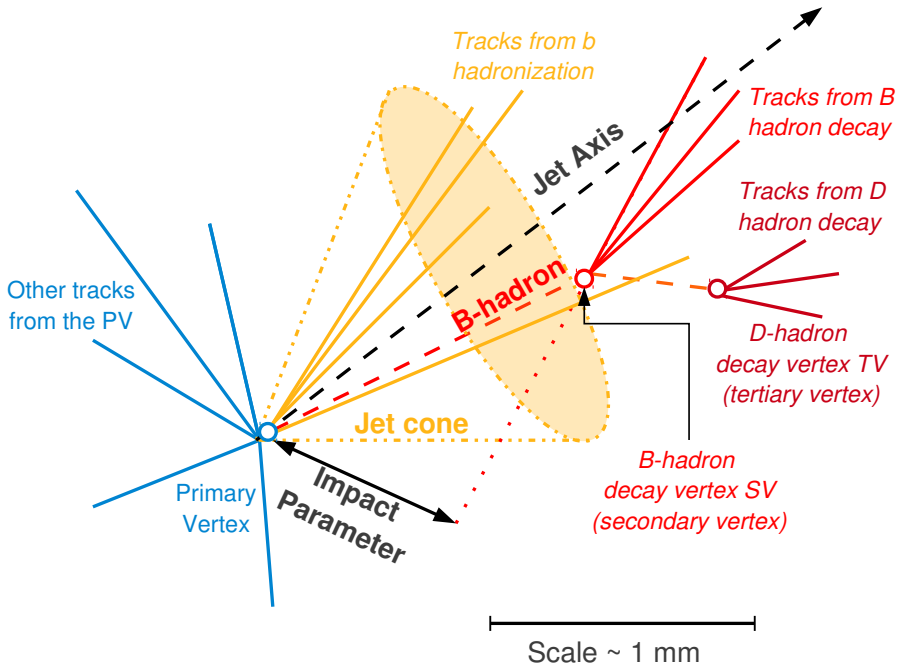

Figure 3.2: Sketch of an event with a jet

not containing good tracks are given a default value of the discriminant (-100). In Figures 3.3 and 3.4 few distributions of the two discriminants, divided by flavors, obtained using Monte Carlo QCD events and data recorded during the Spring 2010 are shown.

- **Jet probability algorithm**: it calculates for each jet the probability that the set of tracks associated to it have been originated from the primary vertex using the impact parameter significance of the tracks. The algorithm JP can be considered as an extension of the algorithm TC because it takes into account all the tracks associated to a jet and not only the first n tracks. The first step of this algorithm is to calculate the probability P_{tr} for each track with defined impact parameter significance S to originate from the primary vertex. This probability is essentially uniform for tracks coming from a primary vertex, while it is peaked at zero in presence of a secondary vertex. In order to increase the discrimination power tracks are divided into different *bins* and for each category (p_T, η, χ^2) the computation of the probability is performed. The second step is to combine the probabilities of the tracks associated to the jet. The jet probability for a given jet with

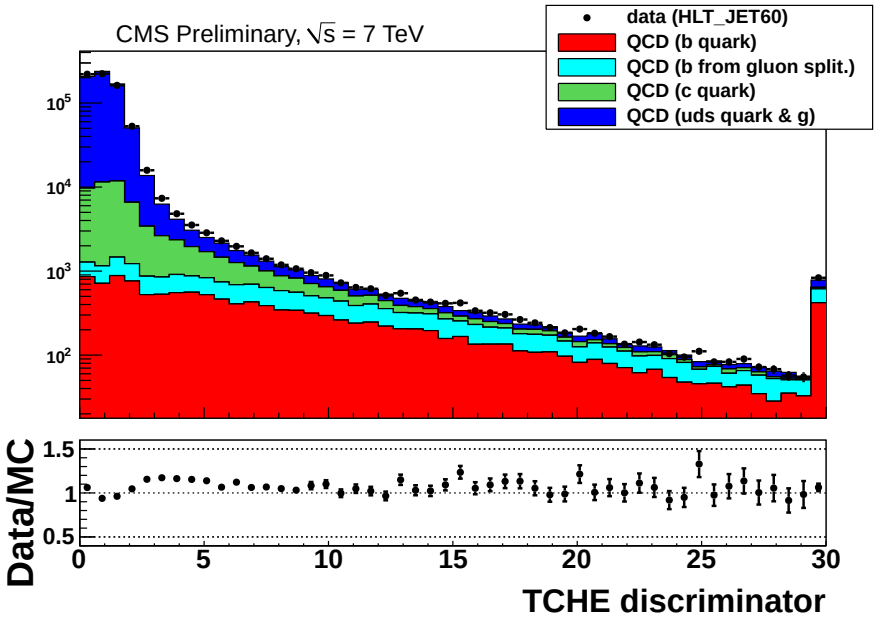

Figure 3.3: Shape of the discriminator TCHE divided for the different jet flavors.

N associated tracks is defined as the confidence level that any set of N tracks coming from primary vertex would give the observed track probability, namely:

$$P = \Pi \sum_{j=0}^{N-1} \frac{(-\ln \Pi)^j}{j!}$$

where:

$$\Pi = \prod_{i=1}^{N} P_{tr}$$

3.2.2 Secondary vertex based b-tagging algorithm

The idea of these algorithms is to use variables related to a reconstructed secondary vertex:

- **Simple secondary vertex b-tagging algorithm**: this is defined with the significance of the flight distance $\frac{D}{\sigma_D}$, where D is the distance

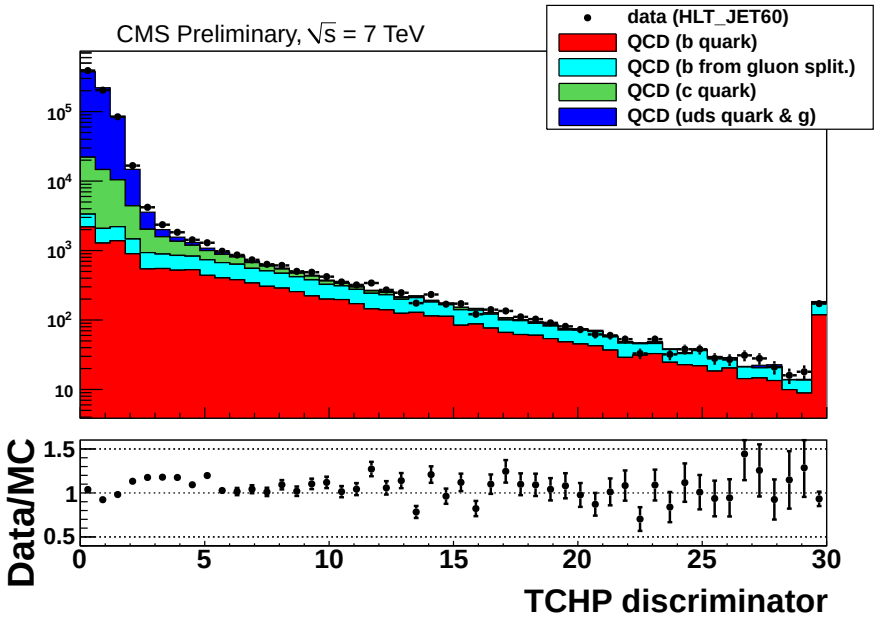

Figure 3.4: Shape of the discriminator TCHP divided for the different jet flavors.

between the reconstructed primary vertex and the secondary vertex calculated in the 3 dimensions. The discriminant is defined as $\log(1+\frac{D}{\sigma_D})$. The algorithm is called SSVHE (simple secondary vertex high efficiency) when choosing the second highest significance track, while it is called SSVHP (simple secondary vertex high purity) when choosing the third highest significance track. This is algorithm is robust against possible miscalibration of the detector because it requires at a reconstructed secondary vertex. In Figure 3.6 and 3.7 some distributions for the discriminants are shown.

- **Combined secondary vertex b-tagging algorithm**: although this algorithm makes explicitly use of secondary vertex related variables, the final discriminant is defined also when a secondary vertex is absent. The combination of these features allows the discriminant to provide a not-default value even when there is not a reconstructed secondary vertex. Variables are combined together using a multivariate analysis tool to perform the computation. In presence of a reconstructed secondary vertex the following variables are used:

  – three dimensional flight significance: the same variable used for the Simple Secondary vertex algorithms (SSV);

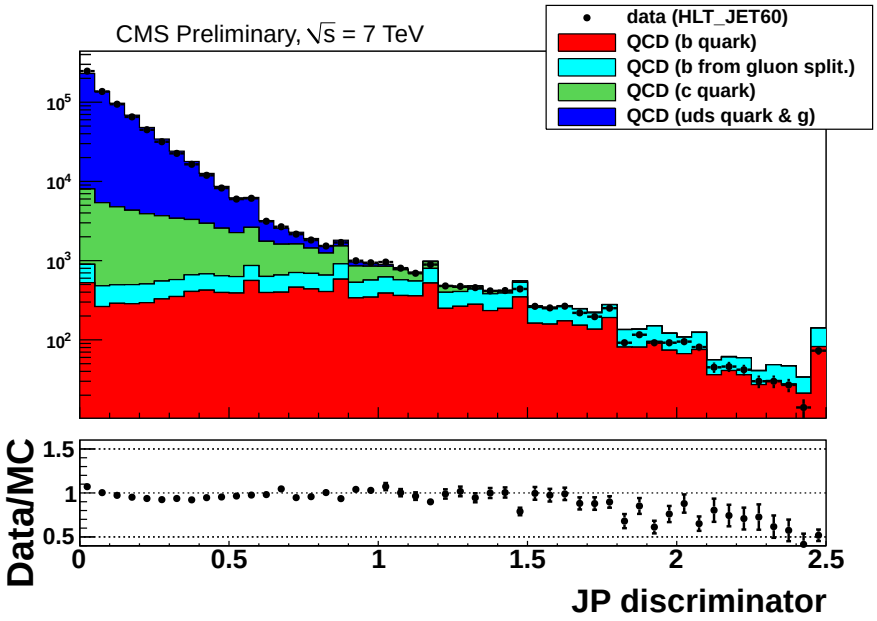

Figure 3.5: Shape of the discriminator JP divided for the different jet flavors.

- angle between the vertex flight direction and the jet axis: this angle is related to the energy carried by the bottom quark, in fact the most energetic the quark is, the most collinear is with the jet;
- vertex mass: invariant mass of the charged particles associated to the primary vertex;
- numbers of tracks associated with the vertex: vertex corresponding to a B-hadron decay has higher track multiplicity than that of a D-hadron decay;
- η_{rel} of all tracks from the vertex: for each track associated to the secondary vertex its pseudorapidty with respect to the jet axis is computed.

In absence of secondary vertex the following variables are used:

- track multiplicity: number of all selected tracks associated to a jet;
- three dimensional SIP: the first three tracks are evaluated using special probability density function.

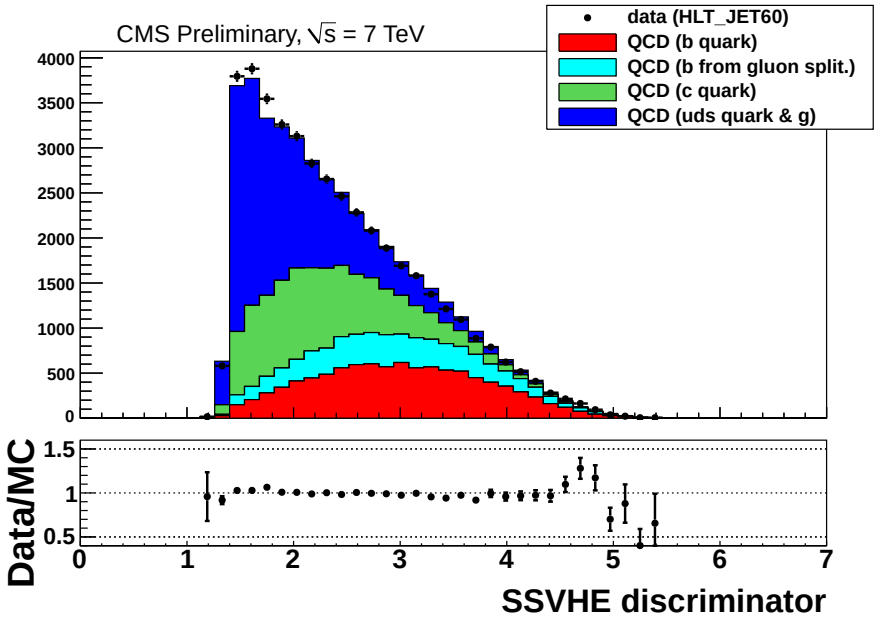

Figure 3.6: Shape of the discriminator SSVHE divided for the different jet flavors.

3.2.3 Soft lepton based b-tagging algorithm

This kind of algorithm exploits the properties of semileptonic decay of B and D-hadrons. The most common sources of objects faking electrons are charged hadrons with significant loss energy in the calorimeter, neutral pions and and photon conversions. About the muons, the main limitation in reconstructing these objects with low p_T is the contribution coming from mis-identification coming from decays of $\pi^{\pm}$ or $K^{\pm}$ in light jets with a muon inside the cone of the jet.
The algorithm makes principally use of four variables:

- p_{Trel}: the transverse momentum of the lepton with respect to to the jet axis;
- SIP: impact parameter of the lepton track from the distance of the closest approach of lepton track to the primary vertex;
- ΔR: the angular distance between the lepton and the axis of the jet;
- momentum ratio: ratio between the transverse momentum of reconstructed lepton track and the transverse energy of the jet.

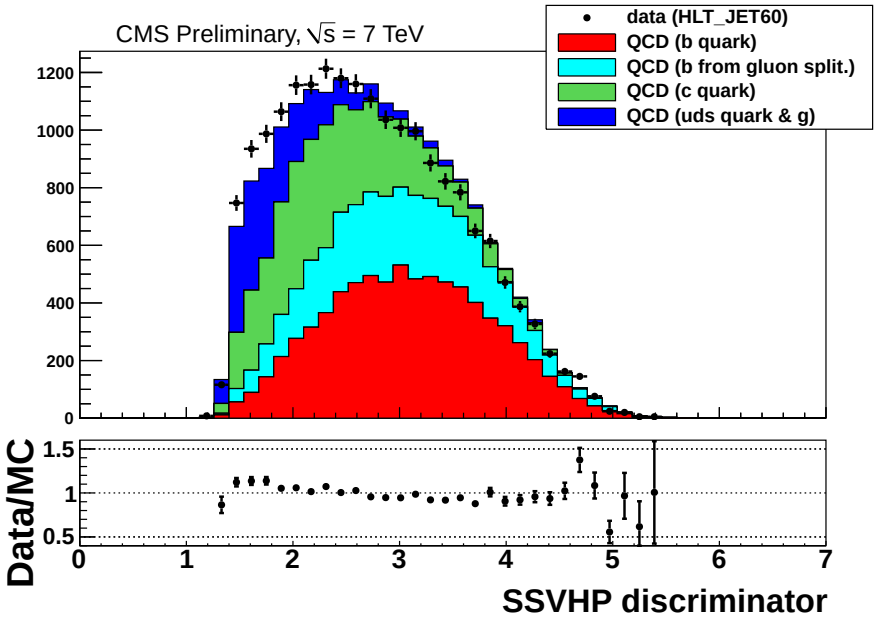

Figure 3.7: Shape of the discriminator SSVHP divided for the different jet flavors.

All these variables are combined into a neural network, which is trained on specific samples. The distributions for some variables are shown in Figure 3.8 and Figure 3.9.

In CMS all analyses are done using for each discriminant a fixed working point which assures a fixed efficiency of identifying (wrongly) a light jet as a b-jet (ϵ_l). The working points are defined as follows:

- tight: $\epsilon_l = 10^{-3}$;
- medium: $\epsilon_l = 10^{-2}$;
- loose: $\epsilon_l = 10^{-1}$.

These mistag rates depend on the p_T and η of the jets and therefore on the sample used. In CMS there is a POG[1] (BTV POG) which provides for these working points the correction factors between Data and Monte Carlo and the relative uncertainties. Performances of each discriminant are shown in distributions showing the ϵ_l as a function of the efficiency of tagging a b-jet (ϵ_b). The Figure 3.10 shows the performance of the different discriminant compared one to each other and it refers to a 3 millions events QCD Monte

[1]Physics object group

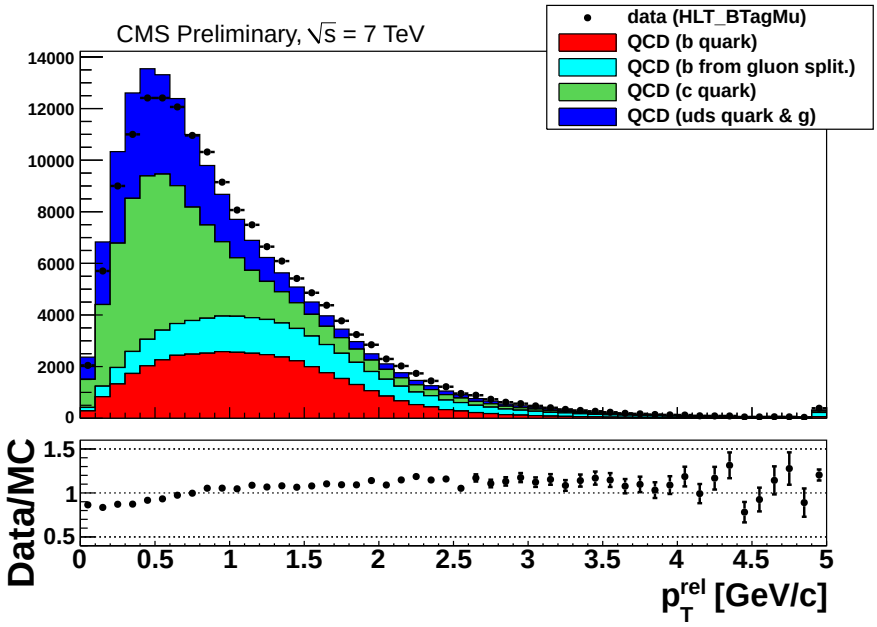

Figure 3.8: Transverse momentum of the muon with respect to the jet axis p_T^{rel}.

Carlo sample generated with a $\hat{p_T} > 80$ GeV/c, where $\hat{p_T}$ is the transverse momentum of the parton participating into a $2 \rightarrow 2$ hard scatter in the rest frame of the hard interaction.

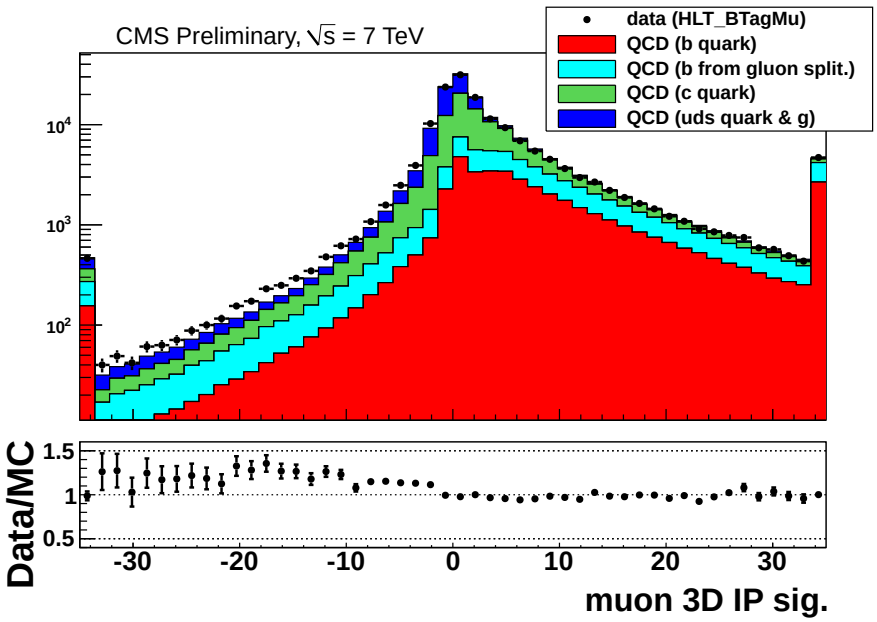

Figure 3.9: 3D impact parameter significance of muons in jets.

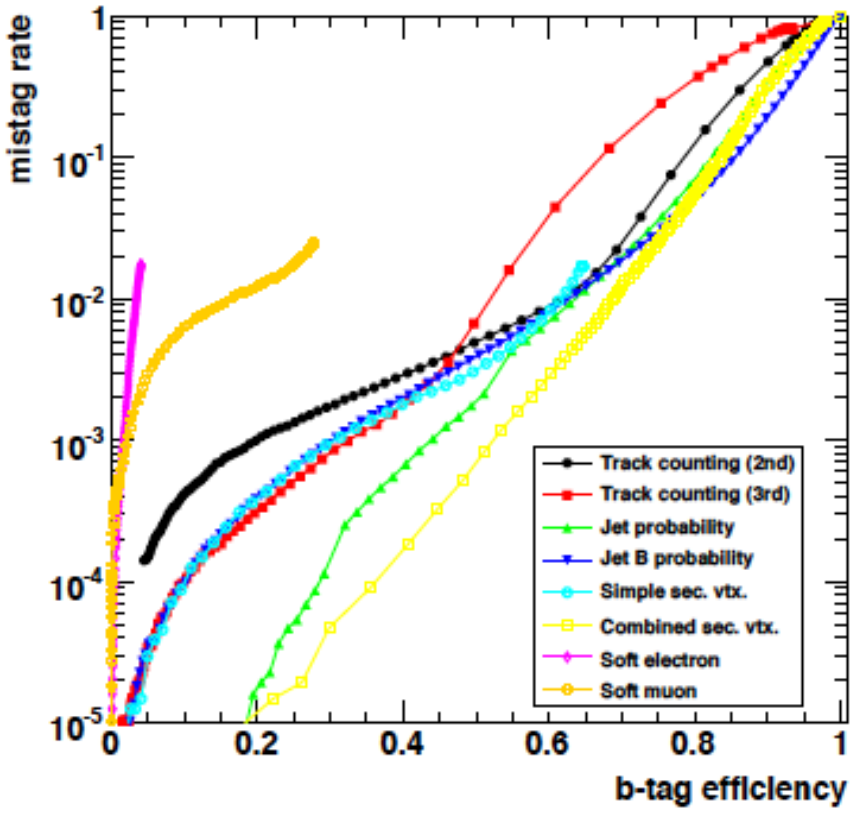

Figure 3.10: Performance of each discriminator in terms of b efficiency with.respect to mistag rate.

Chapter 4

Impact of the alignment of the tracker on the b-tagging performance

4.1 Track based alignment and misalignment scenarios

The goal of the track based alignment procedures is to determine the positions of the sensitive elements from a large sample of reconstructed charge particle trajectories (tracks). This can reduce the uncertainty on the position of the sensitive elements to the level of few microns. Track based alignment is based on the principle of minimization of track-to-hit residuals. Assuming some input geometry, for instance the design geometry or the one determined by the surveys available after the construction of the detector, each trajectory is built from charge deposits on individual detectors (that is called *hit*). Five track parameters describe the trajectory of a charged particle at the point closest to the nominal interaction point: the distance of the closest approach in the transverse plane d_{xy} and along the longitudinal direction d_z, the track azimuthal angle ϕ, the track polar angle θ and the ratio, signed curvature, between the charge and the transverse momentum $\frac{q}{p_T}$. A comparison between the measurement of the hits and the prediction from the track trajectory is then performed: the differences expressed in terms of local coordinates between the two quantities are called track-to-hit residuals. Ideally all measured points of the trajectory have normally distributed residuals, centered at zero and with the width corresponding to the nominal detector resolution. In reality the detector is misaligned: the measured points along a trajectory differ by how much the position of the module is shifted away from

its ideal position. The residual will be higher than in the ideal case. Track based alignment redifines the position of all the modules minimizing the sum of normalized residuals. The procedure is summarized in Figure 4.1.

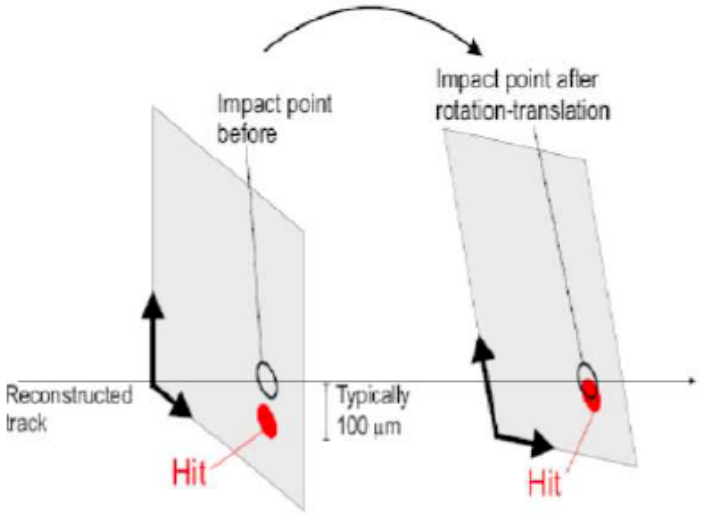

Figure 4.1: Track impact point and hit position on the module surface before and after alignment.

A set of corrections to the positions of the modules as determined by the alignment is usually defined as *geometry*. Even if the alignment performed before the first collisions already gave very good results, the position of the sensitive areas needs to be checked continuously during collisions because the geometry can potentially change and influence the reconstruction of tracks. To study the impact of possible remaining misalignment of the Tracker on track and vertex reconstruction in physics analysis, models of misalignment, called *scenarios*, were produced and made available in the standard CMS reconstruction software. A scenario allows to determine biases and deterioration of the resolution of the reconstructed physics observables. Different misalignment scenarios were produced to model the knowledge of the Tracker geometry at different stages of data taking. Instead of implementing the misalignment effects when generating simulated events, in the CMS offline it was decided to introduce the displacement of the detector modules directly at the reconstruction level. Thus the workflow is:

1. generate simulated events with the design geometry;

2. apply the shifts of the detector modules during the reconstruction of the events.

This approach is assumed to be equivalent to the situation found in the data where the true position of the real modules after an alignment remains unknown, but the shifts of the modules with respect to the geometry determined by the alignment are expected to be not too large because of the mechanical tolerances adopted during the construction. On the contrary, from the computational point of view, this approach is more convenient than generating simulated events with a deformed geometry and then reconstruct them with the design one. Finally one should mention that together with a set of corrections for the position, the alignment procedure provides also a set of Alignment Parameters Errors (APE).

4.2 Impact of misalignment on b-tagging

4.2.1 Validation on data

A set of different geometries were produced during 2010 and 2011 based on data, to best reproduce not only the knowledge of the real position of the modules, but also to introduce new degrees of freedom describing the deformation of the sensitive areas. The geometries available were:

- *plain*: a geometry where modules are considered flat and their position is set after the alignment procedure run on 2011 data;
- *KB (kinks and bows)*: with respect to the *plain*, in this geometry new degrees of freedom are added considering silicon modules bowed and not flat;
- *GR10v4*: this geometry was produced in 2010 using both cosmic rays and data tracks, but not considering kinks and bows;
- *ideal*: the design geometry, chosen as a cross check to evaluate the effect, expected to be large, of applying no alignment corrections at all.

The impact of different alignment conditions on the b-tagging performance needs to be estimated using the official tools provide by the BTV group because there is a standard selection based on specific requirements for the tracks used in the b-tagging and for the procedure how they get associated to a jet. The official tool to perform studies on the geometries based on data makes use of *standard* ntuples containing informations about the events

track p_T	> 1 GeV
tracker hits	≥ 8
pixel hits	≥ 2
normalized χ^2	< 5
2D impact parameter	< 0.2 cm
longitudinal IP	< 17 cm
ΔR(track,jet)	< 0.5
decay length	< 5 cm

Table 4.1: List of standard cuts applied for the production of the standard ntuples with the b-tagging validation tool.

selected according to the criteria summarized in Table 4.1. The BTV group provides also a set of ROOT scripts to produce plots running on the ntuples. These plots show distributions directly related to b-tagging discriminants (such as SV,IP,SIP), but also the quality of the reconstructed tracks (for example normalized χ^2) comparing two different geometries. The analysis described here was done using the dataset and the selections summarized in Table 4.2.

Data set	/BTag/Run2011A-PromptReco-v4/REC0
Numbers of events	590k events processed
Run range	165100-165500
Jet algorithm	ak5PF with jet energy correction ($E_T > 10$ GeV, $\|\eta\| < 2.4$)
Jet energy correction	L1Fast,L2+L3

Table 4.2: Summary table of the sample of data used for the analysis.

Starting from plots showing track properties, we notice that tracks refitted using the latest geometry *kinks and bows*, shown in Figure 4.2 with respect to the *GR10v4* geometry are better reconstructed because they have a small normalized χ^2. The *ideal* geometry was just chosen as a cross-check: in fact the differences with respect to the *GR10v4* geometry are very large, as shown in Figure 4.3. Figure 4.4 shows the 3 dimensional SIP considering all the tracks for the comparison between the *GR10v4* and the *ideal* geometry evidencing big discrepancies as expected. Figure 4.5 shows the 3 dimensional SIP considering only the second highest significance track, which enters into

the definition of the SSVHE discriminant: the difference is relevant only in the left tail where the latest geometry is better than *GR10v4*.

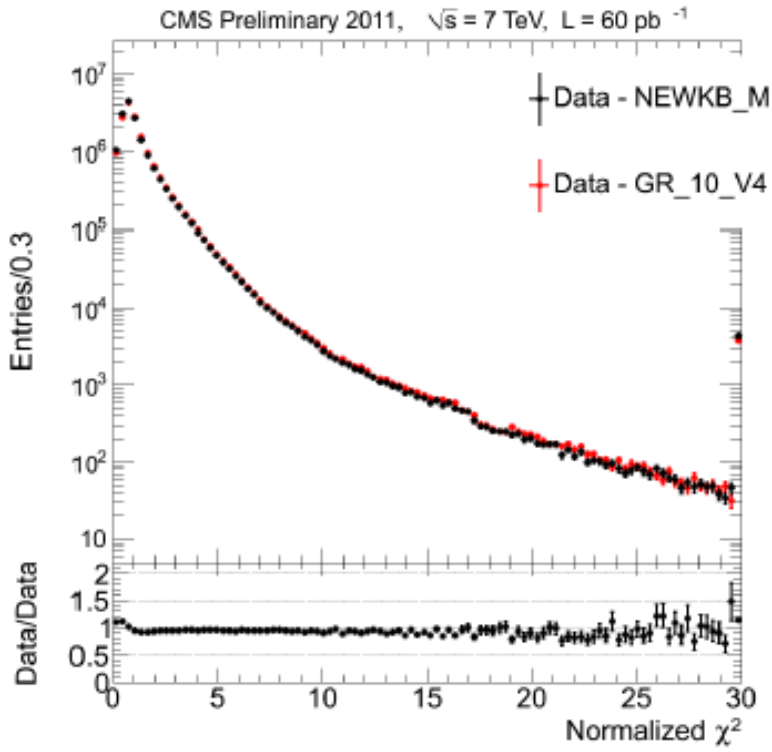

Figure 4.2: Normalized χ^2 of all tracks for scenarios kinks and bows (KB) and GR10v4. The bottom plot shows the ratio of the black curve with respect to the red curve.

The studies on these geometries confirm that the *Kinks and bows* geometry describes better the real detector than the older ones.

4.2.2 Studies on simulated events

Even if the alignment procedure achieved very good results already before the first collisions and it was continually monitored during all the data taking, nevertheless degradations of some variables, related to track reconstruction for example, were observed. This degradation can be interpreted as due to a change of the position of the modules during the data taking (for example movements along the z axis of the two halves composing the BPIX from 20 to

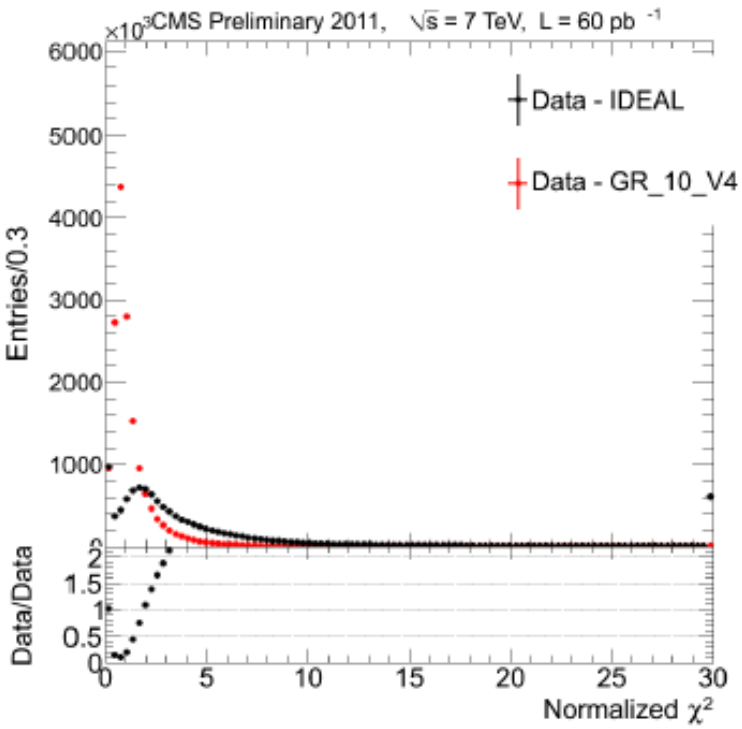

Figure 4.3: Normalized χ^2 of all tracks for scenarios GR10v4 and ideal. The bottom plot shows the ratio of the black curve with respect to the red curve.

100 μm). To study these changes of conditions one can model the movement of the modules in *misalignment scenarios* and study how sensitive variables change using simulated events. The BTV group provides a standard tool, based on DQM (data quality monitor) files, to perform studies on the change of the b-tagging performance introducing systematic misalignment scenarios and using simulated events. The performance of the b-tagging algorithms is evaluated studying the efficiency of tagging a light jet (ϵ_l), related to *mistag*, as a function of the efficiency of tagging a b jet (ϵ_b). The first step for this analysis was to test the robustness of b-tagging algorithms with respect to large deformations in the BPIX considering movements, for each half shell, from 200 μm to 4 mm along the z axis. We refer to them as *BPIX* Δz misalignment scenarios. A sketch illustrating these scenarios is shown in Figure 4.6.

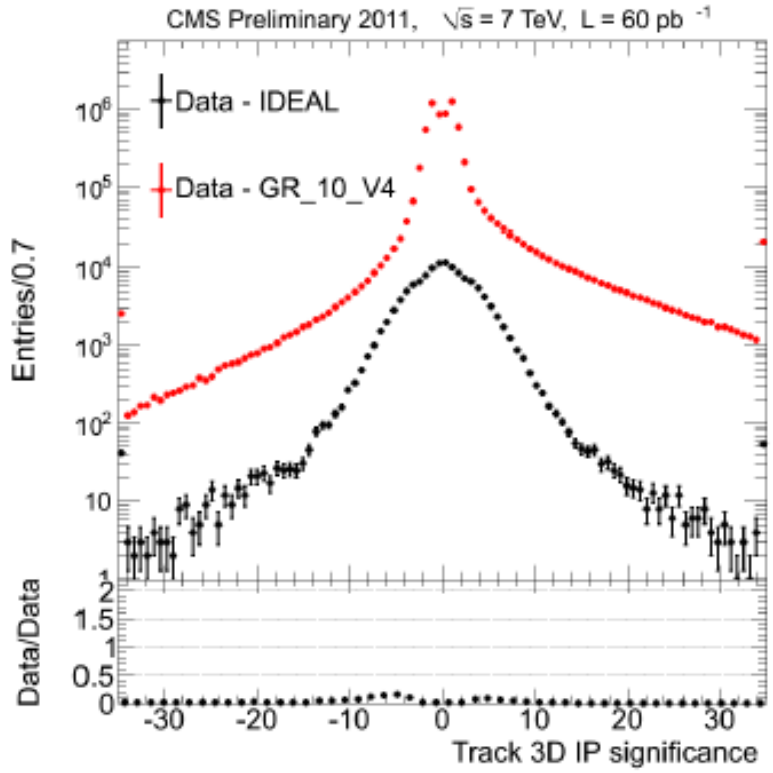

Figure 4.4: 3 dimensional SIP considering all the tracks for scenarios GR10v4 and Ideal. The bottom plot shows the ratio of the black curve with respect to the red curve.

In the following plots, obtained using 1 million $t\bar{t}$ events, the more the curve is shifting to the top left of the figure (large ϵ_l, small ϵ_b), the more the discriminant is loosing its performance. The results for the discriminants based on the impact parameter of the tracks are shown in Figure 4.7 for the TCHE and Figure 4.8 for the TCHP: there is a degradation of the performance of both TCHE and TCHP discriminants and for movements of some millimeters the shape of the discriminants breaks down.[1]

[1]These plots do not show any errors that are correlated among the different curves (as they are obtained from the same samples of events) and among different points of the same curve.

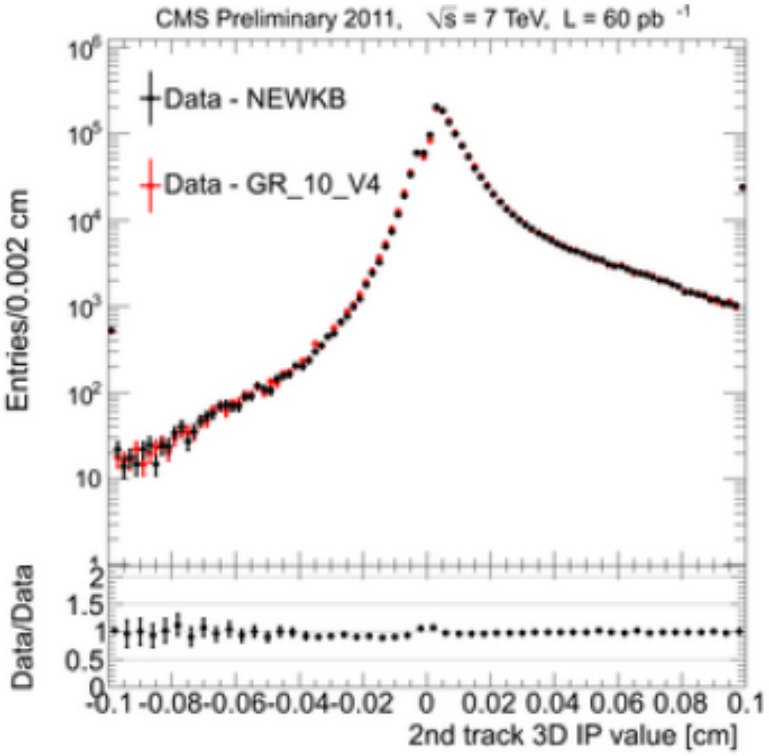

Figure 4.5: 3 dimensional SIP considering only the second highest significance track for scenarios kinks and bows and GR10v4. The bottom plot shows the ratio of the black curve with respect to the red curve.

The situation is very different when considering Figure 4.9 and Figure 4.10 which show the same plots but for the SSVHE and SSVHP discriminants. Althougt there is a global deterioration of performance, the discriminants are very robust even against misalignments of some millimeters, because they explicitly require a presence of a reconstructed secondary vertex.

It is difficult instead to draw any conclusion from Figure 4.11 and Figure 4.12 for JP and CSV discriminants: the former is calibrated on a particular set of conditions for reconstructing simulated events and so big differences are expected when these conditions change, the latter combines all the variables in a multivariate analysis and it is not straight-forward to assess how their combination is affected by a distortion of the geometry.

After these preliminary studies on how large misalignments could affect the discriminants the next step was to study more realistic scenarios. *BPIX* Δz scenarios were studied together with a new misalignment scenario called

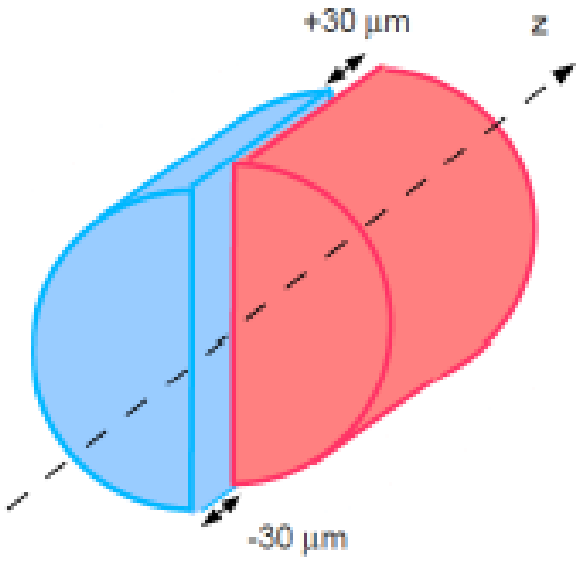

Figure 4.6: Example of a BPIX movement along z axis.

scissors like, which was created to explain a ϕ modulation observed in the analysis of the track residuals with respect to the beam spot. A sketch representing the *scissors like* scenario is shown in Figure 4.13.

The parameterization of the position in the transverse plane of the modules of the BPIX which corresponds to this scenario depends on the angle ψ between the z BPIX halves:

$$\begin{cases} x \to x' & = x + \Delta x(\varphi, \psi) \\ y \to y' & = y + \Delta y(\varphi, \psi) \end{cases}$$

$$R = \sqrt{x^2 + y^2}$$

$$\begin{cases} \Delta x & = R \cdot \sqrt{2(1 - \sin\varphi)} \cdot \sin\psi \\ \Delta y & = R \cdot \sqrt{2(1 - \sin\varphi)} \cdot (1 - \cos\psi) \end{cases}$$

Where $x, (x^{'}), y, (y^{'})$ are the original (modified) positions in the transverse plane of the active areas of the BPIX modules. For the study of this scenario

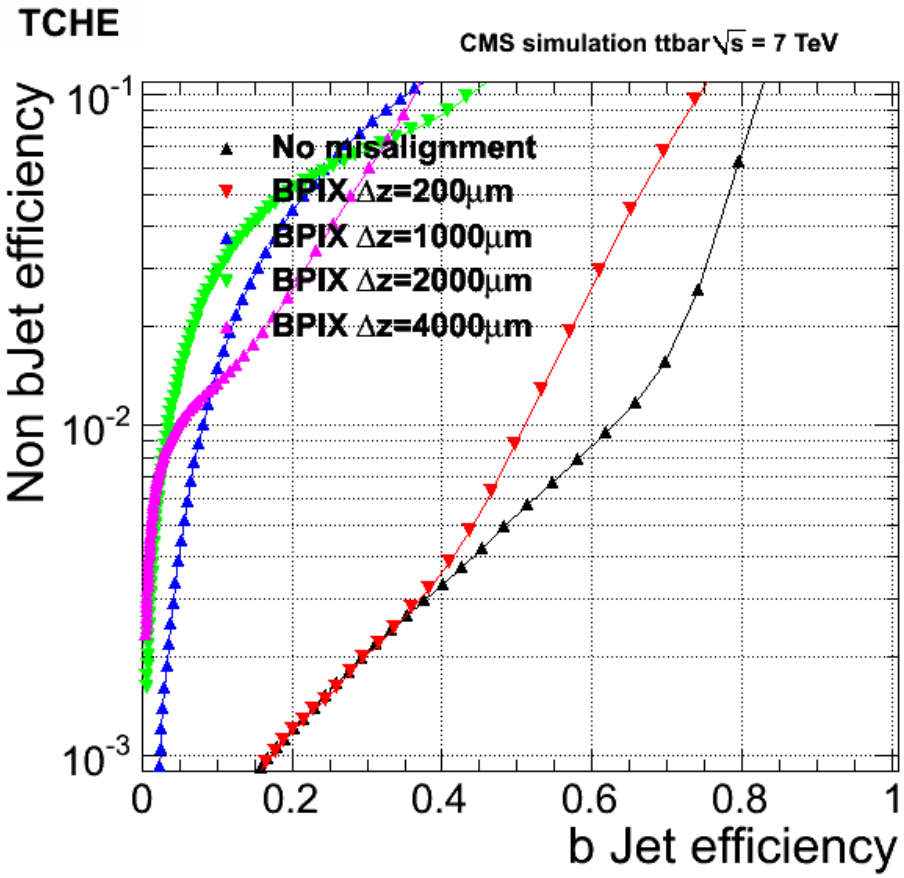

Figure 4.7: b jet efficiency with respect to light jet efficiency for $|\eta| < 2.4$ and $p_T > 30$ GeV for TCHE discriminant.

we considered ψ small, so we could use a Taylor expansion: $\sin\psi \simeq \psi$; $\cos\psi \simeq 1 - \psi^2/2$.
The results for the discriminants JP and TCHP are shown in Figures 4.14 and 4.15: we notice a larger degradation of performance for the 0.50 mrad than the 0.25 mrad.

Figures 4.16 shows the result for the SSVHE: in this case there is no evidence for a sensible degradation of performance.

For the studies on simulated events it was decided to present the results as the absolute difference in ${\epsilon_b}^2$ at fixed ϵ_l. The motivation is that the BTV group provides the full set of systematic uncertainty and DATA/Monte Carlo correction factors for values of the discriminants (working points) which guar-

[2] A b from a gluon splitting is considered in this category.

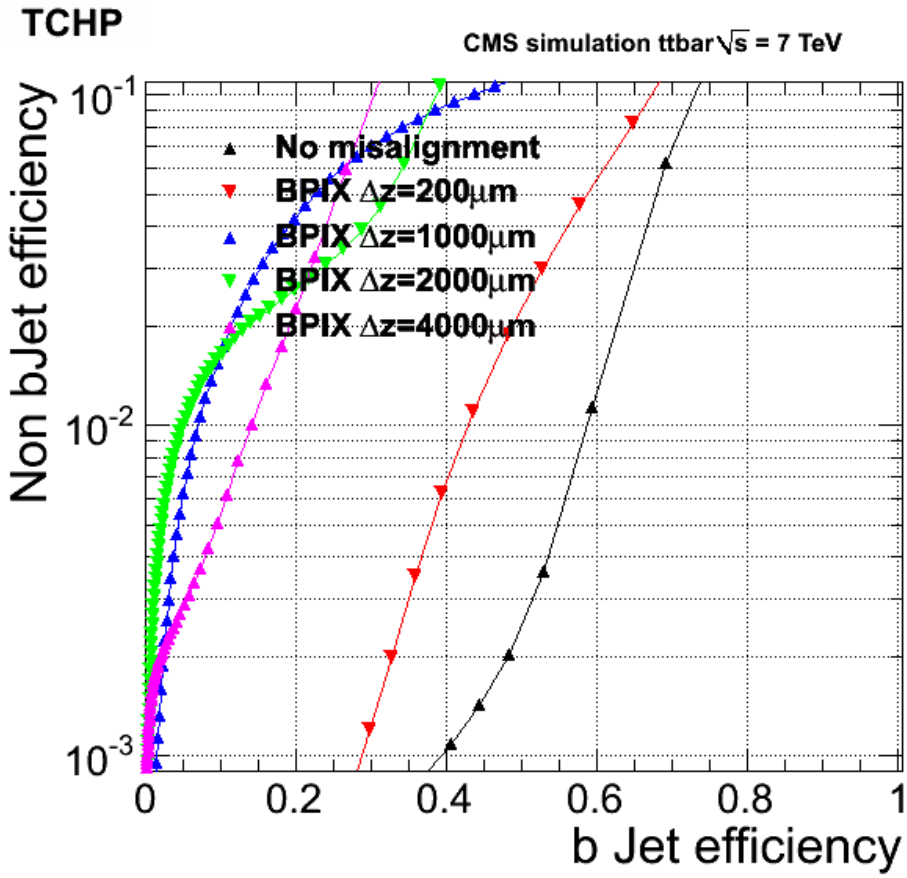

Figure 4.8: b jet efficiency with respect to light jet efficiency for $|\eta| < 2.4$ and $p_T > 30$ GeV for TCHP discriminant.

antee a fixed mistagging efficiency for light jets (ϵ_l). The results, expressed as the absolute difference in the efficiency of a tagging a b-jet as a function of the mistagging efficiency (ϵ_l) are shown from Figures 4.17 to 4.20 for Δz scenarios from 40 to 160 μm. For the TCHE and TCHP discriminants differences are very small up to $\Delta z = 120$ μm, while are about 10% for a misalignment of 160 μm. For the CSV differences are less than 10% for all the scenarios, while for the JP discriminant differences are larger, up to 20 %.

4.2.3 Validation of MC startup scenarios

During 2011 data taking a new scenario was studied, using one million $t\bar{t}$ events, to better reproduce in the simulation the evolved understanding

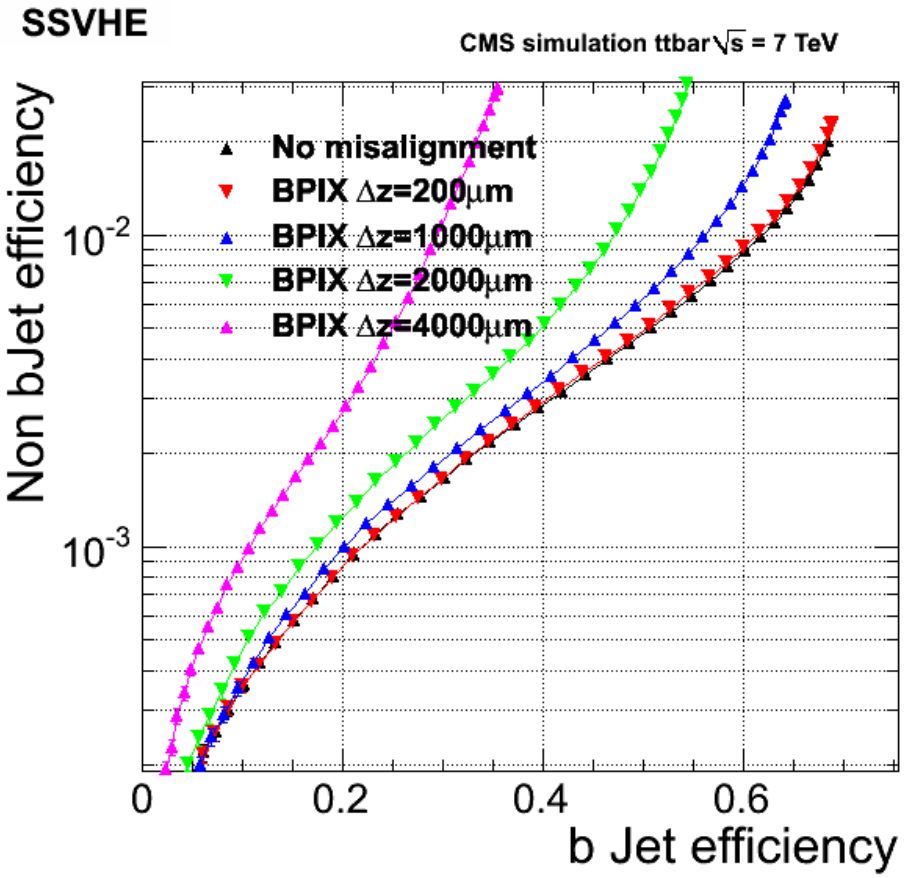

Figure 4.9: b jet efficiency with respect to light jet efficiency for $|\eta| < 2.4$ and $p_T > 30$ GeV for SSVHE discriminant.

of the real geometry of the tracker. This scenario was called Monte Carlo *2011 startup*.[3] We have decided to perform the validation of this scenario using the DQM tool. We have also decided to evaluate influence of APE on different scenarios, separating the 2011 Monte Carlo *startup* scenario with one considering APE different from zero and one considering APE set equal to zero. APE play a fundament role in b-tagging because the variable SIP is related to the uncertainty on the measurement of the impact parameter. The results are shown in Figure 4.23 for the TCHE discriminant, in Figure 4.21 for the JP discriminant and in Figure 4.22 for the CSV discriminant.

Differences are always less than 2% for all the discriminants, only for the

[3] In general a *startup* scenario is a scenario considering the uncertainties on the position of the modules after an alignment procedure.

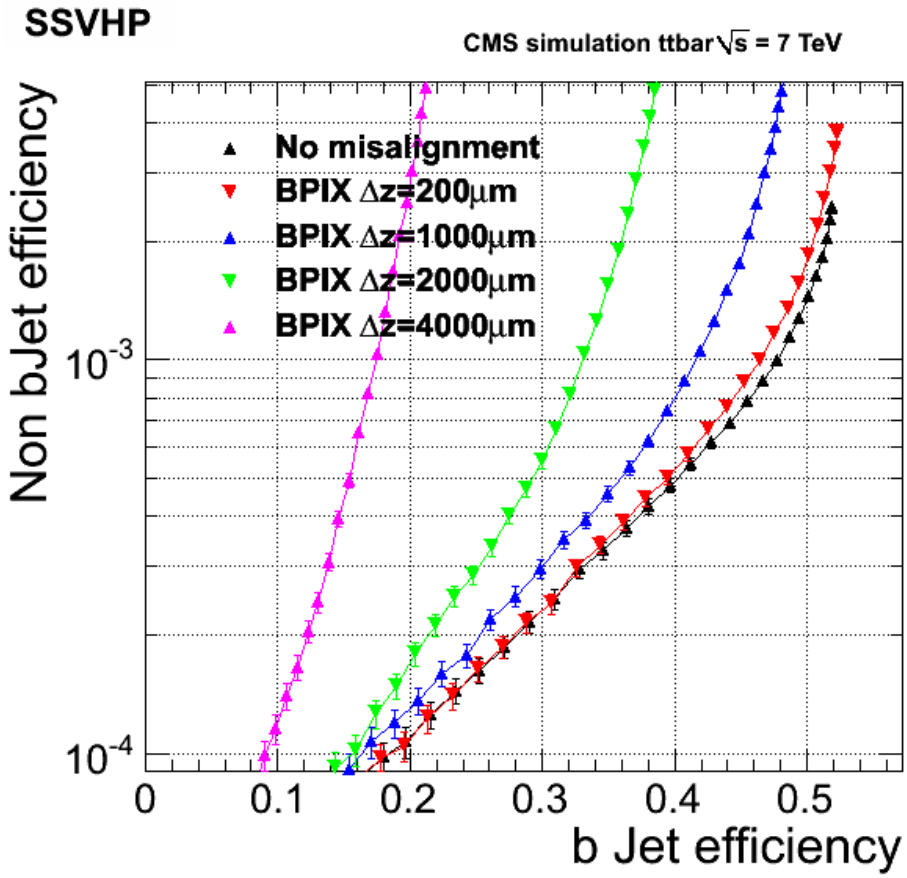

Figure 4.10: b jet efficiency with respect to light jet efficiency for $|\eta| < 2.4$ and $p_T > 30$ GeV for SSVHP discriminant.

TCHE discriminant the 2011 Monte Carlo *startup* with APE equal to zero looks better than the previous 2010 Monte Carlo *startup* especially for the workings point medium and loose.

4.2.4 Comparison of the b-tagging variables jet-by-jet

The results described in the previous sections showed that the DQM-based validation does not offer a sensible analysis tool to evaluate how a possible misalignment could affect the performance of b-tagging. Specifically it does not allow to compare how the variables, for example track or SIP, change with the different conditions used for the reconstruction event-by-event. The reason is that the standard CMS software allows the reconstruction of events with only a set of conditions, for example one geometry, at time. This means

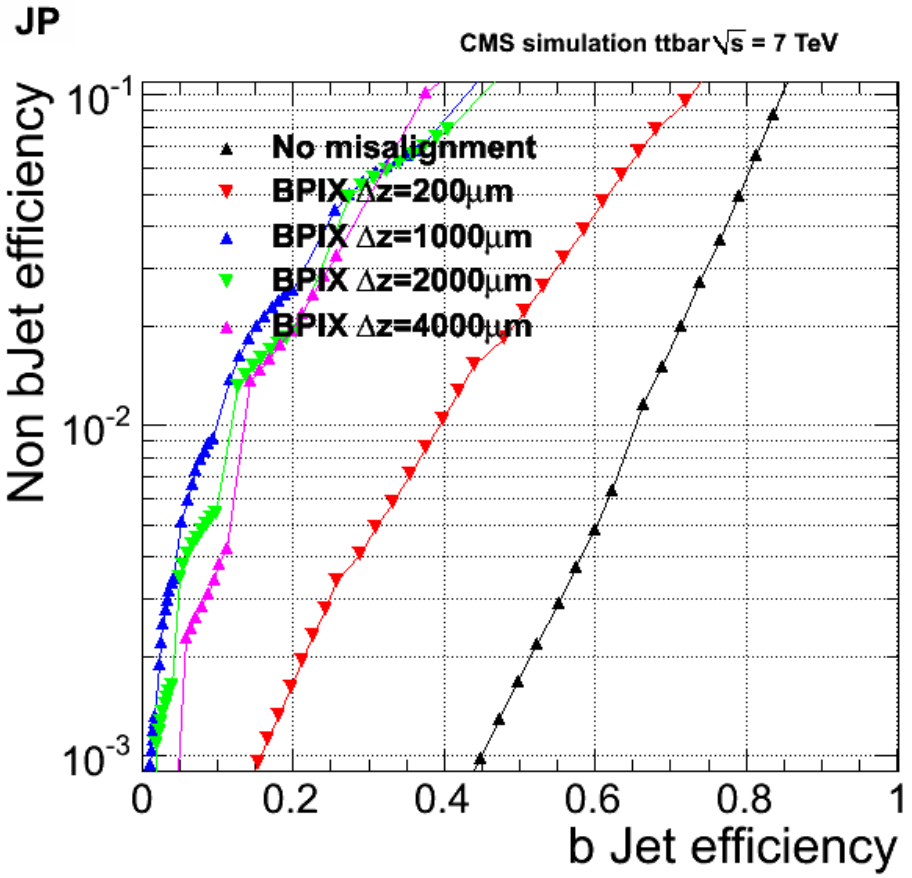

Figure 4.11: b jet efficiency with respect to light jet efficiency for $|\eta| < 2.4$ and $p_T > 30$ GeV for JP discriminant.

not to have the possibility of taking into account correlations. For this reason we have developed a jet-by-jet comparison tool. The workflow of the tool can be summarized as:

- produce standard b-tagging ntuples on different geometries, or in general two different conditions say *CondA* and *CondB*;
- match events between the two ntuples;
- match each jet of a given event in the two ntuples by using a cone matching algorithm, with a tight resolution in $\Delta R < 0.1$;
- produce a TTree containing the basic informations (η,ϕ,p_T) about the jets matched and storing them into a new class (called *Jet Info*). This

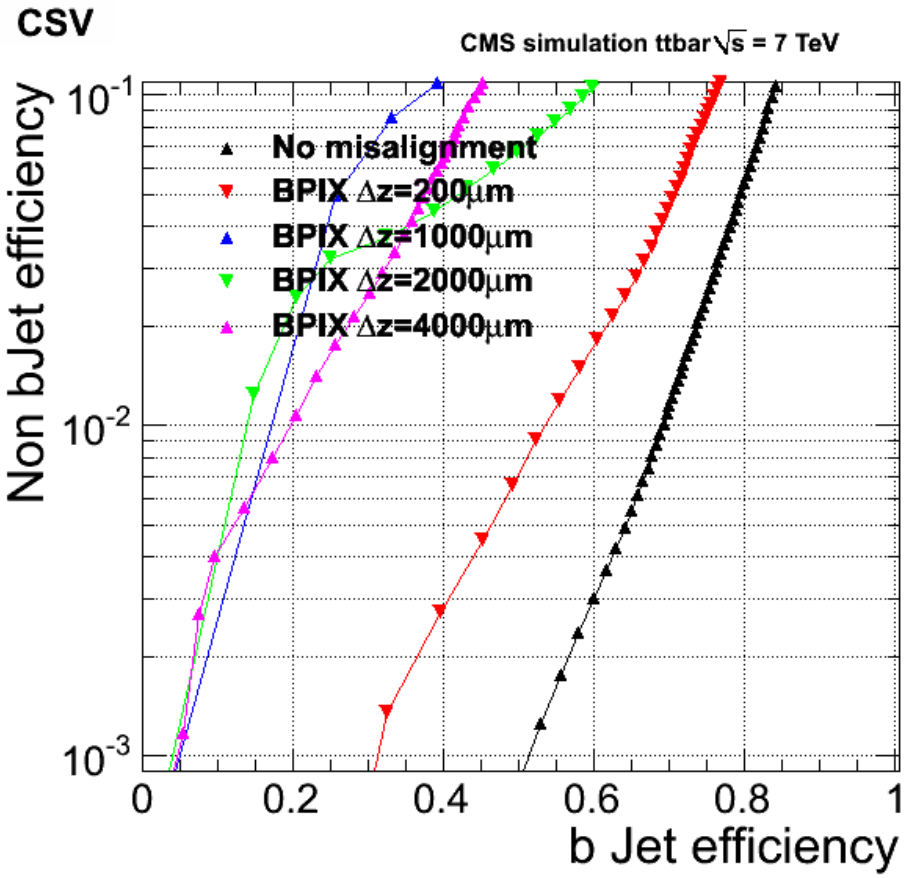

Figure 4.12: b jet efficiency with respect to light jet efficiency for $|\eta| < 2.4$ and $p_T > 30$ GeV for CSV discriminant.

class contains also informations about the discriminants and the most important variables used by the b-tagging algorithms;

- produce a set of reference plots with the difference for the most relevant per jet quantities as a function of some characteristic variable.

A possible concern for this approach was the time required to run it, as a typical test was performed on about 500k $t\bar{t}$ events. The results on the time performance of the tool were anyway conforting:

- submission of the ntuplizer jobs using GRID: about one day;
- scripts for matching events on a standard lxplus node: about 15 minutes;

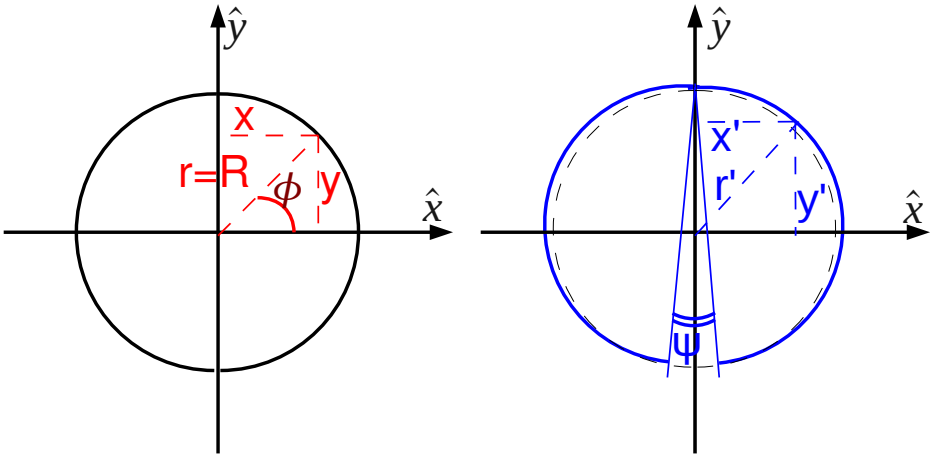

Figure 4.13: Sketch of the *scissor like* geometry: the parameter ψ can be changed to produce different misalignment conditions.

- script for jet matching and production of the TTree: one hour;
- final script to produce plots: 30 minutes.

A delicate point to consider in the analysis of the results is that, due to the change of reconstruction conditions, a discriminant may not provide a value associated to a jet. This requires to organize the results into different categories:

- not default-not default: it is the case when the algorithm successfully produces values for the discriminant different from the default for both the set of conditions;
- default-not default: the algorithm is providing a value for the discriminant only for one of the two set of conditions. For example this is the case for the SSV algorithms when, with the conditions applied, the algorithm fails to reconstruct a secondary vertex;
- default-default: both the algorithm fail to deliver a value for the discriminant in both the sets of conditions.

To take this fact into account, the tool allows a quantitative study on the numbers of jets moving from a default value of the corresponding discriminant to a not default value, producing migration matrices for each discriminant, as shown in Figure 4.24 for the comparison between the *Ideal* geometry and the *Ideal* geometry setting APE equal to zero. The fraction of migrating jets is about 1%.

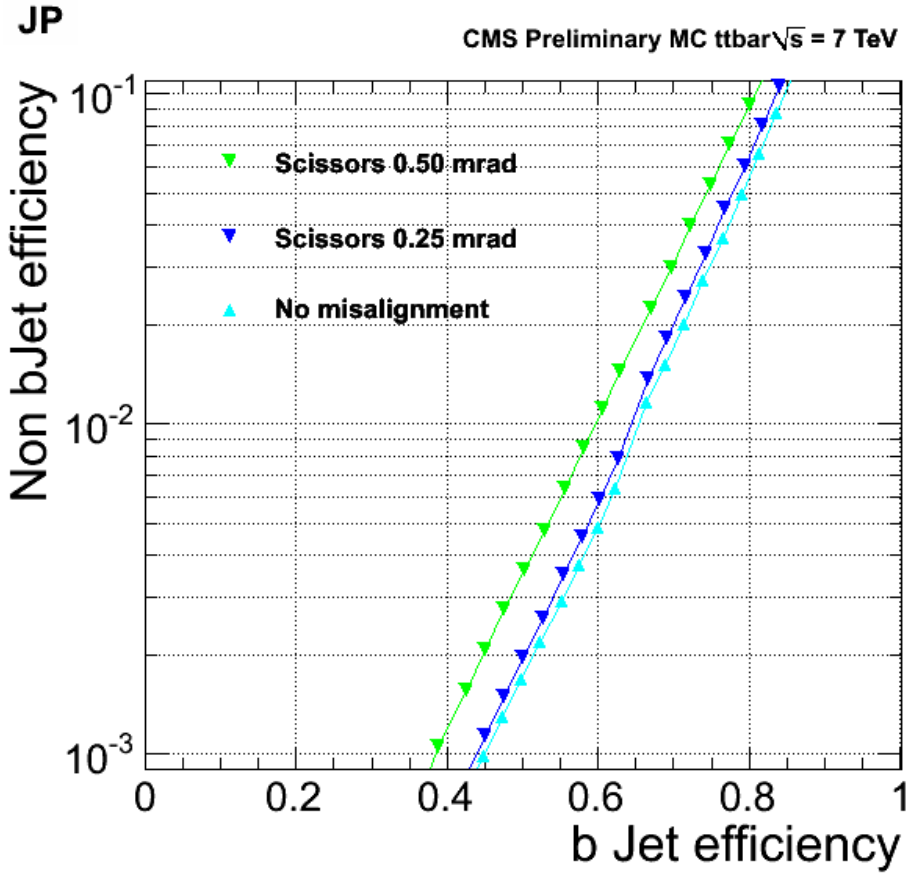

Figure 4.14: b jet efficiency with respect to light jet efficiency for $|\eta| < 2.4$ and $p_T > 30$ GeV for JP discriminant.

In the following we will always refer to the case not default-not default for all the distributions.
Distributions were also divided by flavors, one category for b-tagged jets (including gluon splitting) and one for light jets. In Figures 4.25 and 4.26 some distributions for the discriminant as a function of the shift of the position of the Primary Vertex (PV) along the z axis and in the transverse plane are shown.

Figures 4.28 and 4.27 show distributions for the discriminants TCHE and TCHP for $\Delta z = 40$ μm with respect to the ideal geometry.

The results above indicates that even if the tool can provide a very deep diagnostic study, correlation or evidences of big collective behaviors were not found, meaning that variables used for analyses in CMS are robust against

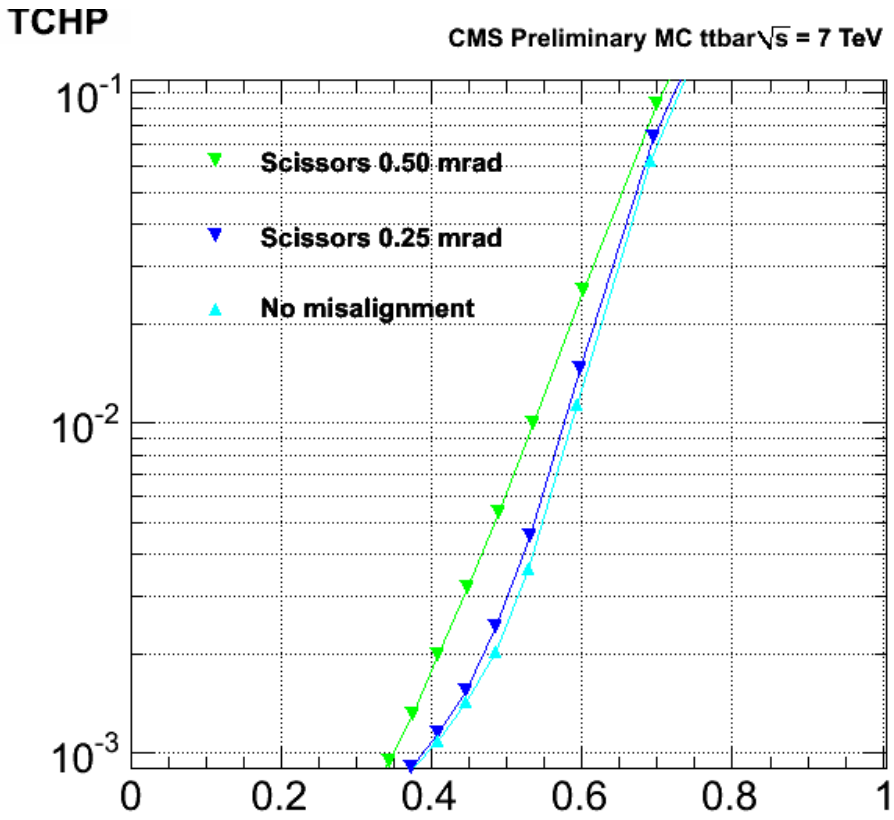

Figure 4.15: b jet efficiency with respect to light jet efficiency for $|\eta| < 2.4$ and $p_T > 30$ GeV for TCHP discriminant.

possible biases. Nevertheless this is the first tool in CMS which allows to compare two different set of conditions at the same time and it also very versatile. Altough conceived originally to test different alignment conditions, the tool can be used to compare any different reconstruction conditions and after its official presentation in the BTV group it has been used by other groups for comparing the performance of b-tagging algorithms in different pile-up scenarios and using fast and full simulation. Appendix A collects a set of plots for the different cases which were investigated.

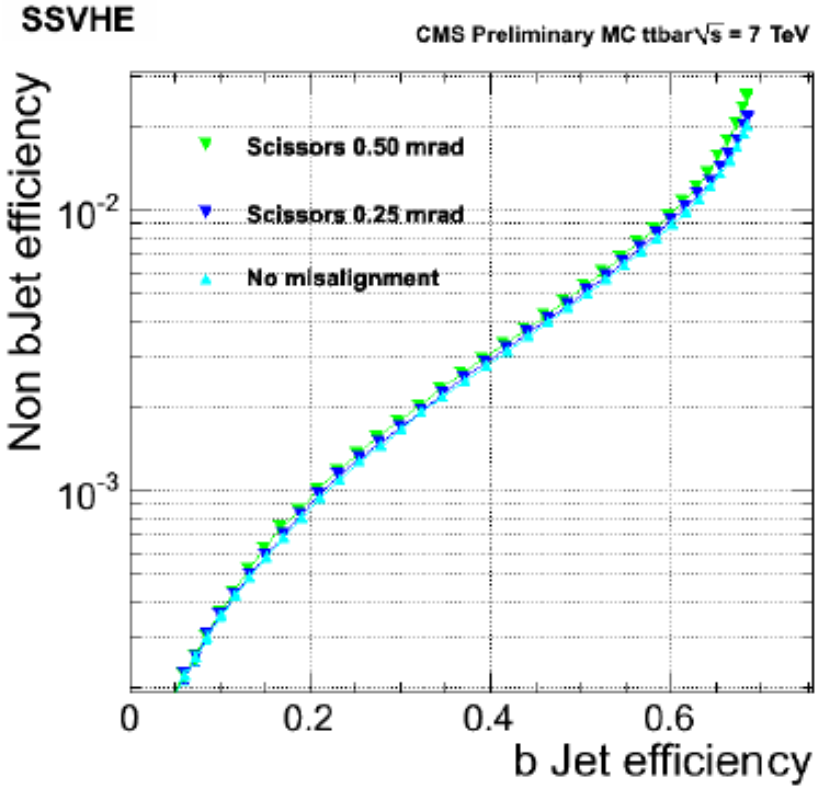

Figure 4.16: b jet efficiency with respect to light jet efficiency for $|\eta| < 2.4$ and $p_T > 30$ GeV for SSVHE discriminant.

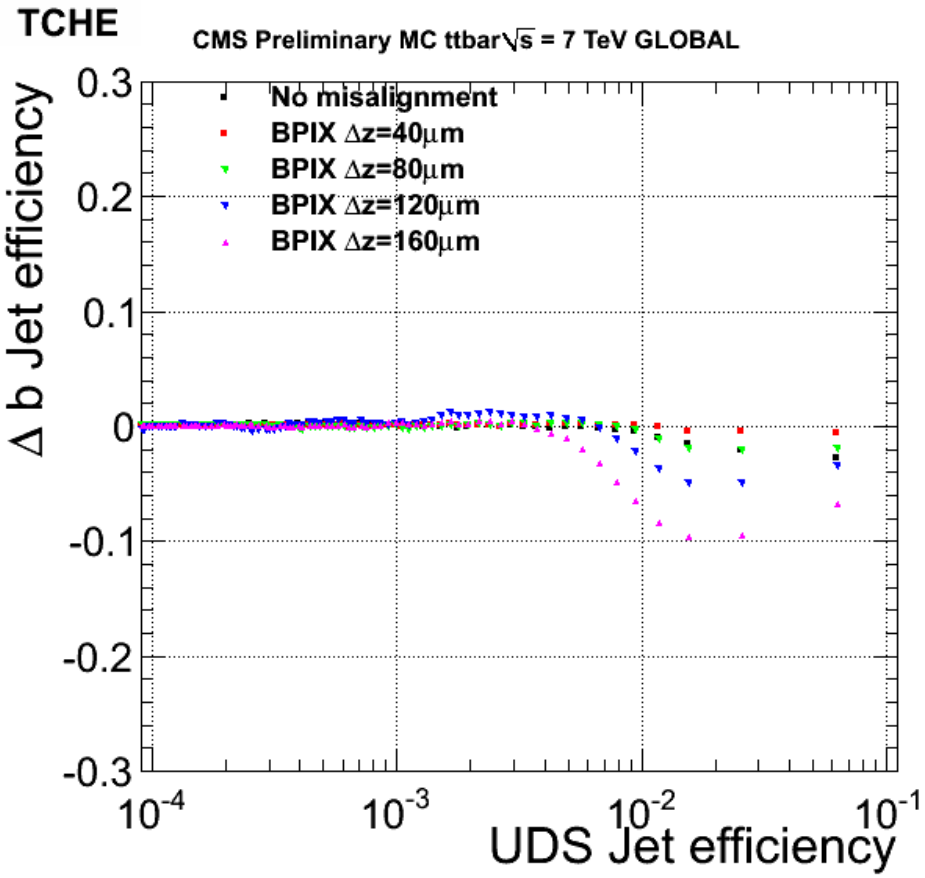

Figure 4.17: Absolute variation of b jet efficiency with respect to light jet efficiency for $|\eta| < 2.4$ and $p_T > 30$ GeV for TCHE discriminant.

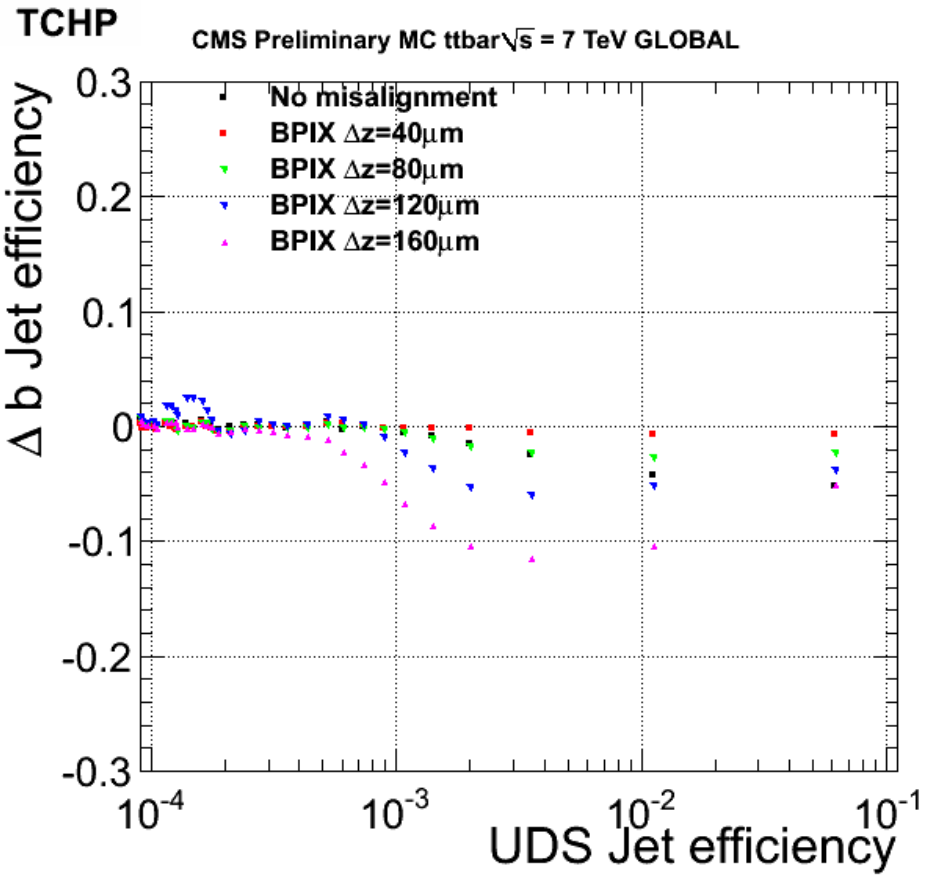

Figure 4.18: Absolute variation of b jet efficiency with respect to light jet efficiency for $|\eta| < 2.4$ and $p_T > 30$ GeV for TCHP discriminant.

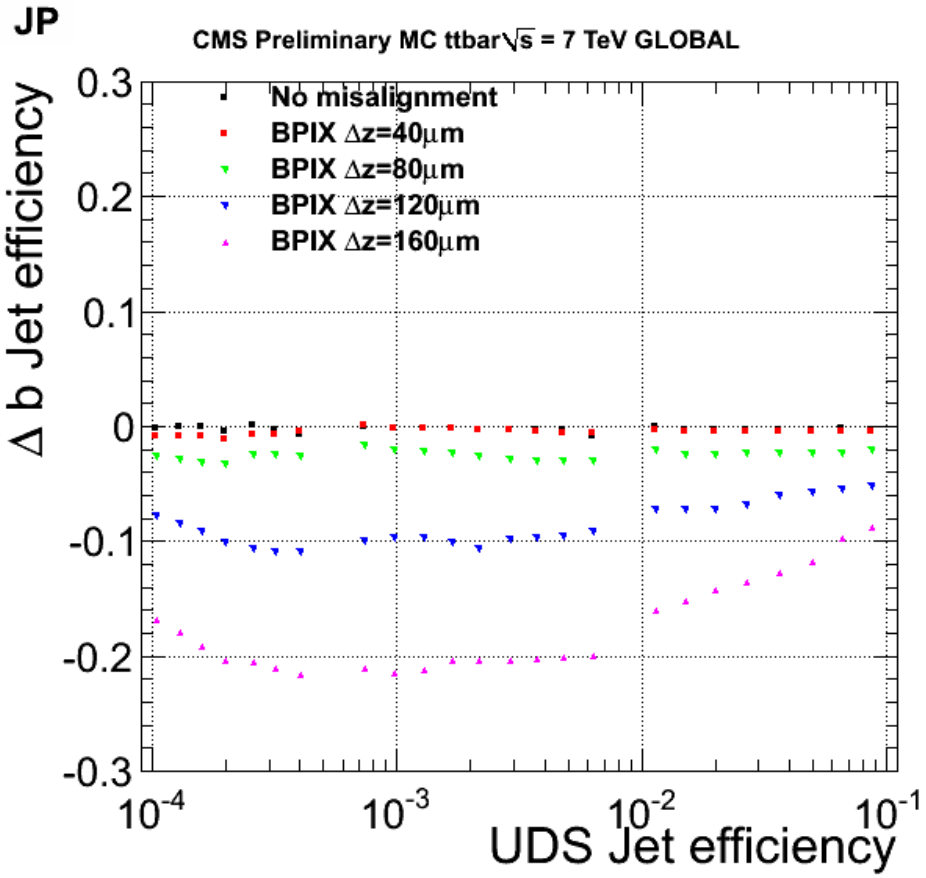

Figure 4.19: Absolute variation of b jet efficiency with respect to light jet efficiency for $|\eta| < 2.4$ and $p_T > 30$ GeV for JP discriminant.

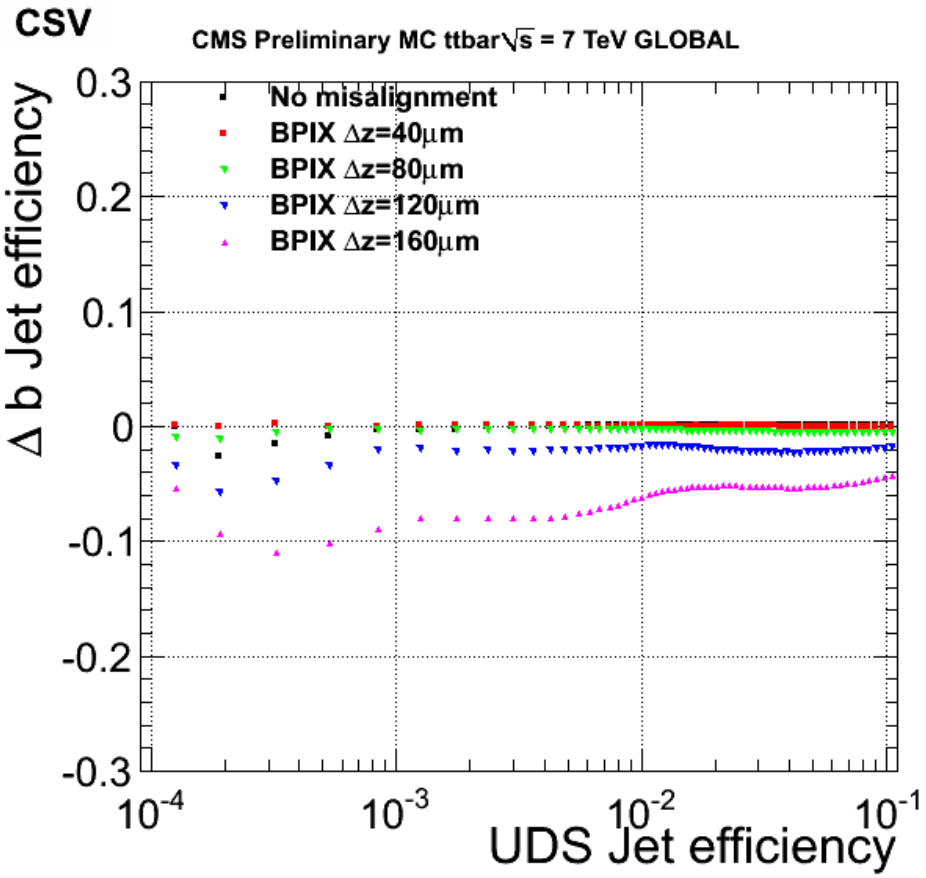

Figure 4.20: Absolute variation of b jet efficiency with respect to light jet efficiency for $|\eta| < 2.4$ and $p_T > 30$ GeV for CSV discriminant.

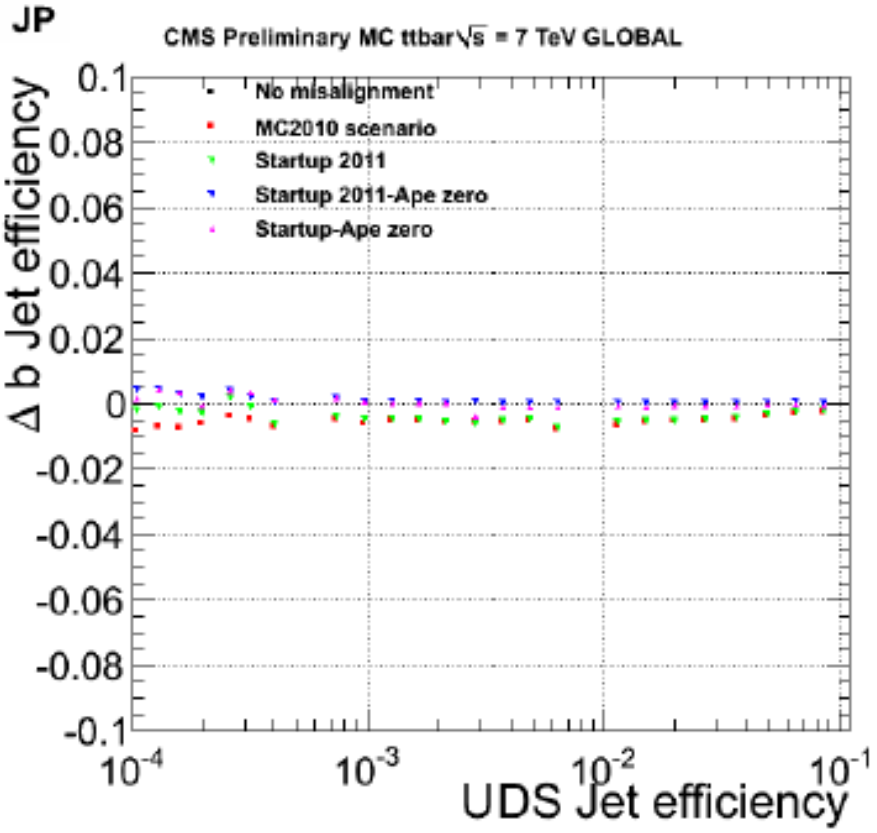

Figure 4.21: Absolute variation of the b jet efficiency with respect to light jet efficiency for $|\eta| < 2.4$ and $p_T > 30$ GeV for JP discriminant.

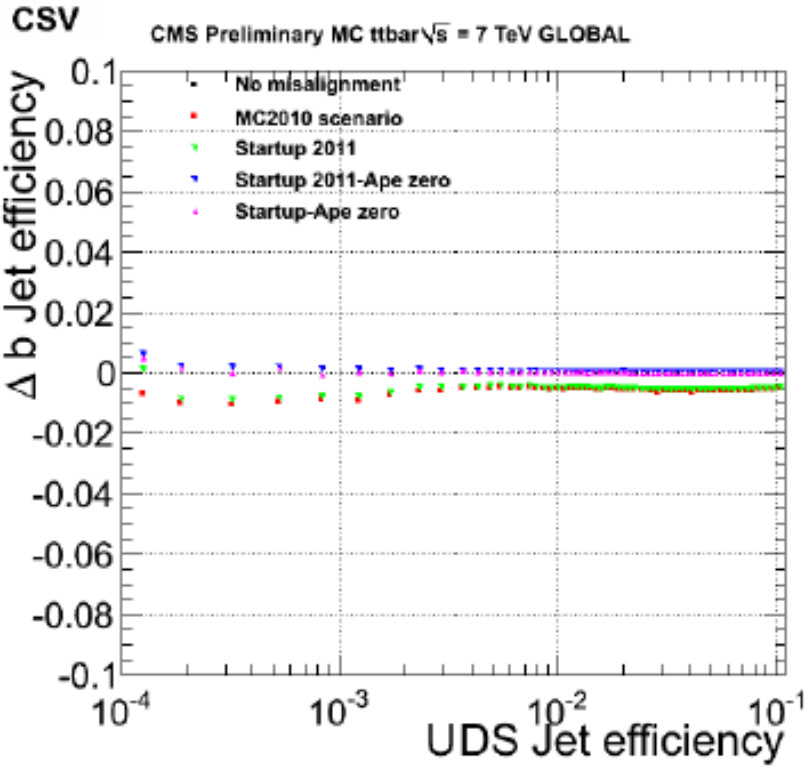

Figure 4.22: Absolute variation of the b jet efficiency with respect to light jet efficiency for $|\eta| < 2.4$ and $p_T > 30$ GeV for CSV discriminant.

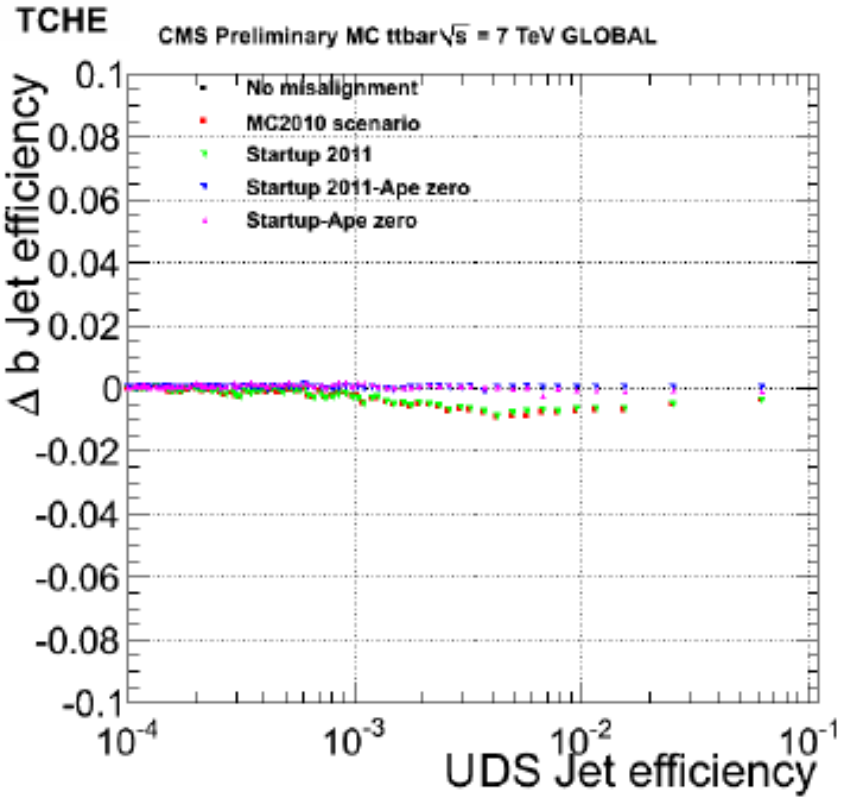

Figure 4.23: Absolute vacation of the b jet efficiency with respect to light jet efficiency for $|\eta| < 2.4$ and $p_T > 30$ GeV for TCHE discriminant.

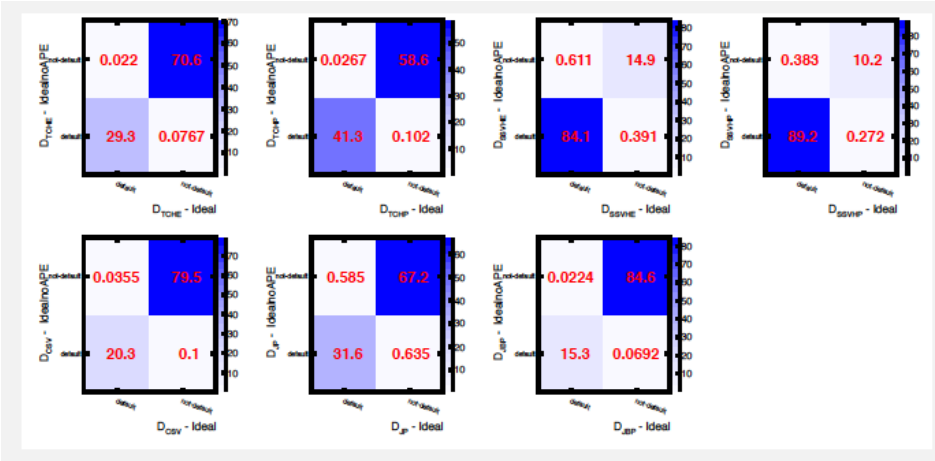

Figure 4.24: Table of migration matrices from default to not-default values of the discriminants, for Ideal geometry versus Ideal geometry with APE = 0.

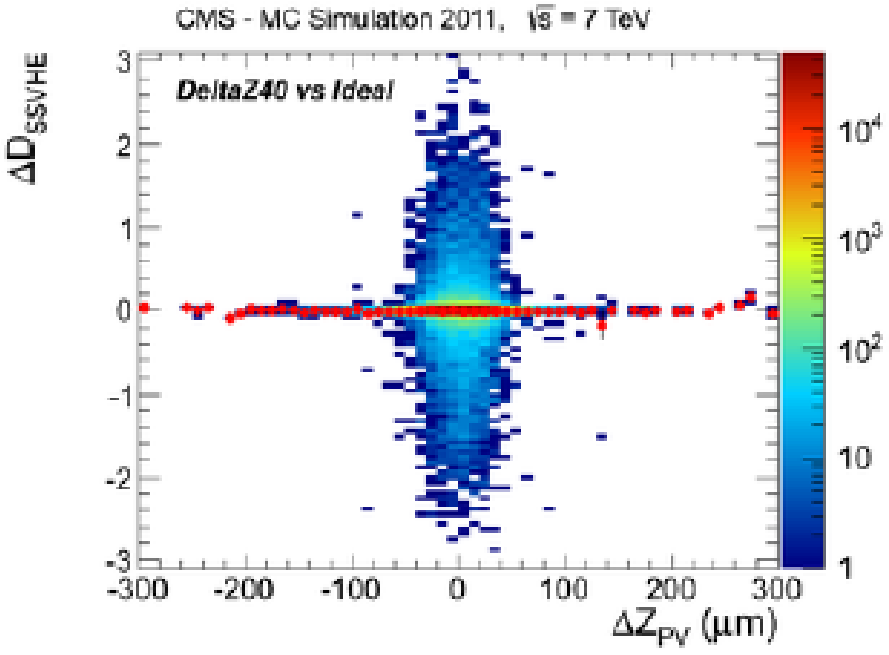

Figure 4.25: Jet-by-jet difference of SSVHE discriminant as a function of shift on the position of the PV longitudinal projection.

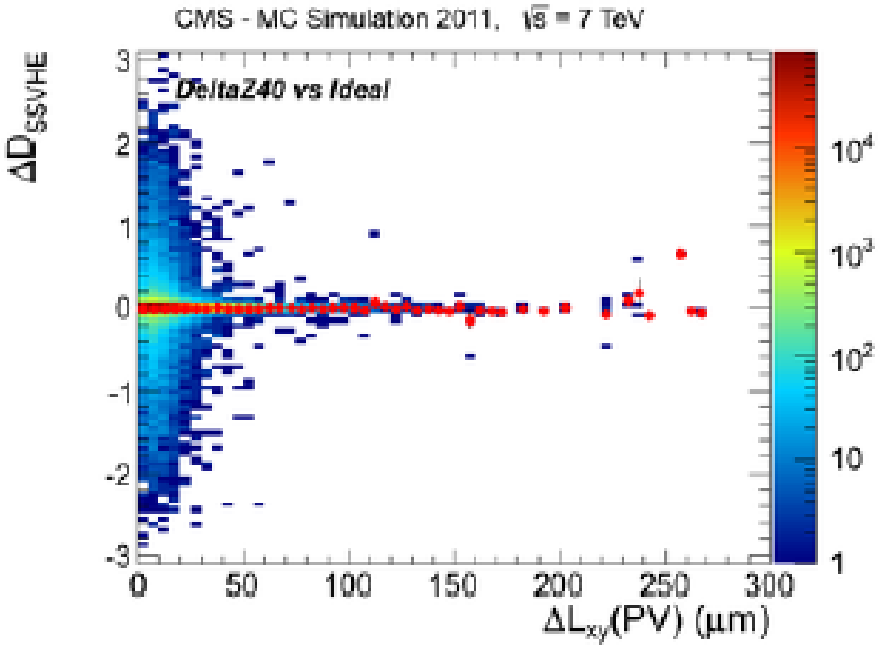

Figure 4.26: Jet-by-jet difference of SSVHE discriminant as a function of shift on the position of the PV projection in transverse plane.

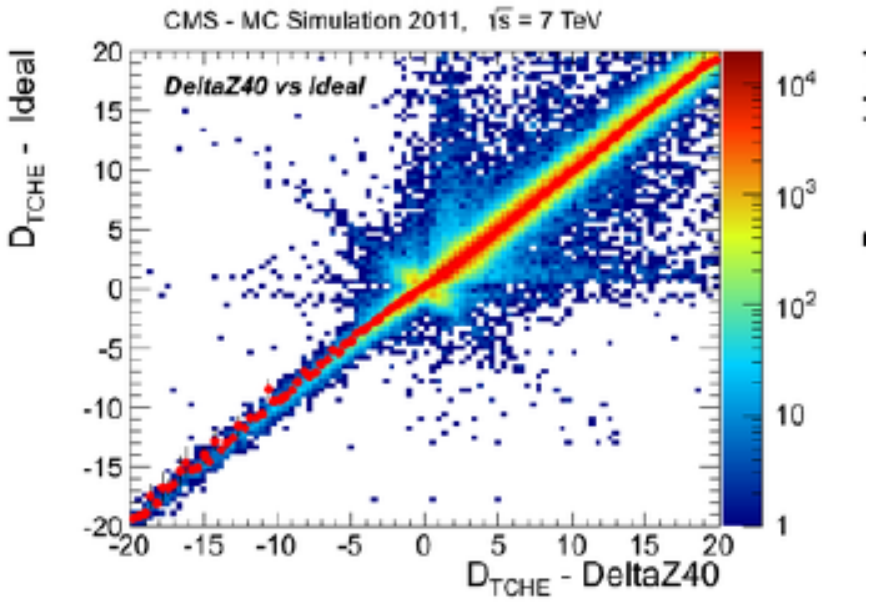

Figure 4.27: Correlation of the TCHE discriminant in case not default-not default for the scenarios Ideal (y axis) and Δz 40 μm (x axis).

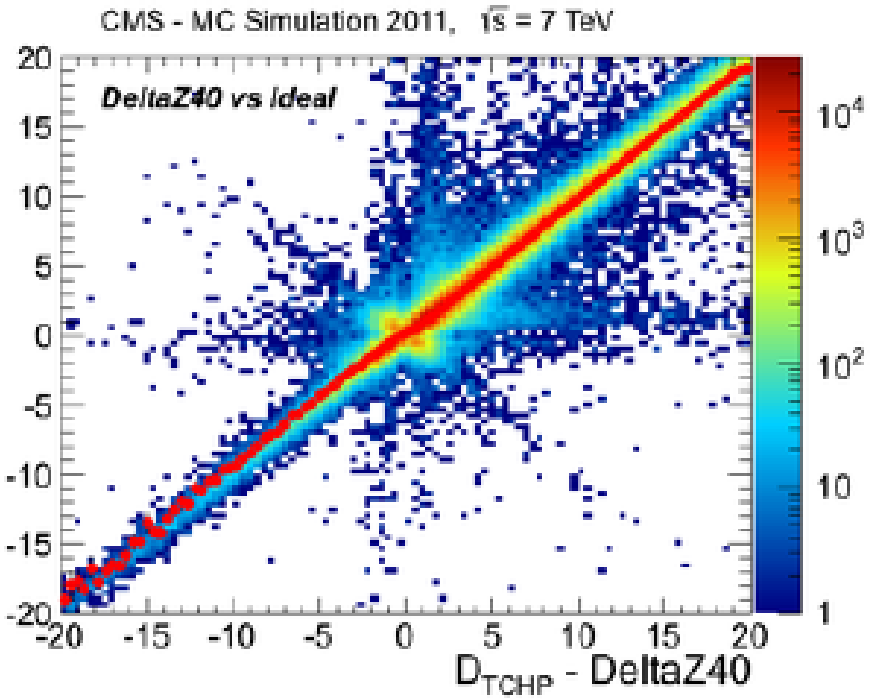

Figure 4.28: Correlation of the TCHP discriminant in case not default-not default for the scenarios Ideal (y axis) and Δz 40 μm (x axis).

Chapter 5

Production of a Z boson in association with heavy flavors

In the previous chapter a detailed study about the b-tagging algorithms has been presented. It was then decided to apply one of this b-tagging algorithm, the SSVHP, to measure a process of interest at LHC: the production of a Z boson in association with bottom quarks, usually called $Z+b$. The first part of this chapter presents a theoretical introduction to this process, while the second part is devoted to the description of the measurement of $Z+b$.

5.1 Standard model Higgs search

The measurement of the Z+b production is important, both as a benchmark channel when studying the production of a light SUSY Higgs boson in association with b quarks, and as an electroweak background in searches for the standard model Higgs boson. The Z+b process is a relevant background for the SM Higgs boson searches especially in two channels:

- $H \to ZZ^* \to 4l$, the *golden channel*, for which it represents a reducible background if besides the two leptons coming from the Z boson the b quarks decay semileptonically, being the total branching ratio for the semileptonic decay of about 20%;

- the production of a low mass Higgs in association with a Z boson, *Higgstrahlung*, in which the Higgs boson decays into a $b\overline{b}$ pair (the most probable decay mode in the low mass range). The signature for this process is represented by two leptons coming from the Z decay and two b-tagged jets: this is the same signature of a $Z+b$ event with two b-tagged jets.

5.2 Minimal supersymmetric extension of the Standard Model (MSSM)

In constructing the MSSM, both Y= -1 and Y= +1 Higgs complex doublets need to be considered to obtain an anomaly-free theory. Two types of models can be built corresponding to different scalar-fermion couplings:

1. Type I model: all fermions couple to the same doublet and the Yukawa couplings are rescaled to $\tan\beta = \frac{v_2}{v_1}$, where v_1 and v_2 are the vacuum expectation values of the two doublets;

2. Type II model: the up fermions couple to one doublet with an intensity proportional to $\cot\beta$, while the down fermions couple to the other doublet with intensity proportional to $\tan\beta$.

In the Type II model the doublets are written as $\Phi_d = (\Phi_d^0, \Phi_d^-)$ coupled to the down-type fermions and $\Phi_u = (\Phi_u^+, \Phi_u^0)$ coupled to the up-type ones. Once the Higgs potential is minimized, their respective vacuum expectation values are:

$$< \Phi_d >= \begin{pmatrix} v_d \\ 0 \end{pmatrix}$$
$$< \Phi_u >= \begin{pmatrix} 0 \\ v_u \end{pmatrix}$$

where the vacuum expectation values satisfy: $v^2 = v_d^2 + v_u^2 = (246\ \text{GeV})^2$ and $\tan\beta = \frac{v_d}{v_u}$. The spontaneous symmetry breaking mechanism results in three Goldstone bosons and five physical Higgs particles: a charged pair $H^\pm$, one CP-odd neutral scalar A and two CP-even neutral scalars h and H. Without any claim of completeness, we only underline that quantities in the MSSM Higgs sector depend on two parameters which are $\tan\beta$ and one mixing angle α.

$\Phi b\bar{b}$ production in the MSSM

For large values of $\tan\beta$, the production of a neutral Higgs boson in MSSM ($\Phi = h/H/A$) at the LHC is dominated by the Higgs-radiation off bottom quarks, leading to the associated production $\Phi b\bar{b}$ shown in Figure 5.1.

The Z+b measurement is important as a background for the decay exclusive channel $pp \to \Phi b\bar{b} \to (\mu^+\mu^-)b\bar{b}$.

5.3 Z+b production al LHC

Inclusive events with a Z boson and one b jet identified in the acceptance of the detector are expected to give the main contribution to the cross section

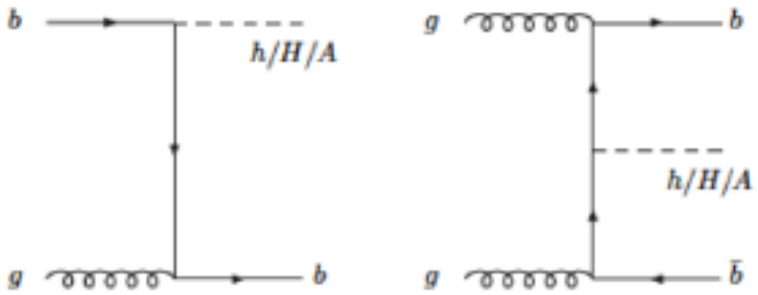

Figure 5.1: Feynman diagrams at LO for the production of Φ in association with bottom quarks.

for the process Z + $b\bar{b}$. In literature it is common to define two different approaches for the calculation of the inclusive cross section of Z+b: the *fixed flavor number scheme* (FFNS) and the *variable flavor number scheme* (VFNS). The former is often referred as *4 flavor scheme* (4FS) because only massless partons densities are considered for the initial state, while the latter is often called *5 flavor scheme* because a b-quark density of the proton is introduced. The main difference between the two schemes when used to compute the Z+b cross section is that in the FFNS the perturbative series is ordered by powers of the strong coupling constant α_S while in the VFNS the presence of the PDF of the b quark allows to resum terms like $\alpha_S^n \log(\frac{m_b^2}{M^2})^m$, where M is the upper integration limit of the transverse momentum of the b quark. At leading order the two schemes give different results, but at NLO the two schemes should give compatible results.

5.3.1 Z+b production in the FFNS

In the FFNS, at the leading order there are two subprocesses contributing to Z+b: $q\bar{q} \to Zb\bar{b}$ and $gg \to Zb\bar{b}$, where q can be a u,d,s,c quark. The tree level diagrams contributing to the LO cross section are shown in Figure 5.2 and Figure 5.3.

At NLO one has to consider also one-loop virtual corrections to the processes, proportional to α_S^2, and real subprocesses, proportional to α_S^3, as shown in Figure 5.4 and Figure 5.5.

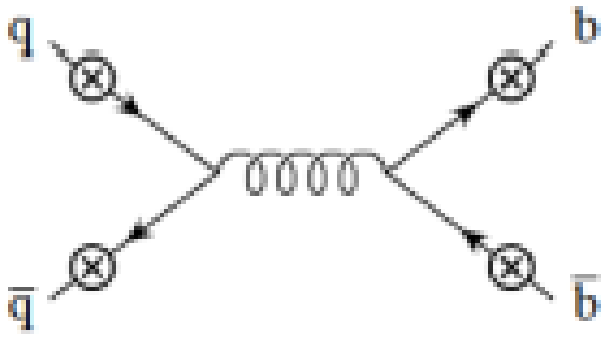

Figure 5.2: Tree level Feynman diagram for $q\bar{q} \rightarrow Zb\bar{b}$. The circled crosses correspond to all the possibilities to insert a Z boson, each one representing a different diagram (FFNS).

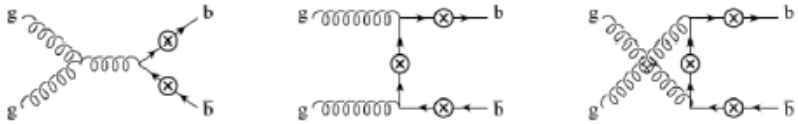

Figure 5.3: Tree level Feynman diagrams for $gg \rightarrow Zb\bar{b}$. The circled crosses correspond to all the possibilities to insert a Z boson, each one representing a different diagram (FFNS).

5.3.2 Z+b production in the VFNS

In the VFNS, the LO process is $gb \rightarrow Zb$ shown in Figure 5.6. Other compatible diagrams are those with the production of a $b\bar{b}$ pair from gluon splitting where the $b\bar{b}$ pair is collinear and it is reconstructed as a unique b-tagged jet. NLO corrections can be applied, such that additional partons are produced in the final state in process like $bg \rightarrow Zbg$, $bq \rightarrow Zbq$ or $qg \rightarrow Zb\bar{b}q$, represented in Figure 5.7.

5.4 Measurement of the Z+b cross section

CMS has published the measurement of the cross section for the inclusive process Z+b using data recorded at a center of mass energy of 7 TeV, corre-

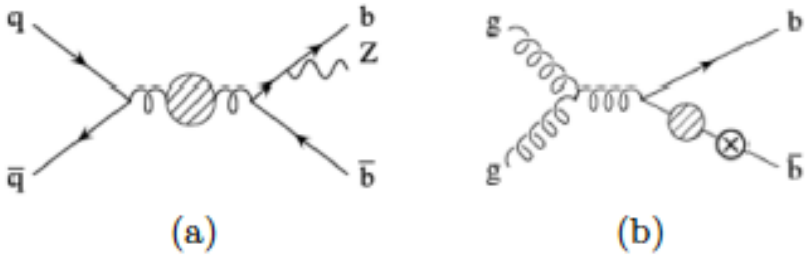

Figure 5.4: One loop virtual corrections to $q\bar{q} \to Zb\bar{b}$ (a) and to $gg \to Zb\bar{b}$ (b) (FFNS).

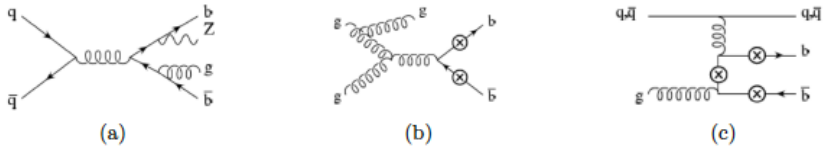

Figure 5.5: Real corrections to the cross section $Zb\bar{b}$ (FFNS).

sponding to an integrated luminosity of 2.2 fb^{-1}. The measured cross section with $Z \to ll$ (where ll = ee or $\mu\mu$) for events with the invariant mass of the dileptons between 60 and 120 GeV, at least one b-jet at the particle level [1] with $p_T > 25$ GeV and $|\eta| < 2.1$, and a separation between the leptons and the jets of $\Delta R > 0.5$ is found to be $5.84 \pm 0.08(stat) \pm 0.72(syst)^{+0.25}_{-0.55}(theory)$ pb. This cross section is higher than the one predicted by the program MCFM[2], corrected at particle level, being (3.97 ± 0.47) pb. Furthermore there are few differences in the differential distributions of same variables, as the azimuthal angle between the b quark and the dilepton, between the experimental measurement and the prediction. One point to be noticed is that the theoretical predictions are more accurate when the b-jet is the leading jet in the event. This requirement was not used in the reference CMS analysis. Therefore for this thesis it was decided to determine an inclusive cross section requiring explicitly the b-jet to be the leading p_T jet in the event. Different from the published analysis, here we will consider only final state with a dimuon. The measurement is performed using reconstructed (*reco*) quantities and has to be compared with the theoretical predictions, which are made in terms of

[1] The particle level considers hadrons before their decay.
[2] In Appendix B there as some studies about performed using this tool.

Figure 5.6: LO diagrams for $gb \to Zb$ where Q represents a b quark (VFNS).

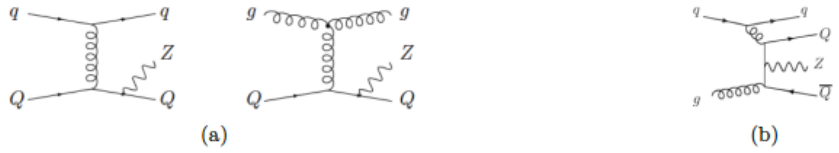

Figure 5.7: NLO diagrams for $qQ \to gZQ$ (a) and $gQ \to gZQ$ (VFNS).

leptons (before or after final state radiation emission) and a jets made of hadrons. The formula to compute the cross section from the numbers of selected events containing different corrections factors is:

$$\sigma = \frac{N(\mathcal{P} - f_{t\bar{t}})}{\epsilon_\beta \epsilon_l \epsilon_b \mathcal{A}_l \mathcal{C}_{had} \mathcal{L}}$$

where:

- N is the number of selected events (yield);
- $\mathcal{P}$ is the purity of b-tagging;
- $f_{t\bar{t}}$ is the background fraction from $t\bar{t}$ events;
- ϵ_β is efficiency for the association jet-to-vertex;
- ϵ_l is the lepton identification efficiency;
- ϵ_b is the efficiency of b-tagging;
- $\mathcal{A}_l$ is the acceptance of the leptons;
- $\mathcal{C}_{had}$ is a correction factor which takes into account the passage from generated Monte Carlo quantities to reconstructed quantities;

- $\mathcal{L}$ is the integrated luminosity.

The term ϵ_β is introduced as a consequence of the requirement on the leading jet, because jets not coming from the same PV of the dilepton, called *pile up* vertices, have to be rejected. In the data used for the analysis the mean of number of reconstructed PV was 6. The jet-to-vertex association fraction β is defined as the total p_T of the jet carried away by the tracks associated to the dilepton vertex:

$$\beta = \frac{\sum_{i \in vtx,jet} p_T^i}{\sum_{i \in jet} p_T^i}$$

where i represents the tracks associated to the jet. A sketch illustrating possible events coming from vertices of *pile up* is shown in Figure 5.8. In

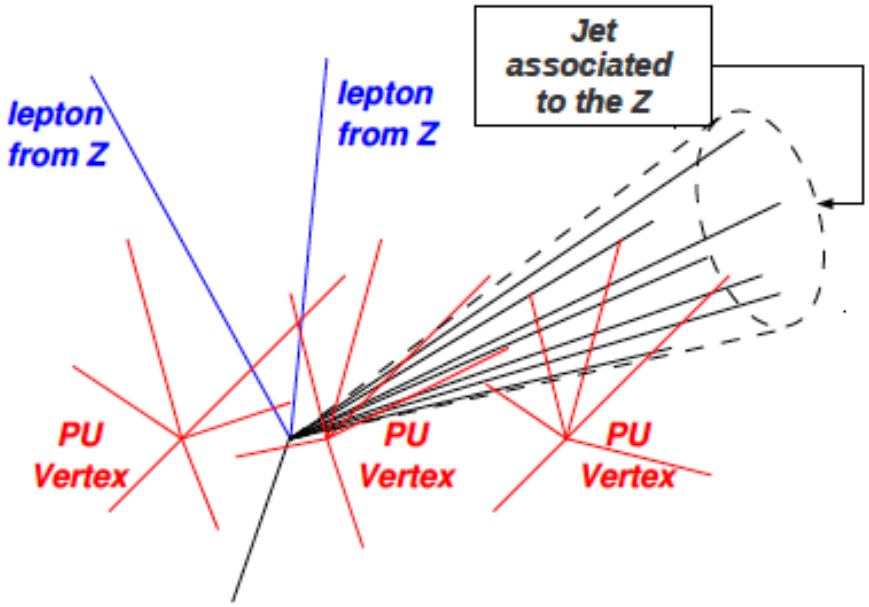

Figure 5.8: Simple sketch for possible events coming from vertices of *pile up*.

order to describe how to compute these correction factors in the cross section it is useful to define first the following categories of events:

- gen$_Z$: events with leptons, before final state radiation emission, of desired flavor (μ) with opposite charge and with the invariant mass of the dilepton $60 < M_{ll} < 120$ GeV;

- gen$_{Z_{kin}}$: gen$_Z$ events with the generated leptons passing η and p_T acceptance cuts;

- gen_J: events with the generated jet[3] passing η and p_T cuts isolated with a cone of $\Delta R > 0.5$ with respect to the generated leptons;
- gen_b: gen_J events with the jet matched [4] with a B-hadron;
- $\text{reco}_{Z_{kin}}$: reconstructed leptons matched with generated leptons passing all η and p_T cuts including the cut in the invariant mass $60 < M_{ll} < 120$ GeV;
- reco_J: reconstructed jet passing all η and p_T acceptance cuts, out of a cone of radius $\Delta R = 0.5$ with the reconstructed leptons and matched with a generated jet of any flavor;
- reco_b: reconstructed jet passing all η and p_T cuts, out of a cone of ΔR=0.5 with the reconstructed leptons and matched with a gen_b.

Finally the previous factors in the cross section can be evaluated from counting events in the categories described above:

$$\mathcal{A}_l = \frac{N(gen_{Zkin} \& gen_b)}{N(gen_Z \& gen_b)}$$

$$\mathcal{C}_{had} = \frac{N(reco_{Zkin} \& reco_b)}{N(gen_{Zkin} \& gen_b)}$$

$$\epsilon_l = \frac{N(reco_{Zkin+ID+ISO} \& reco_b)}{N(reco_{Zkin} \& reco_b}$$

$$\epsilon_b = \frac{N(reco_{Zkin+ID+ISO} \& reco_{btag})}{N(reco_{Zkin+ID+ISO} \& reco_b)}$$

where *ID + ISO* means that the leptons passed the identification and isolation cuts and *b-tag* means that the jet passed the cut for the working point of the discriminant chosen for the analysis.

5.4.1 Determination of the correction factors

Jet-vertex association

For jets not coming from PU events, β follows over a wide range essentially an exponential distribution $\sim \exp \frac{\beta}{\beta_0}$. For $\beta \to 0$ the peak from PU vertices dominates while for $\beta \to 1$ the distribution reaches a saturation. To compute ϵ_β we perform the following steps:

[3] A generated jet is a jet made at hadron level using the simulation.

[4] The matching is done using a cone algorithm with a cone size $\Delta R < 0.3$.

- determine β_0 from a fit in a fiducial region far from saturation and pile-up events;
- evaluate the cumulative distribution function from 0 up to an upper limit β_{MAX} below the saturation;
- add the content of the histogram for bins with $1 < \beta < \beta_{MAX}$.

The formula for ϵ_β is:

$$\epsilon_{beta} = \frac{\int_0^{\beta_{MAX}} PDF(\beta)}{\int_0^{\beta_{MAX}} PDF(\beta) + N_{dataabove}}$$

The background is defined as all the data in excess to the exponential in the fit region and below. Figure 5.9 shows, as an example, how the different regions are selected.

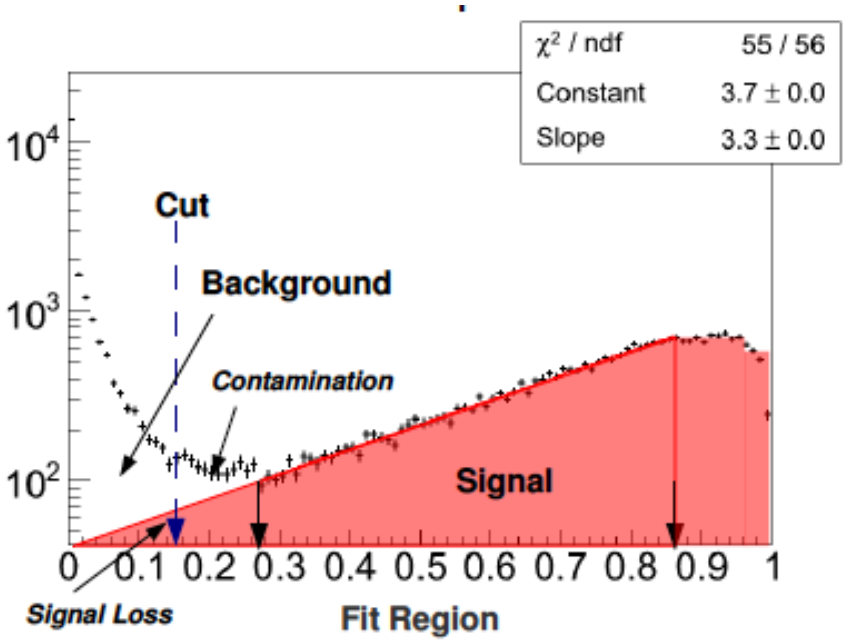

Figure 5.9: Example of definition of the regions from the exponential fit for β.

Exponential fits of the signal region are done in bins of number of primary vertices, from 1 to 16 primary vertices. The mean value $< \epsilon_\beta >$ is the weighted mean of the efficiencies for each bin:

$$< \epsilon_\beta >= \frac{\sum_{i=0}^{nPU} (\epsilon_\beta^i \times N_{DATA}^i)}{\sum_{nPU}^{i} N_{DATA}^i}$$

B-tagging efficiency ϵ_b

Because the efficiencies of tagging a b-jet depend on analysis specific selections, the BTV group provides only scale factors, SF_b, defined as the ratio between the efficiency of b-tagging in data with respect to the efficiency of b-tagging in the Monte Carlo samples. The advantage is that these scale factors can be used to estimate the efficiency for a specific selection. The scale factors are computed in two bins of the η of the jet: $|\eta| < 1.2$ and $1.2 < |\eta| < 2.4$. They are provided for different discriminants and working points. The final efficiency of the b-tagging selection is extracted from the signal Monte Carlo sample, correcting the MC using the scale factors. The procedure for applying the SF_b to the analysis can be summarized as follows:

1. the jets selected by the cut of the b-tag discriminant are classified according to the generator level information into b,c, or light flavors;
2. the scale factors defined above are associated to these jets, depending on their η and p_T;
3. select the desire b-jet multiplicity: at least one in the case of inclusive Z+b measurement or exactly one in the case of an exclusive measurement;
4. if n_{jets} is the total number of jets in the final state, there are $2^{n_{jets}}$ possible combinations of tagged-mistagged states for each event;
5. determine the probability for this event to be reconstructed, using the tag efficiency and the mistag-fraction;
6. repeat this operation including the scale factors;
7. the total weight per combination is given by the product over all jet states;
8. sum only the possible combinations matching the selected multiplicity;
9. define the per event-weight as follows:

$$w_{evt} = \frac{\prod_{i=N_{jets}}(\epsilon_i(p_t, \eta_j) \times SF(p_T, \eta_j))}{\prod_{l=N_{jets}}(\epsilon_l(p_{T,k}, \eta_k))}$$

The final efficiency ϵ_b used for the measurement of the cross section is defined as the ratio of events after b-tag cut over the ratio of events before b-tag cut, when there is at least one jet matched to a generated-level b in the acceptance.

Lepton efficiency ϵ_l

To estimate the lepton identification efficiency Monte Carlo events need to be rescaled according to some correction factors, in order to match the data. The efficiency can be factorized into five sequential relative measurements, the five steps are: tracking efficiency, reconstruction efficiency, identification efficiency, isolation efficiency and the total trigger efficiency, underlying that in the Monte Carlo there are no trigger requirements. The total efficiency is given by the product:

$$\epsilon_l = \epsilon_{tracking} \times \epsilon_{RECO} \times \epsilon_{ID} \times \epsilon_{ISO} \times \epsilon_{trigger}$$

Each term is considered separately and the values are estimated using a *tag and probe* method, which uses known resonances in bins of p_T and η of the leptons, while the efficiency of tracking is assumed to be 100%. In this analysis the Z resonance has been chosen to measure the efficiency with this method. It requires a mass constraint from a pair of same flavor leptons where one of the two is tightly selected to ensure a certain purity, *the tag*, and the other, the *probe*, is used to measure the efficiencies for a given identification criterium. The efficiency is defined as the ratio of number of probes passing a set of cuts over the total number of probes before the cuts. A set of correction factors can be extracted in order to properly rescale simulated events for matching the data. There is a set of correction factors for each of the five efficiencies described above. In order to provide the final weight w for each event all the correction factors are combined into two terms:

1. a contribution related to the offline reconstruction: reco, ID and isolation ($w_{offline}$);
2. a contribution related to the trigger ($w_{trigger}$);

A scale factor is defined as the ratio between the efficiency in data over the efficiency in the Monte Carlo. Since there are two leptons in each event the final weight to due to the offline reconstruction is:

$$w_{event}^{offline} = SF_{lepton1}^{offline} \times SF_{lepton2}^{offline}$$

The formula for a generic trigger having asymmetrical thresholds, labelled *high* and *low*, is instead:

$$w_{event}^{trigger} = SF_{lepton1}^{High} \times SF_{lepton2}^{low} + SF_{lepton1}^{low} \times SF_{lepton2}^{high} - SF_{lepton1}^{high} \times SF_{lepton2}^{high}$$

Ultimately the final weight can be expressed as:

$$w_{lepton}(\eta, p_T) = w_{offline}(\eta, p_T) \times w_{trigger}(\eta, p_T)$$

The total lepton efficiency ($\epsilon_{leptons}$) is the ratio between the total number of events, $N^{tot,w}$, passing the kinematical cuts multiplied by the respective weights with respect to the total number of events passing the kinematical cuts N_{tot}:

$$\epsilon_{leptons} = \frac{N^{tot,w}}{N_{tot}}$$

where:

$$N^{tot,w} = \sum_{ev} \mu^+\mu^-(p_T > 20, \eta < 2.1) \times w^{\mu}_{event}$$

$$N_{tot} = \sum_{ev} \mu^+\mu^-(p_T > 20, \eta < 2.1)$$

and the sum is intended for all the lepton events. Each of the lepton events entering in the efficiency computation need to have a reconstructed jet matching a generated one in the acceptance window.

b-tagging purity $\mathcal{P}$

The use of the SSVHP discriminant allows to build the distribution of the invariant mass of the set of tracks associated to the secondary vertex both in the data and in the Monte Carlo. From Drell Yan inclusive (Z+b/c/l) Monte Carlo samples it is possible to assign the flavor of the jet from the flavor of the corresponding generated jet: this allows to build the templates of the distribution of the SV invariant mass for light jets, b-jets and c-jets. These templates are compared to the distribution of the SV invariant mass obtained from the data. There are two different computational approaches to extract the b-purity. The first method performs an extended maximum likelihood fit: using the tool *RooFit* to fit the areas of the different templates for each flavor, it is possible to extrapolate the fraction of different flavors for the working point chosen. The results obtained here are the fractions per jets which in our case, since there is only the leading jet selected per event, correspond to the fractions per event. The other method uses the ROOT framework class TFractionFitter: it takes Monte Carlo templates and tries to find components of them in a data histogram. This method, in addition with respect to the other, lets each bin of the templates fluctuating within its error, while in the former only the fractions fluctuate.

Fraction $t\bar{t}$ background

After the selection, in the mass region $60 < M_{ll} < 120$ GeV, there is still a contribution coming from the two muons from the decay chains of the top and anti-top quark. These $t\bar{t}$ events contain genuine b jets, so they need

to be subtracted after applying $\mathcal{P}$. This fraction is evaluated using a data driven method looking at the dilepton invariant mass distribution: the main idea is to estimate the $t\bar{t}$ contribution in the mass region $M_{ll} > 120$ GeV and extrapolate it to the signal region. The ratio of the number of $t\bar{t}$ events is taken from Monte Carlo simulation and then corrected from discrepancies between data and simulation. The estimation of the $t\bar{t}$ background is done using the formula:

$$N_{t\bar{t}}^{ext}(in) = \frac{\mathcal{R}^{MC}}{\mathcal{R}_Z^{MC} - \mathcal{R}^{MC}} \times (\mathcal{R}_Z^{MC} N_{obs}(out) - N_{obs}(out))$$

where:

$$\mathcal{R}^{MC} = \frac{N_{t\bar{t}}^{MC}(in)}{N_{t\bar{t}}^{MC}(out)}$$

In the previous formula the region with $60 < M_{ll} < 120$ GeV is defined as *in* (inside) and the region with $M_{ll} > 120$ GeV as *out* (outside).

5.4.2 Selection of events using the CMS detector

Monte Carlo and data samples

The reference Monte Carlo events used in the analysis were generated using MADGRAPH, including the parton shower and hadronization using PYTHIA:

1. Z+b, Z+c and Z+j samples: they are all extracted from the data set corresponding to the inclusive Drell Yan;

2. $t\bar{t}$ + j sample: events with a $t\bar{t}$ pair in association with any number of jets;

3. WW, ZZ and ZW samples: they represent events with the presence of two diboson.

All the samples need to be rescaled to the equivalent luminosity which is defined as:

$$\mathcal{L}_{eq} = \frac{N_{evt}}{\sigma}$$

where the cross section used to obtain $\mathcal{L}_{eq}$ was taken from the state-of-art from NLO calculations. In Figures 5.10 and 5.11 the properties of Monte Carlo samples and data samples, including trigger informations, are summarized.

Dataset	Sample	N_{evts}	NLO σ (pb)	$\mathcal{L}_{eq}$ (pb^{-1})
ZbbToLL	Z+b	26840	26.3	1020
ZccToLL	Z+c	40000	41.1	973
DYJetsToLL	Z+j	5172930	3048 (NNLO)	1697
TTJets-tauola	tt+j	200000	165 (NNLO)	1212

Figure 5.10: Monte Carlo samples used in the analysis.

Muon	135821-149442	HLT_Mu9 OR HLT_Mu11 OR HLT_Mu15

Figure 5.11: 2011 data sample used and triggers.

Dilepton selection

The Z boson is reconstructed through its decay into a pair of muons with opposite charge and same flavor. The acceptance cuts applied are the same used for the analyses in CMS dealing with massive vector bosons are: $p_T^{\mu} > 20$ GeV, $|\eta| < 2.1$. The isolations cuts are based on the sum of p_T and E_T of tracks lying in a cone of dimension $\Delta R = 0.3$ centered around the leptons: $\frac{\sum_{TRK} p_T + \sum_{ECAL} p_T + \sum_{HCAL} p_T}{p_T} < 0.15$, where TRK stands for the tracker, ECAL for the electromagnetic calorimeter, HCAL for the hadronic calorimeter; For the identification muons are required to be reconstructed in the muon chambers and then matched with a track in the tracker (referring to them as *global muons*). In the data there is also a trigger match: muons need to be matched into a $\Delta R < 0.3$ cone. Finally the Z boson is reconstructed into a mass range from 60 GeV to 120 GeV.

Jet selection

The jets are reconstructed using an anti k_T cone algorithm with *particle flow* objects, choosing $\Delta R = 0.5$. A *good jet* is defined according to the following cuts:

- kinematics cuts: $p_T^j > 25$ GeV, $|\eta| < 2.1$;
- standard jet identification cut based on the deposits into ECAL and HCAL of charged and neutral fractions of jets.

Finally a jet is considered as a b-jet if it is selected by the SSVHP discriminant at a tight working point:

$$\log(1 + \frac{D}{\sigma_D}) > 2.0$$

leading to a mistag rate of about 0.1%.

5.4.3 Results

The results for the evaluation of the correction factors in the cross section are summarized in Table 5.1.

Correction Factor	Value
ϵ_b	(0.35 ± 0.05)
ϵ_l	(0.92 ± 0.03)
$\mathcal{A}_l$	(0.59 ± 0.03)
C_{hadr}	(0.73 ± 0.04)
ϵ_β	(0.99 ± 0.02)
$\mathcal{P}$	(0.81 ± 0.02)
$f_{t\bar{t}}$	(0.147 ± 0.01)

Table 5.1: Correction factors for the channel $\mu\mu$b: only statistical errors are reported

Figure 5.12 shows the fit to estimate the b-tagging purity.

Table 5.2 shows the numbers of selected events: each row represents a different step in the selection, while each column represents the different Monte Carlo samples, excluding the last column which represents the data sample.

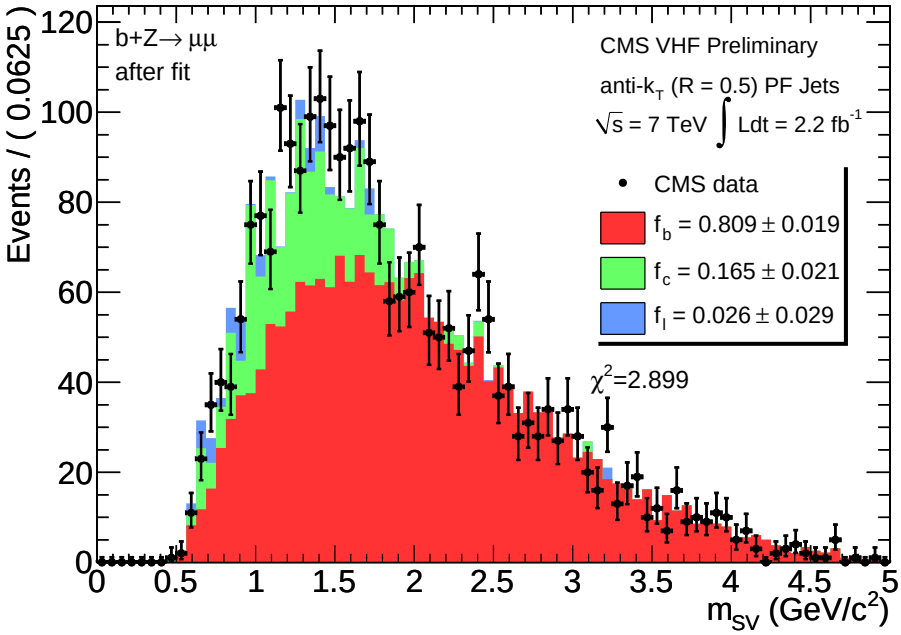

Figure 5.12: Final fit for the estimation of $\mathcal{P}$ for the channel $\mu\mu$.

Cut/Sample	$\tau\tau$	$Z(\tau\tau)$	ZZ	ZW	Z+b	Z+c	Z+l	SumMC	Data
ll	2624 ± 36	5309 ± 122	445 ± 2	728 ± 6	45323 ± 122	207914 ± 830	541672 ± 1342	80405 ± 1588	824790 ± 908
ll tight	2038 ± 32	1930 ± 83	386 ± 2	633 ± 5	40635 ± 118	188681 ± 802	494860 ± 1300	729162 ± 1535	760630 ± 855
l-HLT match	2038 ± 32	1930 ± 83	386 ± 2	633 ± 5	40635 ± 118	188681 ± 802	494860 ± 1300	729162 ± 1535	760630 ± 855
Z(ll)	1008 ± 23	880 ± 56	371 ± 2	616 ± 5	39819 ± 116	183012 ± 790	481630 ± 1282	707337 ± 1512	700836 ± 837
Z+j	973 ± 22	146 ± 22	265 ± 2	435 ± 5	10279 ± 59	38884 ± 364	81997 ± 528	132979 ± 645	126868 ± 356
Z+j(vtx)	973 ± 22	146 ± 22	265 ± 2	435 ± 5	10279 ± 59	38884 ± 364	81997 ± 528	132979 ± 645	126868 ± 356
Z >=1b (HP)	376 ± 13	0 ± 0	19 ± 0	5 ± 0	1768 ± 23	442 ± 37	33 ± 10	2642 ± 47	2592 ± 51
Z=1b (HP)	376 ± 13	0 ± 0	19 ± 0	5 ± 0	1768 ± 23	442 ± 37	33 ± 10	2642 ± 47	2592 ± 51

Table 5.2: Selected events from Monte Carlo samples and data sample for each step of selection

Figure 5.13 shows the p_T of the b-jet: there is a good agreement between data (black points) with respect to the simulated events. Figures 5.14 and 5.15 show respectively the p_T of the dimuon and the azimuthal distance between the dimuon and the b-jet: there is a good agreement between data (black points) with respect to the prediction for the simulation.

The number of selected events in data was found to be (2592 ± 51) which has a good agreement with the number of data expected by the Monte Carlo which is (2642 ± 47). The value for the leading cross section is: $\sigma_{leading} = (5.69 \pm 0.11(stat))$ pb which needs to be compared to the inclusive predicted cross section with MCFM, corrected at particle level, being $\sigma_{MCFM} = (3.56 \pm 0.11)$ pb. The most interesting result is that in MCFM the cross section in which the leading jet is a b jet represents 90% of the total inclusive cross section. The ratio between the exclusive cross section and the total inclusive cross section ($\sigma = 6.29 \pm 0.11$) pb , both measured experimentally, is of about 10%, with a good agreement with the prediction.

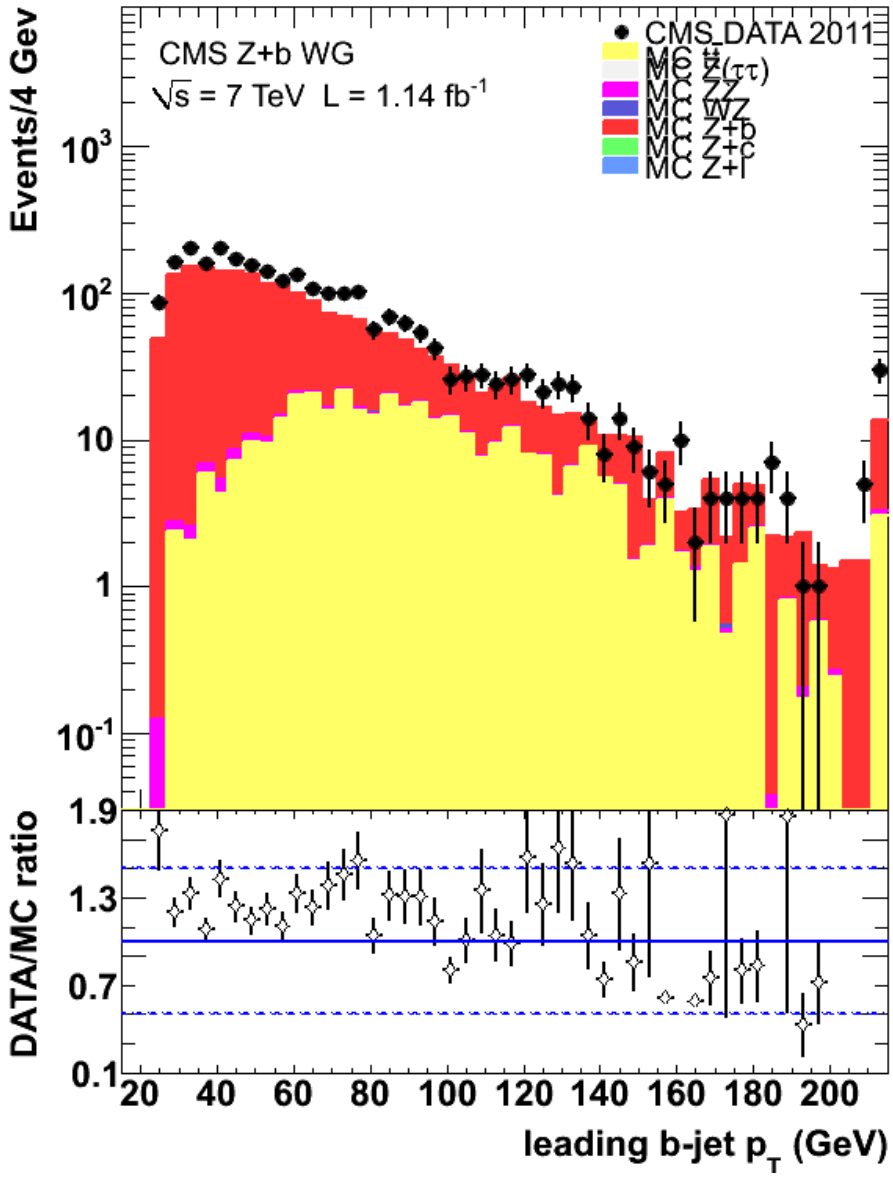

Figure 5.13: Distribution of the p_T of the leading b-jet. The ratio is intended as the data over the Monte Carlo.

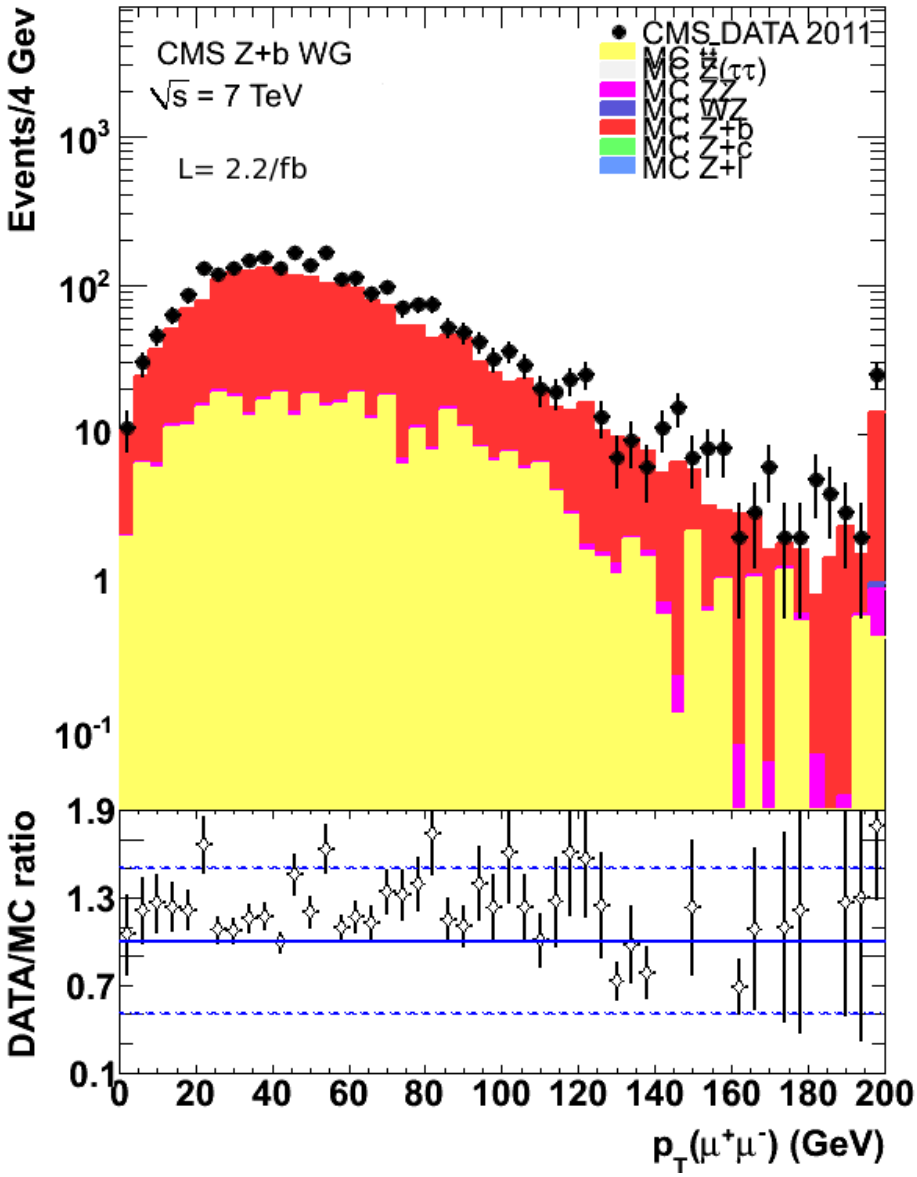

Figure 5.14: Distribution of the p_T of the dimuon. The ratio is intended as the data over the Monte Carlo.

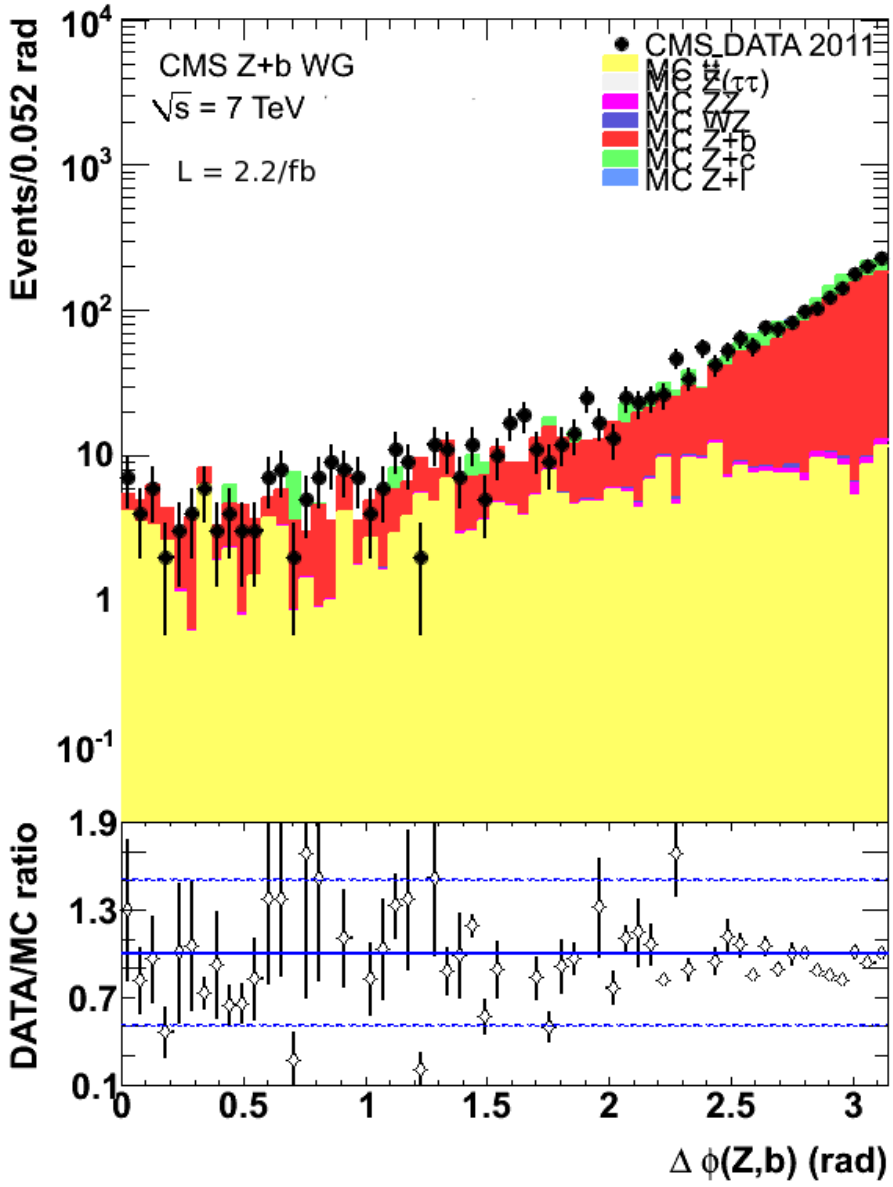

Figure 5.15: Distribution of the azimuthal distance $\Delta\Phi$ between the dimuon and the b-jet. The ratio is intended as the data over the Monte Carlo.

Chapter 6

Conclusions

The work of this thesis has allowed to improve the knowledge about the b-tagging algorithms used by the CMS experiment and to apply them to perform a measurement of a physical process interesting both as a reducible background for the searches of a Standard Model Higgs boson and for SUSY scenarios. In the first part a deep validation of the b-tagging algorithms has been performed introducing systematic misalignment conditions in phase of the reconstruction of the tracks. It was shown that the official tools provided by the CMS to perform these studies are not sensible for more realistic misalignment scenarios and do not permit to evaluate correlations between all the variables from different scenarios: this fact motivated the decision to build a diagnostic tool to evaluate differences jet-by-jet. The second part of the work was devoted to the construction of this algorithm and to apply it on the scenarios studied before. This tool has shown a very powerful versatility, because it allows to make comparisons jet-by-jet of two files which differ for any kind of condition when reconstructing tracks: it has been used by several groups of CMS at CERN doing studies on b-tagging variables. The global result of applying this tool to different misalignment scenarios confirmed that the global variables, such as track or reconstructed vertices, used by the CMS experiment are very robust, even when applying a deep diagnostic and not only when looking at global performances of b-tagging variables, such as the efficiency of tagging a light jet (related to the mistag rate) as a function of the efficiency of tagging a b-jet. After this increased knowledge about all the b-tagging algorithms it was decided to use it to measure a process in which the tagging of a b quarks is fundamental: the Z+b cross section. An aspect which was not investigated in the previous measurement of the cross section of this process by the CMS group is that the theoretical predictions are more precise where the b-jet is the leading jet in the event. The computation for this cross section, only for the channel $\mu\mu b$ was performed with

good results between theory and the experimental value. There are still open points: understand why the experimental cross section is always bigger than the predicted and perform the calculation for the channel eeb.

Appendix A

Studies of misalignment scenarios with the jet-by-jet comparison tool

This appendix presents a set of correlation plots for the scenarios studied with the jet-by-jet comparison tool. The first row shows some distributions for the differences of the discriminant TCHE as a function of the SIP of the second track, the difference of the position of the primary vertex in the transverse plane (ΔL_{XY} (PV)) and the longitudinal difference of position of the primary vertex (Δz (PV)). The second row shows differences of the discriminant SSVHP as a function of the mass of the SV, the three dimensional SIP and the difference of the position of the primary vertex in the transverse plane and in the longitudinal plane. The third row shows differences of the discriminant SSVHE as a function of the mass of the SV and the three dimensional SIP. The first set of figures refers to the design geometry (*Ideal*) compared with the design geometry setting APE equal to zero (*Ideal No APE*). The second set of figures refers to the Monte Carlo *2011 Startup* scenario with respect to the *ideal* geometry. The third set of figures refers to the scenario BPIX $\Delta z = 40$ μm with respect to the *ideal* geometry.

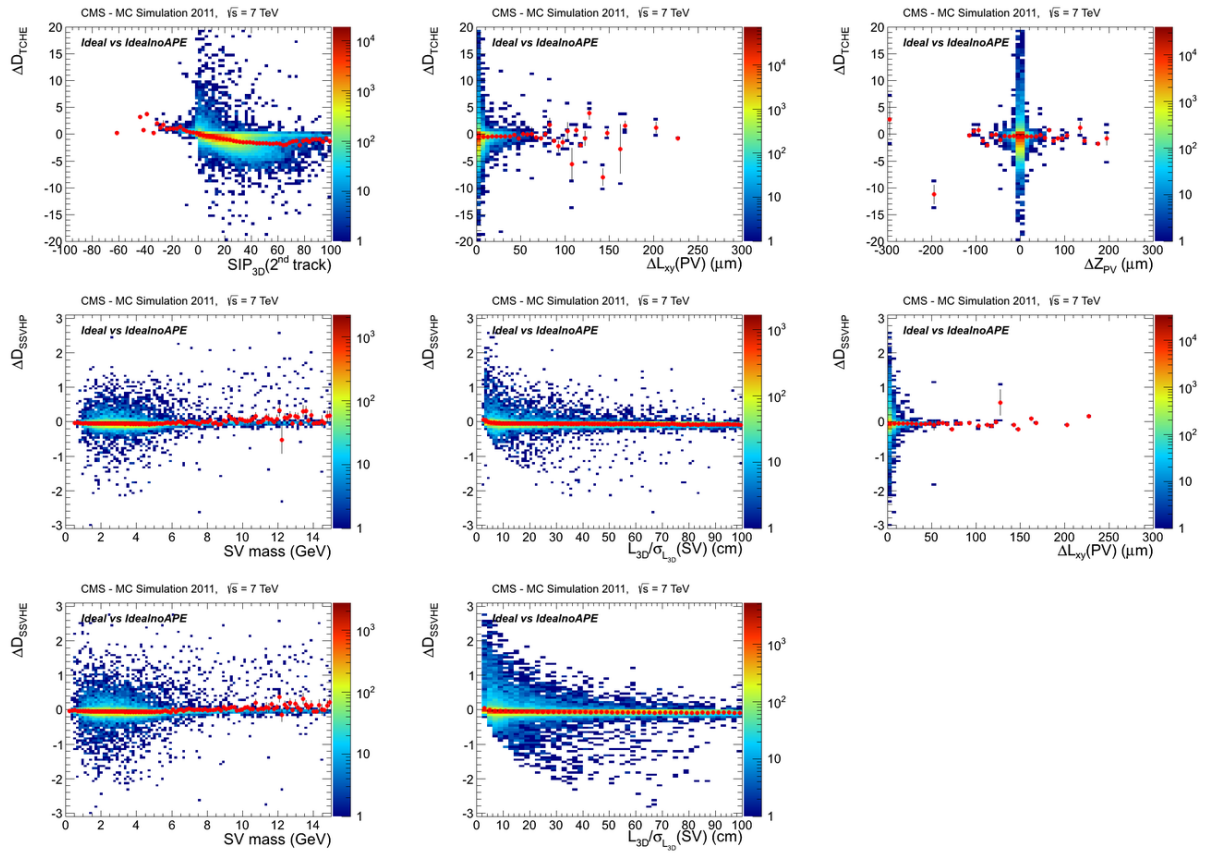

Figure A.1: Jet-by-jet differences for the TCHE,SSVHP,SSVHE discriminants as a function of few characteristic variables for the *Ideal* geometry versus *Ideal* geometry with APE = 0.

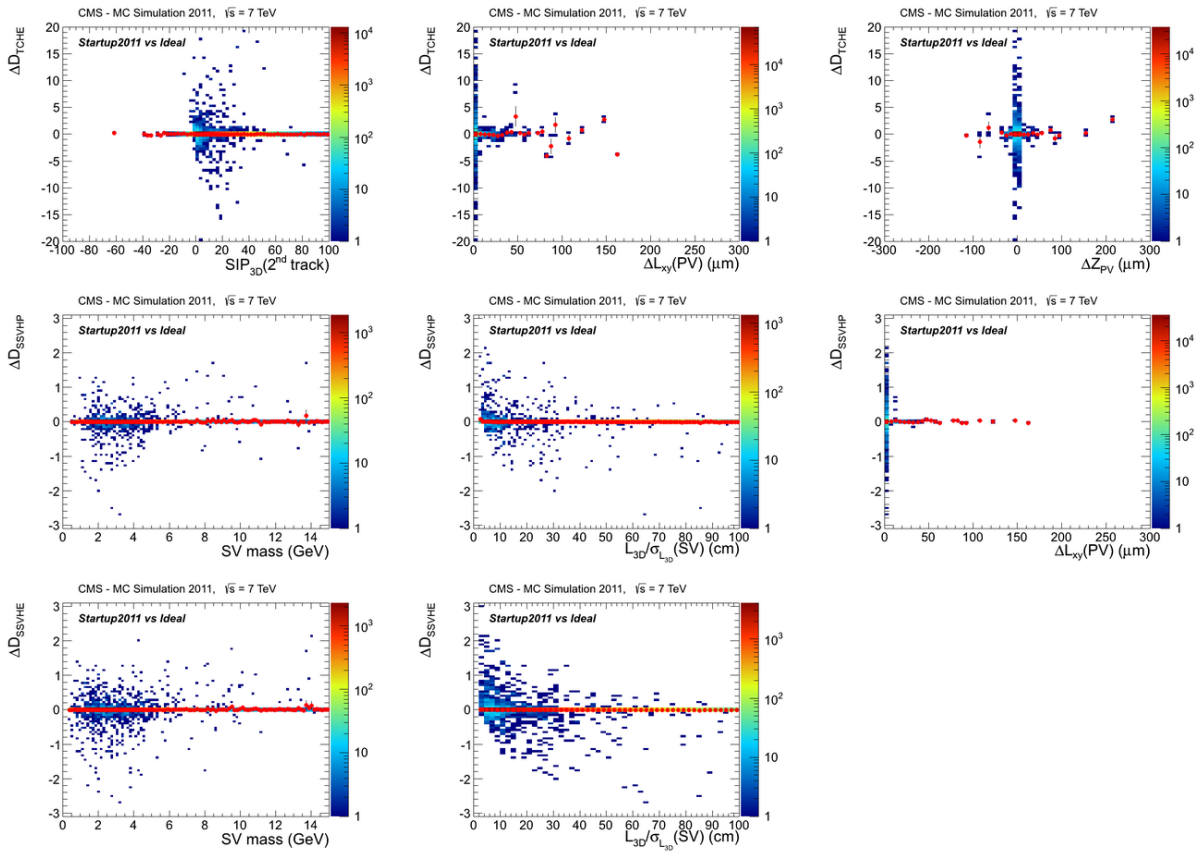

Figure A.2: Jet-by-jet differences for the TCHE,SSVHP,SSVHE discriminants as a function of few characteristic variables for the *Startup 2011* geometry versus *Ideal* geometry.

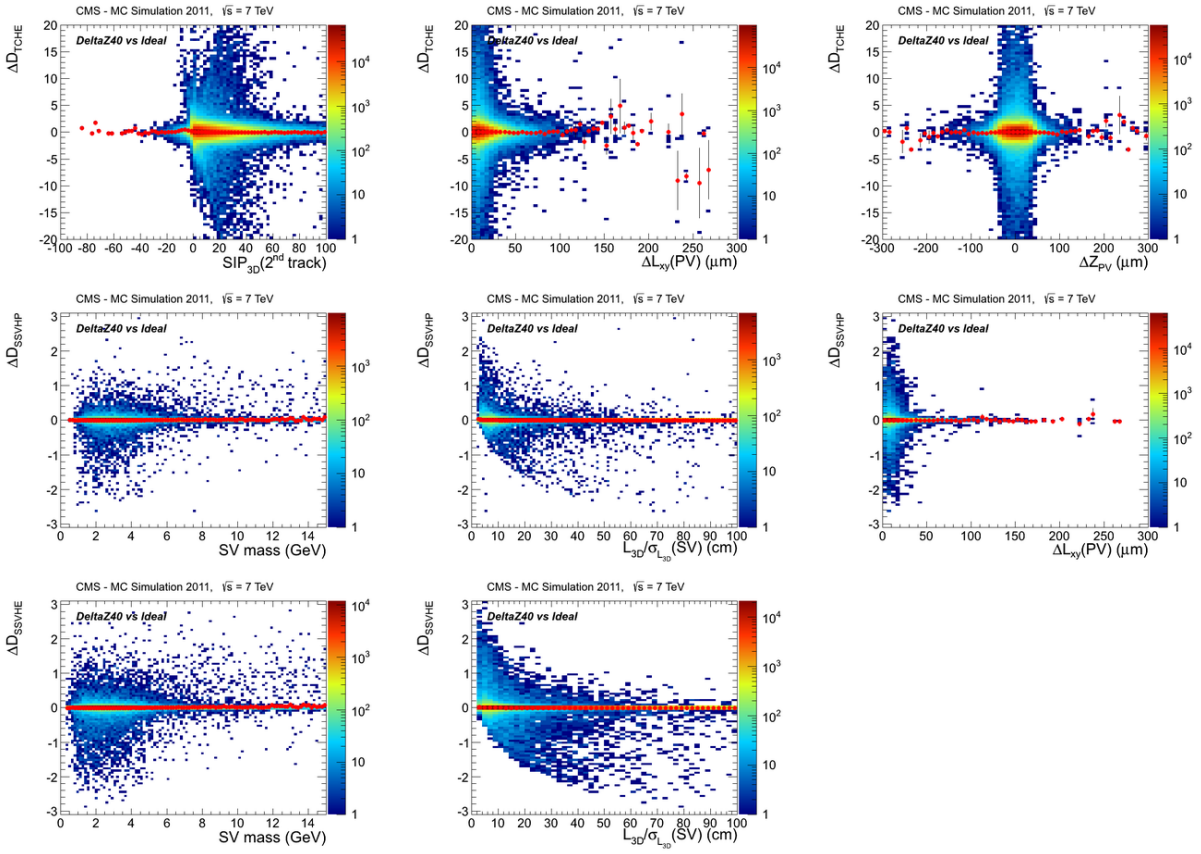

Figure A.3: Jet-by-jet differences for some discriminants as a function of few characteristic variables for the Δz 40 μm geometry versus Ideal geometry.

Appendix B

Parton level studies with the MCFM tool.

The MCFM (Monte Carlo for FEMtobarn processes at hadron colliders) is a parton-level program which gives NLO predictions for a range of processes at hadron colliders. MFCM allows the user to choose between a number of schemes for defining the electroweak couplings as shown in Figure B.1. The

Parameter	Name (_inp)	Input Value	Output Value determined by ewscheme			
			-1	0	1	2
G_F	Gf	1.16639×10^{-5}	input	calculated	input	input
$\alpha(M_Z)$	aemmz	1/128.89	input	input	calculated	input
$\sin^2\theta_w$	xw	0.2312	calculated	input	calculated	input
M_W	wmass	80.398 GeV	input	calculated	input	calculated
M_Z	zmass	91.1876 GeV	input	input	input	calculated
m_t	mt	input.DAT	calculated	input	input	input

Figure B.1: Different options for the scheme used to fix the electroweak parameters of the Standard Model and the corresponding default input values.

process to compute is selected by the integer *nproc*. Processes denoted as *LO* can only be calculated using the Born approximation, while for processes denoted as *NLO+F* the calculation is done using NLO calculations including the fragmentation of the photons. The processes of interest for the production of a Z boson in association with heavy quarks are summarized in the Table B.1.

Others important parameters to the describe the diagrams included in the computations are:

- *lord*: the calculation is performed at the leading order only;

nproc	$f(p_1) + f(p_2) \to \ldots$	Order
261	$Z^0(\to e^-(p_3) + e^+(p_4)) + b(p_5)$	NLO
262	$Z^0(\to e^-(p_3) + e^+(p_4)) + c(p_5)$	NLO
263	$Z^0(\to e^-(p_3) + e^+(p_4)) + \bar{b}(p_5) + b(p_6)$[1 b-tag]	LO
264	$Z^0(\to e^-(p_3) + e^+(p_4)) + \bar{c}(p_5) + c(p_6)$[1 c-tag]	LO
266	$Z^0(\to e^-(p_3) + e^+(p_4)) + b(p_5)(+\bar{b}(p_6))$	NLO
267	$Z^0(\to e^-(p_3) + e^+(p_4)) + c(p_5)(+\bar{c}(p_6))$	NLO
341	$f(p_1) + b(p_2) \to Z^0(\to e^-(p_3) + e^+(p_4)) + b(p_5) + f(p_6)[+f(p_7)]$	NLO
342	$f(p_1) + b(p_2) \to Z^0(\to e^-(p_3) + e^+(p_4)) + b(p_5) + f(p_6)[+\bar{b}(p_7)]$	(REAL)
346	$f(p_1) + b(p_2) \to Z^0(\to e^-(p_3) + e^+(p_4)) + b(p_5) + f(p_6) + f(p_7)$	LO
347	$f(p_1) + b(p_2) \to Z^0(\to e^-(p_3) + e^+(p_4)) + b(p_5) + f(p_6) + \bar{b}(p_7)$	LO

Table B.1: Processes of interest for the production of a Z boson in association with heavy quarks.

- *virt*:NLO virtual loop contributions are calculated;
- *real*: NLO contributions for diagrams including real emission of gluons;
- *tota*: this options runs the *virt* and the *real* processes in series before performing a sum to obtain the NLO result;
- *inclusive*: it is a logical parameter to decide if the cross section should be inclusive in the number of jets found at NLO calculations.

There is also a set of parameters defining some properties such as the renormalization scale, or some characteristics related to the initial beam particles:

- *sqrts*: represent the center of mass energy of the colliding particles;
- *ih1,ih2*: identities of the incoming hadrons. For LHC the value +1 for both the parameters correspond to proton-proton collisions;
- *scale*: it is used to adjust the value of the renormalization scale, such as the scale at which the strong coupling constant α_S is evaluated. For processes involving vector bosons, the scale is choose as the average mass of the bosons involved;

Finally, the parameters related to the cuts one can apply to define the volume of the phase space where the cross section is computed. The default MCFM output of the process is an ASCII file containing the values of the computed cross section, with in addition few histograms related to some variables of

the jet and of the leptons. For this thesis we have also produced tuples of weighted evens,checking that the sum of all the weights was equal to the computed cross section. The processes chosen to compare the experimental results obtained in Chapter 4 to the predictions are labelled as *Z+Q production*. These processes represent the production of a Z boson that decays into a pair of leptons in association with a heavy quark Q. For processes 261,262,266 and 267 the initial state at the lowest order is the heavy quark and a gluon an the calculation may be performed at NLO. For the NLO calculation the sequence is:

1. run $nproc = 261$ using $part = tota$;
2. run $nproc = 266$ using $part = real$;
3. sum the two results.

The configuration of parameters chosen to run process 261 is listed below:

```
[General options to specify the process and execution]
261 [nproc]
'tota'   [part 'lord','real' or 'virt','tota']
'tota14TeV'['runstring']
14000d0 [sqrts in GeV]
+1 [ih1 =1 for proton and -1 for antiproton]
+1 [ih2 =1 for proton and -1 for antiproton]
120d0 [hmass]
91d0 [scale:QCD scale choice]
91d0 [facscale:QCD fac_scale choice]
.false. [dynamicscale]
.false. [zerowidth]
.false. [removebr]
10 [itmx1, number of iterations for pre-conditioning]
20000 [ncall1]
10 [itmx2, number of iterations for final run]
20000 [ncall2]
1089 [ij]
.false. [dryrun]
.true. [Qflag]
.true. [Gflag]

[Heavy quark masses]
172.5d0 [top mass]
```

```
4.75d0 [bottom mass]
1.5d0 [charm mass]

[Pdf selection]
'cteq6_m'[pdlabel]
4 [NGROUP, see PDFLIB]
46 [NSET - see PDFLIB]
cteq6mE.LHgrid [LHAPDF group]
-1             [LHAPDF set]

[Jet definition and event cuts]
60d0 [m34min]
120d0 [m34max]
0d0 [m56min]
14000d0 [m56max]
.true. [inclusive]
'ankt'[algorithm]
25d0     [ptjet_min]
0d0 [|etajet|_min]
2.1d0    [|etajet|_max]
0.5d0 [Rcut_jet]
.true. [makecuts]
0d0 [ptlepton_min]
100d0 [|etalepton|_max]
0d0 [ptmin_missing]
0d0 [ptlepton(2nd+)_min]
100d0 [|etalepton(2nd+)|_max]
0d0 [minimum (3,4) transverse mass]
0.5d0 [R(jet,lept)_min]
0.0d0 [R(lept,lept)_min]
0d0 [Delta_eta(jet,jet)_min]
.false. [jets_opphem]
0 [lepbtwnjets_scheme]
25d0     [ptmin_bjet]
2.1d0    [etamax_bjet]
```

while below for the process 266.

```
[General options to specify the process and execution]
```

```
266 [nproc]
'real'   [part 'lord','real' or 'virt','tota']
'real14TeV'['runstring']
14000d0 [sqrts in GeV]
+1 [ih1 =1 for proton and -1 for antiproton]
+1 [ih2 =1 for proton and -1 for antiproton]
120d0 [hmass]
91d0 [scale:QCD scale choice]
91d0 [facscale:QCD fac_scale choice]
.false. [dynamicscale]
.false. [zerowidth]
.false. [removebr]
10 [itmx1, number of iterations for pre-conditioning]
20000 [ncall1]
10 [itmx2, number of iterations for final run]
20000 [ncall2]
1089 [ij]
.false. [dryrun]
.true. [Qflag]
.true. [Gflag]

[Heavy quark masses]
172.5d0 [top mass]
4.75d0 [bottom mass]
1.5d0 [charm mass]

[Pdf selection]
'cteq6_m'[pdlabel]
4 [NGROUP, see PDFLIB]
46 [NSET - see PDFLIB]
cteq6mE.LHgrid [LHAPDF group]
-1            [LHAPDF set]

[Jet definition and event cuts]
60d0 [m34min]
120d0 [m34max]
0d0 [m56min]
14000d0 [m56max]
.true. [inclusive]
'ankt'[algorithm]
25d0     [ptjet_min]
```

```
0d0 [|etajet|_min]
2.1d0    [|etajet|_max]
0.5d0 [Rcut_jet]
.true. [makecuts]
0d0 [ptlepton_min]
100d0 [|etalepton|_max]
0d0 [ptmin_missing]
0d0 [ptlepton(2nd+)_min]
100d0 [|etalepton(2nd+)|_max]
0d0 [minimum (3,4) transverse mass]
0.5d0 [R(jet,lept)_min]
0.0d0 [R(lept,lept)_min]
0d0 [Delta_eta(jet,jet)_min]
.false. [jets_opphem]
0 [lepbtwnjets_scheme]
25d0         [ptmin_bjet]
2.1d0     [etamax_bjet]
```

Figure B.2 shows the distributions of the leading parton transverse momentum p_T^b for the sum of the processes 261 and 266: we notice that about the 10% of the total cross section is represented by events in which the leading jet comes from a light parton and that the p_T spectrum of the light jets is harder than the spectrum of the b jets. Figure B.3 shows the distribution of the azimuthal distance ($\Delta\phi$) between the dilepton and the parton: when the leading jet is a light jet there is cut a $\phi = \frac{\pi}{2}$ due to kinematical constraints. Figure B.4 shows the distribution of the azimuth distance ($\Delta\phi$) between the two partons: the distribution underlines the cut at $\Delta R = 0.5$. The cross section for events in which there are one light jet and one b-jet is found to be $\sigma_{Z+l+b} = 0.93$ pb, while the cross section for events in which there are two b-jets is $\sigma_{Z+b+b} = 0.23$ pb.

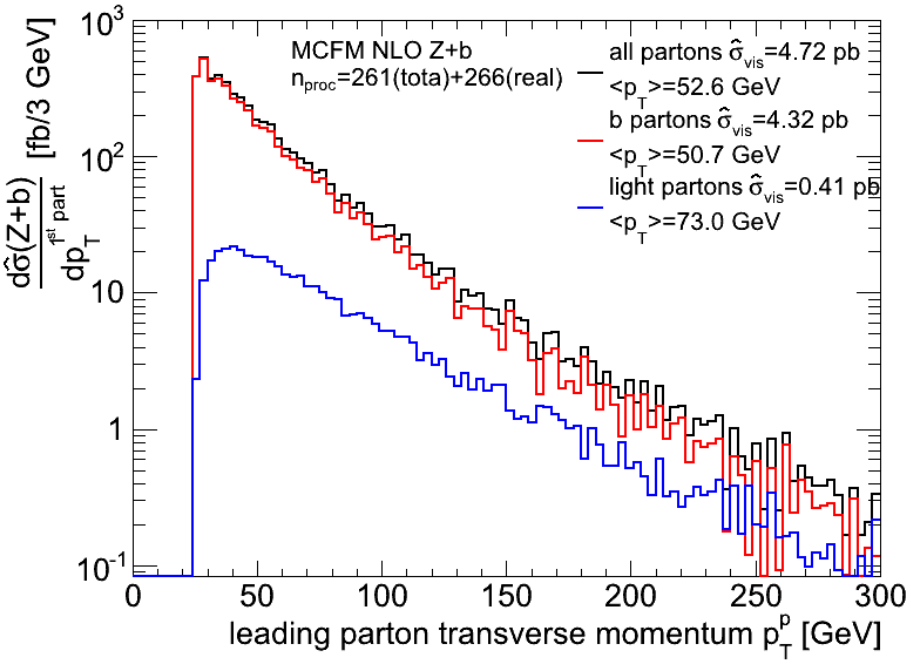

Figure B.2: Spectrum of the leading parton transverse momentum for the sum of the processes 261 and 266: the red curve refers to b partons, while the blue curve to light patrons.

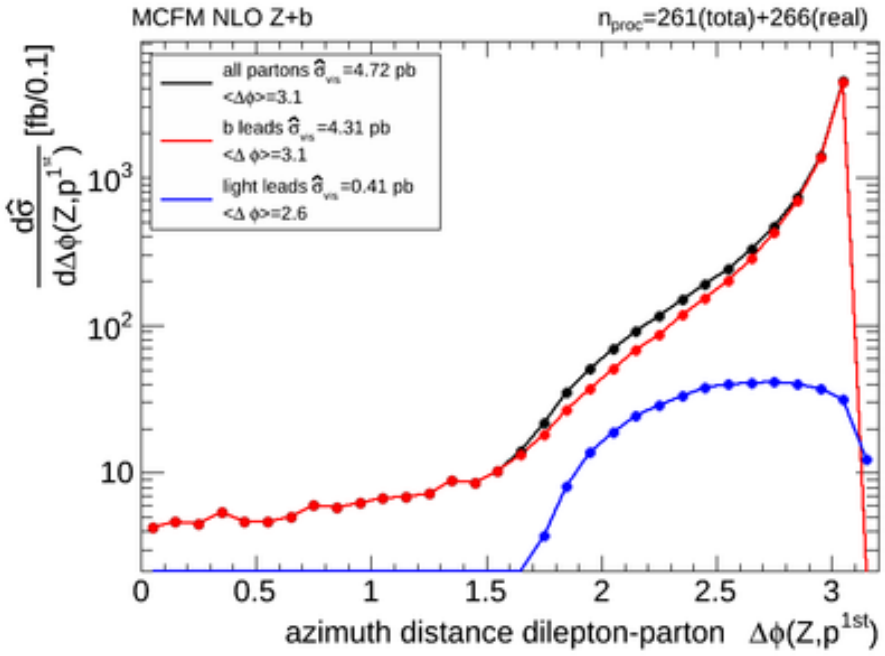

Figure B.3: Distribution of the azimuthal distance between the dilepton and the parton: the red curve refers to b partons, while the blue curve to light partons.

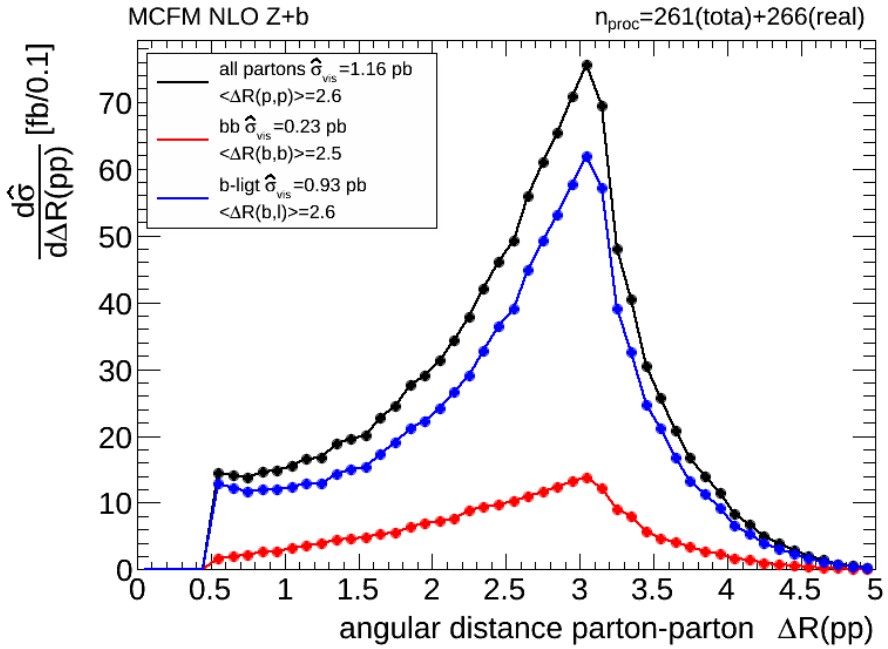

Figure B.4: Distribution of the azimuthal distance between the two partons: the red curve refers to b partons, while the blue curve to light partons.

Bibliography

[1] Alessandro Bettini. *Introduction to Elementary Particle Physics*. Cambridge University Press, 2012.

[2] Cristina Botta. *Search for a SM Higgs boson in the decay channel $H \to ZZ^* \to 4l$ with the CMS experiment*. PhD thesis, 2011.

[3] J. Campbell, R. K. Ellis, F. Maltoni, and S. Willenbrock. Associated production of a Z boson and a single heavy-quark jet. *Phys. Rev. D*, 69:074021, Apr 2004.

[4] J. Campbell, R. K. Ellis, F. Maltoni, and S. Willenbrock. Production of a Z boson and two jets with one heavy-quark tag. *Phys. Rev. D*, 73:054007, Mar 2006.

[5] Roberto Castello. *Alignment of CMS Tracker detector using cosmic ray particles and its impact on early physics performance*. PhD thesis, 2009.

[6] The CMS collaboration. CMS Physics Technical Design Report. *CERN/LHCC, 2006-001,2006-021*, 2006.

[7] The CMS collaboration. Commissioning of b-jet identification with pp collisions at $\sqrt(s) = 7$ TeV, CMS-PAS-BTV-10-001. 2010.

[8] The CMS collaboration. Performance of the b-jet identification in CMS, CMS-PAS-BTV-11-001. 2011.

[9] The CMS collaboration. b-jet identification in the CMS experiment, CMS-PAS-BTV-11-003. 2012.

[10] The CMS collaboration. Measurement of btagging efficiency using ttbar events, CMS-PAS-BTV-11-004. 2012.

[11] The CMS collaboration. Measurement of the z/γ^*+b-jet cross section in pp collisions at 7 TeV. *Jhep*, 043P:0412, 2012.

[12] K. Nakamura et al. (Particle Data Group). *J. Phys. G 37, 075021*, 2010.

[13] Francis Halzen and Alan D. Martin. *Quarks and Leptons: An Introductory Course in Modern Particle Physics*. Wiley, 1984.

[14] J.Maes. *Estimation of b-tag efficiency using top quarks*. PhD thesis, 2010.

[15] Marco Musich. *The Alignment of the CMS Tracker and its Impact on the Early Quarkonium Physics*. PhD thesis, 2010.

[16] George Sterman. *An Introduction to Quantum Field Theory*. Cambridge University Press, 1993.